HAL
open science

Cartographie de l'occupation des sols à partir de séries temporelles d'images satellitaires à hautes résolutions Identification et traitement des données mal étiquetées

Charlotte Pelletier

To cite this version:

Charlotte Pelletier. Cartographie de l'occupation des sols à partir de séries temporelles d'images satellitaires à hautes résolutions Identification et traitement des données mal étiquetées . Interfaces continentales, environnement. Universite Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), 2017. Français. NNT : . tel-01665121v1

HAL Id: tel-01665121

https://theses.hal.science/tel-01665121v1

Submitted on 15 Dec 2017 (v1), last revised 6 Mar 2018 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *11/12/2017* **par :**

CHARLOTTE PELLETIER

Cartographie de l'occupation des sols à partir de séries temporelles d'images satellitaires à hautes résolutions

Identification et traitement des données mal étiquetées

JURY

Laurent POLIDORI	Directeur de Recherche Université Toulouse 3	Président
Samia BOUKIR	Professeure INP Bordeaux	Rapporteure
Sébastien LEFEVRE	Professeur Université Bretagne Sud	Rapporteur
Devis TUIA	Professeur Université de Wageningen	Examinateur
Mathieu FAUVEL	Maître de Conférences INRA	Examinateur
Grégoire MERCIER	Ingénieur SAS eXO maKina	Examinateur
Gérard DEDIEU	Directeur de Recherche CNES	Co-directeur
Silvia VALERO	Maître de Conférences Université Toulouse 3	Co-directrice
Nicolas CHAMPION	Ingénieur IGN Espace	Membre invité

École doctorale et spécialité :
SDU2E : Surfaces et interfaces continentales, Hydrologie

Unité de Recherche :
Centre d'Études Spatiales de la Biosphère (CESBIO) – UMR 5126

Directeur(s) de Thèse :
Gérard DEDIEU et *Silvia* VALERO

Co-encadrant :
Nicolas CHAMPION

Résumé

L'étude des surfaces continentales est devenue ces dernières années un enjeu majeur à l'échelle mondiale pour la gestion et le suivi des territoires, notamment en matière de consommation des terres agricoles et d'étalement urbain. Dans ce contexte, les cartes d'occupation du sol caractérisant la couverture biophysique des terres émergées jouent un rôle essentiel pour la cartographie des surfaces continentales.

La production de ces cartes sur de grandes étendues s'appuie sur des données satellitaires qui permettent de photographier les surfaces continentales fréquemment et à faible coût. Le lancement de nouvelles constellations satellitaires – Landsat-8 et Sentinel-2 – permet depuis quelques années l'acquisition de séries temporelles à hautes résolutions. Ces dernières sont utilisées dans des processus de classification supervisée afin de produire les cartes d'occupation du sol. L'arrivée de ces nouvelles données ouvre de nouvelles perspectives, mais questionne sur le choix des algorithmes de classification et des données à fournir en entrée du système de classification.

Outre les données satellitaires, les algorithmes de classification supervisée utilisent des échantillons d'apprentissage pour définir leur règle de décision. Dans notre cas, ces échantillons sont étiquetés, *i.e.* la classe associée à une occupation des sols est connue. Ainsi, la qualité de la carte d'occupation des sols est directement liée à la qualité des étiquettes des échantillons d'apprentissage. Or, la classification sur de grandes étendues nécessite un grand nombre d'échantillons, qui caractérise la diversité des paysages. Cependant, la collecte de données de référence est une tâche longue et fastidieuse. Ainsi, les échantillons d'apprentissage sont bien souvent extraits d'anciennes bases de données pour obtenir un nombre conséquent d'échantillons sur l'ensemble de la surface à cartographier. Cependant, l'utilisation de ces anciennes données pour classer des images satellitaires plus récentes conduit à la présence de nombreuses données mal étiquetées parmi les échantillons d'apprentissage. Malheureusement, l'utilisation de ces échantillons mal étiquetés dans le processus de classification peut engendrer des erreurs de classification, et donc une détérioration de la qualité de la carte produite.

L'objectif général de la thèse vise à améliorer la classification des nouvelles séries temporelles d'images satellitaires à hautes résolutions. Le premier objectif consiste à déterminer la stabilité et la robustesse des méthodes de classification sur de grandes étendues. Plus particulièrement, les travaux portent sur l'analyse d'algorithmes de classification et la sensibilité de ces algorithmes vis-à-vis de leurs paramètres et des données en entrée du système de classification. De plus, la robustesse de ces algorithmes à la présence des données imparfaites est étudiée. Le second objectif s'intéresse aux erreurs présentes dans les données d'apprentissage, connues sous le nom de données mal étiquetées. Dans un premier temps, des méthodes de détection de données mal étiquetées sont proposées et étudiées. Dans un second temps, un cadre méthodologique est proposé afin de prendre en compte les données mal étiquetées dans le processus de classification. L'objectif est de réduire l'influence des données mal étiquetées sur les performances de l'algorithme de classification, et donc d'améliorer la carte d'occupation des sols produite.

Abstract

Land surface monitoring is a key challenge for diverse applications such as environment, forestry, hydrology and geology. Such monitoring is particularly helpful for the management of territories and the prediction of climate trends. For this purpose, mapping approaches that employ satellite-based Earth Observations at different spatial and temporal scales are used to obtain the land surface characteristics.

More precisely, supervised classification algorithms that exploit satellite data present many advantages compared to other mapping methods. In addition, the recent launches of new satellite constellations – Landsat-8 and Sentinel-2 – enable the acquisition of satellite image time series at high spatial and spectral resolutions, that are of great interest to describe vegetation land cover. These satellite data open new perspectives, but also interrogate the choice of classification algorithms and the choice of input data.

In addition, learning classification algorithms over large areas require a substantial number of instances per land cover class describing landscape variability. Accordingly, training data can be extracted from existing maps or specific existing databases, such as crop parcel farmer's declaration or government databases. When using these databases, the main drawbacks are the lack of accuracy and update problems due to a long production time. Unfortunately, the use of these imperfect training data lead to the presence of mislabeled training instance that may impact the classification performance, and so the quality of the produced land cover map.

Taking into account the above challenges, this Ph.D. work aims at improving the classification of new satellite image time series at high resolutions. The work has been divided into two main parts. The first Ph.D. goal consists in studying different classification systems by evaluating two classification algorithms with several input datasets. In addition, the stability and the robustness of the classification methods are discussed. The second goal deals with the errors contained in the training data. Firstly, methods for the detection of mislabeled data are proposed and analyzed. Secondly, a filtering method is proposed to take into account the mislabeled data in the classification framework. The objective is to reduce the influence of mislabeled data on the classification performance, and thus to improve the produced land cover map.

Remerciements

L'écriture de ces remerciements est l'occasion de mettre un point final à ces trois dernières années dédiées à la réalisation de ces travaux, mais surtout d'exprimer ma reconnaissance à tous ceux qui ont permis l'aboutissement de ces travaux.

Ainsi, mes premiers remerciements sont destinés à Silvia Valero, co-directrice de ces travaux. Merci de m'avoir laissé l'opportunité de mener à bien ces travaux dans un environnement à la fois riche et paisible. Ton soutien, ton encadrement, et ta confiance ont rendu cette expérience très enrichissante. Mes deuxièmes remerciements sont adressés à Gérard Dedieu, co-directeur. Ta présence et nos échanges fructueux ont permis le déroulement serein de ces trois années. Je suis aussi reconnaissante envers l'ensemble des personnes du CESBIO qui m'ont aidé à accomplir ses travaux, en particulier Jordi Inglada et Claire Marais-Sicre qui ont suivi avec intérêt mes avancées.

Ces travaux de thèse sont aussi le fruit d'une collaboration avec l'IGN. Merci Nicolas Champion pour ton suivi régulier, et ton soutien tout au long de ses travaux. Je tiens également à remercier les membres du MATIS pour leur accueil chaleureux lors de mes séjours parisiens.

Par ailleurs, je tiens à remercier l'ensemble des membres de mon jury pour leur intérêt porté à ces travaux. Je tiens en particulier à remercier les deux rapporteurs de ce manuscrit – Samia Boukir, Professeure INP-ENSEGID à Bordeaux et Sébastien Lefevre, Professeur Université Bretagne Sud à Vannes. Vos remarques, commentaires et suggestions m'ont permis d'avoir un nouveau regard sur l'ensemble de ce manuscrit.

La thèse est aussi une grande aventure humaine : merci à tous les amis et aux collègues du CESBIO. Votre présence a permis de fêter les réussites, mais aussi d'adoucir les périodes plus difficiles.

Je vais finalement conclure en exprimant toute ma gratitude à ma famille, en particulier à mes parents. Merci pour vos encouragements, votre enthousiasme et surtout votre soutien dans l'ensemble de mes choix professionnels, mais aussi personnels. Votre bienveillance et votre optimisme a grandement contribué à l'accomplissement et la réussite de ces travaux. Je compte encore sur vous pour les prochaines aventures.

Table des matières

Table des matières vii

Liste des figures xi

Liste des tableaux xv

I Introduction 1

1 Contexte 3
1.1 Cartographie de l'occupation des sols 4
1.2 Télédétection spatiale 7
1.3 Des images vers la carte 14
1.4 Position du problème 21

II Méthodes et données 25

2 Classification supervisée de séries temporelles d'images satellitaires 27
2.1 Introduction à l'apprentissage supervisé 28
2.2 Classification supervisée de séries temporelles 30
2.3 Choix des algorithmes de classification 32
2.4 Algorithmes de classification supervisée 34
2.5 Évaluation des performances des algorithmes de classification 46

3 Données utilisées 51
3.1 Données satellitaires utilisées 51
3.2 Pré-traitements des données satellitaires 55
3.3 Données de référence utilisées 61

III Stabilité et robustesse des algorithmes de classification 65

4 Études des algorithmes de classification sur de grandes étendues 67
4.1 Données en entrée du système de classification 68
4.2 Présentation des expérimentations 74
4.3 Résultats des expérimentations 79
4.4 Conclusion 91

5 Influence des données mal étiquetées sur la qualité des cartes d'occupation des sols **93**
5.1 Problématique des données mal étiquetées 94
5.2 Présentation des expérimentations 97
5.3 Résultats des expérimentations 105
5.4 Conclusion 115

IV Détection des données mal étiquetées **117**

6 Détection des données mal étiquetées **119**
6.1 Détection de données mal étiquetées 120
6.2 Détection de données mal étiquetées avec le *Random Forest* 132
6.3 Présentation des expérimentations 135
6.4 Résultats des expérimentations 139
6.5 Conclusion 155

7 Prise en compte des données mal étiquetées **157**
7.1 Filtrage des données mal étiquetées 158
7.2 Filtrage itératif des données mal étiquetées 164
7.3 Présentation des expérimentations 171
7.4 Résultats des expérimentations 175
7.5 Conclusion 202

V Conclusion **205**

8 Conclusion générale **207**
8.1 Conclusions 207
8.2 Perspectives 210

Bibliographie **216**

Acronymes **241**

A Données satellitaires et données de référence **247**
A.1 Dates des images satellitaires utilisées 247
A.2 Niveau de traitements des données satellitaires 250
A.3 Tuilage 251
A.4 Nomenclature 251

B Compléments sur le *Random Forest* **255**
B.1 Tirages aléatoires 255
B.2 Matrice de proximité 257

C Liste des publications **259**
C.1 Journal international à comité de lecture 259
C.2 Conférence internationale à comité de lecture 259

Liste des figures

1.1 Exemple de carte d'occupation des sols. 5
1.2 Exemple de cartes d'occupation des sols, Toulouse, France. 6
1.3 Caractéristiques des séries temporelles d'images satellitaires. 11
1.4 Illustration de la notion de résolution spatiale. 11
1.5 Exemple de profils spectraux pour six occupations des sols. 12
1.6 Profils spectraux de deux échantillons, maïs et tournesol, à deux dates différentes. 13
1.7 Exemple de profils *Normalized Difference Vegetation Index* en fonction du jour de l'année pour cinq occupations des sols. 13
1.8 Évolution de la forêt des Landes de Gasgogne, France, entre 2007 et 2016. . 14
1.9 CORINE *Land Cover* 2012 16
1.10 *High Resolution Layers* 2012. 17
1.11 Schémas simplifiés des apprentissages supervisé et non-supervisé pour la cartographie de l'occupation des sols. 19
1.12 Processus de classification supervisée. 22

2.1 Processus de classification supervisée. 28
2.2 Illustration des problèmes de sur-apprentissage et sous-apprentissage. . . . 29
2.3 Exemple de discrimination binaire. 34
2.4 Illustration de la notion de marge. 35
2.5 Illustration de la notion de variables ressort dans le cas de données non-linéairement séparables. 37
2.6 Validation croisée sur cinq partitions. 39
2.7 Exemple d'arbre de décision binaire pour la classification des cultures en fonction de différentes informations caractérisant leur cycle phénologique. . 42
2.8 Principe de construction d'un arbre de décision binaire pour un problème de classification binaire à deux classes. 44
2.9 Exemple d'une carte d'occupation des sols à évaluer à partir d'échantillons test extraits des données de référence. 46

3.1 Visualisation et caractéristiques des trois zones d'étude. 54
3.2 Tuiles utilisées pour les données Landsat-8 et Sentinel-2. 55
3.3 Corrections radiométriques : des comptes numériques aux réflectances *Top-of-Canopy*. 56
3.4 Reconstruction de données manquantes par interpolation temporelle linéaire. 58
3.5 Visibilité des pixels pour une série temporelle d'images Sentinel-2. 59
3.6 Reconstruction de données manquantes et utilisation de dates interpolées pour deux profils représentés par différentes dates. 60
3.7 OCcupation des Sols à Grande Échelle, Tarbes 2013. 63

4.1 Exemple d'indices spectraux. 70
4.2 Modélisation d'un profil de *Normalized Difference Vegetation Index* par une double logistique. 73
4.3 Distribution temporelle des images SPOT-4 et Landsat-8 pour la première zone d'étude en fonction du jour de l'année. 75
4.4 Localisation des zones d'étude. 75
4.5 Caractéristique de la seconde zone d'étude. 79
4.6 Distribution des échantillons test pour chaque classe d'occupation des sols en fonction de la distance à la zone d'apprentissage. 90
4.7 OA obtenues pour les trois jeux de variables en fonction de la distance à la zone d'apprentissage. 91

5.1 Illustration de différentes imperfections dans les bases de données. 95
5.2 Exemple d'un profil de végétation au cours de l'année (DoY : *Day of Year*). 99
5.3 Exemple de profils simulés de *Normalized Difference Vegetation Index* pour l'ensemble du cycle phénologique. 101
5.4 *Overall Accuracy* moyenné sur dix tirages aléatoires en fonction du niveau de bruit. 106
5.5 *Overall Accuracy* moyenné sur dix tirages aléatoires en fonction du niveau de bruit. 108
5.6 *Overall Accuracy* moyenné sur dix tirages aléatoires en fonction du niveau de bruit. 109
5.7 Complexité des algorithmes de classification en fonction du niveau de bruit pour les données simulées et SPOT-Landsat à cinq classes. 110
5.8 *Overall Accuracy* moyenné sur dix tirages aléatoires en fonction du niveau de bruit pour trois algorithmes de classification. 112
5.9 Valeurs d'*Overall Accuracy* obtenues sur la grille de recherche lors de l'optimisation des paramètres du SVM sur les données SPOT-Landsat à cinq classes décrites par les profils de NDVI pour des niveaux de bruit de 0, 20 et 40 %. 115

6.1 Définition d'un *outlier* basée sur la distance selon Knorr and Ng [1998]. . . 124
6.2 Définition d'un *outlier* selon Ramaswamy et al. [2000]. 125
6.3 Illustration de la notion de densité pour deux échantillons 126
6.4 Distribution d'échantillons décrits par deux *clusters* de différentes densité. 127
6.5 Illustration des notions de *semantic* et de *cross-outlier*. 128
6.6 Illustration de la notion de densité pour deux échantillons 130
6.7 Exemples d'arbres de décision. 134
6.8 Exemple de courbes *Receiver Operating Characteristic* et de valeur d'*Area Under the Curve*. 139
6.9 Évolution du ROC-AUC en fonction du paramètre k pour les méthodes kNN, kNW et LOF pour les données simulées et SPOT-Landsat à cinq classes pour quatre niveaux de bruit 10, 20, 30 et 40 %. 141
6.10 Évolution du ROC-AUC en fonction du paramètre k pour les méthodes kNN, kNW et LOF pour les données simulées et SPOT-Landsat à cinq classes pour un niveau de bruit de 30 %. 142
6.11 Évolution du F-Score en fonction des paramètres k et n pour les données simulées à cinq classes en fonction de quatre niveaux de bruit 10, 20, 30 et 40 %. La croix rouge représente le meilleur F-Score. 143

6.12 Évolution du F-Score en fonction des paramètres k et n pour les données SPOT-Landsat à cinq classes en fonction de quatre niveaux de bruit 10, 20, 30 et 40 %. 144
6.13 Évolution du F-Score en fonction du paramètre k pour les méthodes ENN, RENN et All$k$NN pour les données simulées et SPOT-Landsat à cinq classes pour quatre niveaux de bruit 10, 20, 30 et 40 %. 145
6.14 Évaluation de la précision des méthodes de détection de données mal étiquetées pour les données simulées et SPOT-Landsat à cinq classes. 147
6.15 Courbe ROC pour les données Sentinel-2 152

7.1 Processus de classification supervisée avec une étape de filtrage des données mal étiquetées avant l'étape d'apprentissage supervisé. 158
7.2 Illustration du principe général du filtrage des données mal étiquetées. . . . 159
7.3 Stratégies possibles pour la détection des données mal étiquetées. 159
7.4 Principe simplifié des méthodes de détection de données mal étiquetées basées sur un ensemble d'algorithmes de classification. 161
7.5 Illustration du principe du filtrage itératif. 165
7.6 Illustration des types d'erreurs lors du filtrage des données mal étiquetées . 172
7.7 Valeurs d'*Overall Accuracy* et de précision de la méthode *Edited Nearest Neighbor* en fonction du paramètre de voisinage k. 177
7.8 Valeurs d'*Overall Accuracy* pour les méthodes Breiman, DistanceLCA et PuretyLCA en fonction du paramètre de seuil n. 179
7.9 Évolution de la précision du filtre (FP, ER_1 et ER_2) en fonction du paramètre de seuil n pour les méthodes Breiman, DistanceLCA, PuretyLCA. Les données simulées avec 20 et 40 % de données mal étiquetées sont utilisées. 181
7.10 Évolution de la précision du filtre (FP, ER_1 et ER_2) en fonction du paramètre de seuil n pour les méthodes Breiman, DistanceLCA, PuretyLCA. Les données SPOT-Landsat avec 20 et 40 % de données mal étiquetées sont utilisées. 182
7.11 Évolution de l'*Overall Accuracy* et des scores d'*outlier* en fonction du paramètre de seuil n pour les méthodes Breiman, DistanceLCA, PuretyLCA. 184
7.12 Évolution de l'*Overall Accuracy* au cours des itérations pour différents processus itératifs basés sur les méthodes d'édition. 187
7.13 Évolution de l'*Overall Accuracy* au cours des itérations pour les processus itératifs basés sur les scores O_{RF} pour les données simulées. 188
7.14 Évolution de l'*Overall Accuracy* au cours des itérations pour les processus itératifs basés sur les scores O_{RF} pour les données SPOT-Landsat. 189
7.15 Évolution de la précision du filtre (FP, ER_1 et ER_2) en fonction des itérations t pour les deux processus itératifs global et par classe. 190
7.16 Évaluation des critères d'arrêt pour les filtrages itératifs global et par classe pour les données simulées. 192
7.17 Évolution de l'*Overall Accuracy* au cours des itérations pour les processus itératifs basés sur les prédictions des arbres du *Random Forest* appliqués sur les données simulées pour 20 et 40 % de données mal étiquetées. 193
7.18 Évolution de l'*Overall Accuracy* au cours des itérations pour les processus itératifs basés sur les prédictions des arbres du *Random Forest* appliqués sur les données SPOT-Landsat contaminées par 20 et 40 % de données mal étiquetées 194

7.19 Valeurs d'*Overall Accuracy* (OA) et de précision de la méthode *Edited Nearest Neighbor* (ENN) en fonction du paramètre de voisinage k pour les données Sentinel-2. 197
7.20 Valeurs d'*Overall Accuracy* (OA) pour différents filtrages itératifs sur les données Sentinel-2. 197
7.21 Étude du critère d'arrêt sur les données Sentinel-2 pour les filtrages itératifs global et par classe. 201

A.1 Systèmes de tuilage utilisés lors de la distribution des images Landsat-8 et Sentinel-2. 251
A.2 Nomenclature complète *Land Cover Classification System* 252
A.3 Légende 254

B.1 Illustration de différentes matrices de proximité. 257

Liste des tableaux

1.1 Recommandation du *Global Climate Observing System* pour l'observation des occupations des sols dans le cadre des études sur le changement climatique. 7

2.1 Référentiel pour interpréter la valeur de Kappa. 48

3.1 Caractéristiques de l'instrument Haute Résolution Visible et Infra-Rouge embarqué sur le satellite SPOT-4. 52

3.2 Caractéristiques de l'instrument *Operational Land Imager* embarqué sur le satellite Landsat-8. 53

3.3 Caractéristiques de l'instrument *Multi-Spectral Instrument* embarqué sur les deux satellites Sentinel-2. 53

3.4 Données du Registre Parcellaire Graphique utilisées pour les trois zones d'études. 62

4.1 Description des primitives spectrales utilisées. 71

4.2 Nombre total de variables pour chaque zone d'étude en fonction des jeux de variables utilisés. 76

4.3 Nombre d'échantillons d'apprentissage et validation pour différentes configurations. 80

4.4 Précisions, rappels et F-Scores par classe et *Overall Accuracy* moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le RF et le SVM. 81

4.5 Précisions, rappels, et F-Scores par classe et *Overall Accuracy* moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le SVM pour différents nombres d'échantillons d'apprentissage. 83

4.6 Précisions, rappels, et F-Scores par classe et *Overall Accuracy* (OA) moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le RF pour différents nombres d'échantillons d'apprentissage. . 84

4.7 Temps de calcul pour l'apprentissage (avec les écarts-types) en secondes pour le SVM et le RF. 85

4.8 *Overall Accuracy* moyennés sur cinq tirages aléatoires obtenus pour le RF en utilisant différentes valeurs de paramètres. 86

4.9 *Overall Accuracy* (OA) moyennés sur cinq tirages aléatoires obtenus pour le RF en utilisant différentes valeurs de paramètres. 88

4.10 Précisions, rappels, F-Scores par classe et *Overall Accuracy* moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le RF pour différents vecteurs de variables. 89

5.1 Valeurs minimales et maximales des paramètres de la double logistique pour dix classes d'occupation des sols. 100
5.2 Occupation des sols utilisés pour chaque jeu de données simulées. 101
5.3 Nombre de polygones disponibles par classe dans le Registre Parcellaire Graphique (RPG). 102
5.4 Description des jeux de données SPOT-Landsat. 103
5.5 Choix des étiquettes corrompues dans le cas de l'ajout d'un bruit aléatoire systématique. 112
5.6 Valeurs du paramètre γ optimisées par la validation croisée sur cinq partitions pour l'algorithme du SVM-RBF pour les données SPOT-Landsat à cinq classes composées d'un total de 2500 échantillons d'apprentissage. . . 114

6.1 Exemple de matrice de confusion pour un problème de détection de données mal étiquetées. 138
6.2 Valeurs de rappel) et précision obtenues pour les différentes méthodes de détection de données mal étiquetées sur les données simulées. 149
6.3 Valeurs de rappel et précision obtenues pour les différentes méthodes de détection de données mal étiquetées sur les données SPOT-Landsat. 149
6.4 Précisions à n, $P@n$ pour $n = 10$, $n = 50$ et $n = 100$ pour quatre niveaux de bruit (10, 20, 30 et 40 %) pour les données simulées à cinq classes. . . . 150
6.5 Précisions à n, $P@n$ pour $n = 10$, $n = 50$ et $n = 100$ pour quatre niveaux de bruit (10, 20, 30 et 40 %) pour les données SPOT-Landsat à cinq classes. 150
6.6 Nombre d'échantillons pour les données Sentinel-2 en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) 151
6.7 Nombre d'échantillons utilisés pour les expérimentations sur les données Sentinel-2 en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) 152
6.8 F-Scores et précisions à n (avec n égal à 10, 50 et 100) obtenus pour les différentes méthodes de détection de données mal étiquetées sur les données Sentinel-2. 153
6.9 Valeurs de F-Score pour les méthodes de détection de données mal étiquetées. 154
6.10 Valeurs de rappel pour les méthodes de détection de données mal étiquetées. 154
6.11 Valeurs de précision pour les méthodes de détection de données mal étiquetées. 154

7.1 Filtrages itératifs proposés basés sur les scores d'*outlier* du *Random Forest*. 168
7.2 Filtrages itératifs proposés basés sur les prédictions des arbres du *Random Forest*. 170
7.3 Valeurs d'*Overall Accuracy* obtenues pour les données simulées et SPOT-Landsat dans différentes configurations. 173
7.4 Nombre d'échantillons test, nombre d'échantillons d'apprentissage et pourcentage de données mal étiquetées pour les données Sentinel-2. 174
7.5 Valeurs d'*Overall Accuracy* et de F-Scores obtenues pour les données Sentinel-2 lorsque les données sans bruit, sans filtrage et parfaitement filtrées sont utilisées pour l'apprentissage d'un *Random Forest*. 175
7.6 Nombre d'échantillons d'apprentissage par classe restant pour différentes valeurs de k après l'utilisation de la méthode *Edited Nearest Neighbor* pour les données simulées. 178

7.7 Nombre d'échantillons d'apprentissage restant pour différentes valeurs de n après un filtrage basé sur les scores classiques d'*outliers* du *Random Forest* pour les données simulées. 180
7.8 Valeurs de l'*Overall Accuracy* obtenues pour les filtrages non-itératifs basés sur les méthodes d'édition, les scores d'*outlier* du *Random Forest* (RF) et les prédictions des arbres du RF. 185
7.9 Valeurs de l'*Overall Accuracy* obtenues pour les filtrages itératifs basés sur les méthodes d'édition, les scores d'*outlier* du *Random Forest* (RF) et les prédictions des arbres du RF. 195
7.10 Valeurs d'*Overall Accuracy* et de F-Scores obtenues pour les données Sentinel-2 pour différentes stratégies de filtrage. 198
7.11 Valeurs des précisions FP, ER_1 et ER_2 pour le filtrage *Edited Nearest Neighbor* (ENN) et les filtrages itératifs global et par classe. 199
7.12 Nombre d'échantillons d'apprentissage avant l'étape de filtrage en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) 200
7.13 Nombre d'échantillons utilisés pour l'apprentissage, pour les données Sentinel-2, en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) après un filtrage *Edited Nearest Neighbor* ($k = 21$). 200
7.14 Nombre d'échantillons utilisés pour l'apprentissage, pour les données Sentinel-2, en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) après un filtrage itératif global ($t = 11$). 200
7.15 Nombre d'échantillons utilisés pour l'apprentissage, pour les données Sentinel-2, en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) après un filtrage itératif par classe ($t = 7$). 200

A.1 Images disponibles pour la série temporelle composée d'images SPOT-4 et Landsat-8 en 2013. 247
A.2 Images disponibles pour la série temporelle composée d'images Landsat-8 en 2013 pour huit tuiles (nomenclature USGS). 248
A.3 Images disponibles pour la série temporelle composée d'images Sentinel-2A acquises fin 2015 et sur l'année 2016 pour six tuiles (nomenclature ESA). . 249
A.4 Niveaux de pré-traitements effectués sur les images SPOT-5 par l'organisme distributeur Spot Image. 250
A.5 Nomenclature hiérarchique *Land Cover Classification System* 252
A.6 Nomenclature du Registre Parcellaire Graphique 253

Première partie

Introduction

Chapitre 1

Contexte

Sommaire

1.1 Cartographie de l'occupation des sols 4
1.2 Télédétection spatiale 7
1.2.1 Généralités sur les satellites imageurs 8
1.2.2 Évolution des capteurs vers la haute résolution 9
1.2.3 Séries temporelles d'images satellitaires 10
1.3 Des images vers la carte 14
1.3.1 Approches manuelles 15
1.3.2 Approches automatiques 19
1.4 Position du problème 21
1.4.1 Défis 22
1.4.2 Objectifs 23
1.4.3 Organisation du manuscrit 24

Depuis plusieurs décennies, l'observation de la Terre permet de mieux comprendre notre planète. Au cœur des enjeux sur les changements globaux, la caractérisation des dynamiques liées aux transformations des surfaces continentales – consommation des surfaces agricoles, déforestation ou encore étalement urbain – est essentielle.

Dans ce contexte, la télédétection spatiale offre la possibilité de cartographier fréquemment l'ensemble de la planète. Plus spécifiquement, les images issues des acquisitions satellitaires permettent de produire des cartes qui donnent une représentation graphique relative aux surfaces terrestres comme l'occupation des sols.

Dans un premier temps, la notion et les applications de la cartographie de l'occupation des sols sont présentées. Dans un deuxième temps, l'évolution des capteurs satellitaires servant à la production de ces cartes est détaillée. En particulier, les caractéristiques des nouvelles séries temporelles d'images satellitaires sont mises en avant. Puis, les moyens de production des cartes d'occupation des sols à partir d'images satellitaires sont décrits. Enfin, la problématique traitée dans ce manuscrit est introduite, avec notamment le détail des objectifs.

1.1 Cartographie de l'occupation des sols

L'occupation des sols ou la couverture des sols (*land cover* en anglais) décrit la couverture bio-physique de la surface des terres émergées. Elle est identifiée par le programme *Global Climate Observing System* (GCOS) comme l'une des cinq Variables Climatiques Essentielles (VCE) [1] hautement prioritaires [GCOS, 2016]. Ces variables sont sélectionnées pour leur importance dans le cadre de la compréhension et de la prédiction des évolutions du climat, ainsi que pour leur caractère indispensable pour guider les mesures sur l'adaptation et l'atténuation aux changements climatiques. Elles permettent entre autres de soutenir le travail du Groupe d'experts Intergouvernemental sur l'Évolution du Climat (GIEC) [IPCC, 2014].

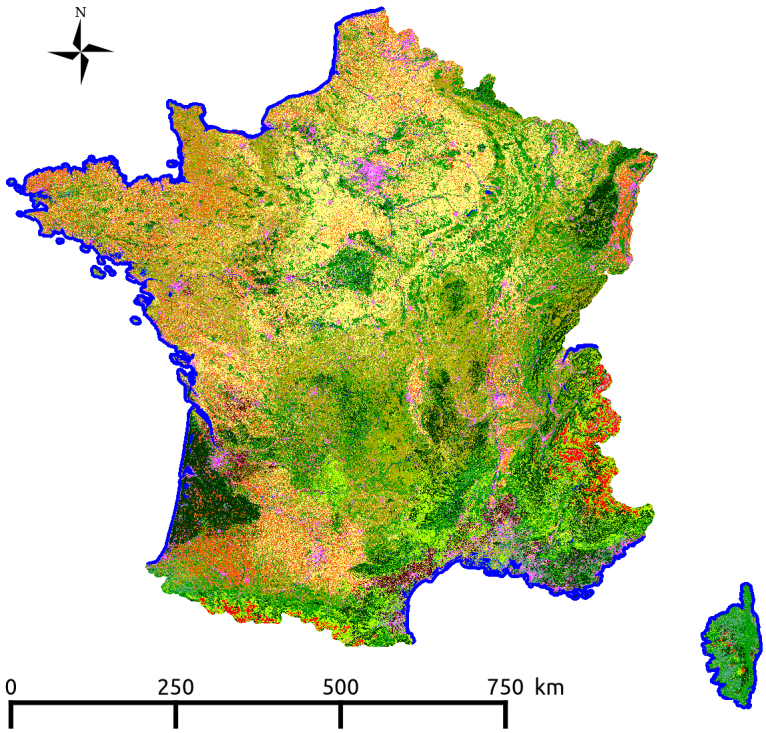
Les cartes décrivant l'occupation des sols sont de puissants outils scientifiques et décisionnels. Elles sont utilisées dans des travaux de recherche comme entrée des systèmes de modélisation des cycles de l'eau et du carbone ou encore pour les bilans d'énergie [Claverie et al., 2012; Houghton et al., 2012]. Elles servent également pour des applications opérationnelles, notamment pour le suivi des changements globaux, et de supports pour appliquer les consignes et recommandations des politiques publiques qui nécessitent une connaissance précise des territoires [Feddema et al., 2005; IPCC, 2014; Pielke, 2005].

L'importance de ces cartes a donc conduit à l'émergence de plusieurs initiatives visant à produire des cartes d'occupation des sols et de changements d'occupation des sols. Au niveau international, le projet *Global Land Programme* (GLP) mené par *Future Earth* [2] a par exemple pour objectif de caractériser les changements d'occupation des sols, mais aussi de favoriser l'émergence d'une communauté et de produits accessibles à un plus grand nombre. Les cartes d'occupation des sols sont aussi un outil clé pour le suivi des cultures à grande échelle [Inglada et al., 2015; Valero et al., 2016], qui est utile dans le cadre de la sécurité alimentaire. Par exemple, le projet *GEO Global Agricultural Monitoring* (GEOGLAM) [3] créé en 2001 vise à fournir de façon opérationnelle des prédictions de récolte aux échelles nationale et mondiale à partir de ces cartes [Whitcraft et al., 2015].

En France, une initiative portée par le pôle de données et de services surfaces continentales Theia a pour vocation de définir et de développer des algorithmes pour automatiser la production des cartes d'occupation des sols à travers le Centre d'Expertise Scientifique de l'Occupation des Sols (CES OSO). Une seconde initiative nationale récente, l'OCcupation des Sols à Grande Échelle (OCS-GE) portée par l'Institut National de l'Information Géographique et Forestière (IGN), vise à répondre aux problématiques d'aménagement du territoire tout en préservant la biodiversité, les terres agricoles et les continuités écologiques - haies, corridors ou encore espaces boisés – pour notamment satisfaire la loi Grenelle II [4]. Autour de ces problématiques, de nombreux travaux de recherche se sont développés pour cartographier les infrastructures urbaines ou la Trame Verte et Bleue [Maire et al., 2012; Masse, 2013; Sheeren et al., 2012].

À titre d'exemple, la Figure 1.1 montre la carte d'occupation des sols pour la France en 2016 obtenue par le CES OSO. Chaque élément de la carte est associé à une couverture des sols représentée par une couleur. De manière générale, les cartes d'occupation des sols sont caractérisées par :

1. Une VCE est une variable ou un groupe de variables physique, chimique ou biologique qui contribue significativement à la caractérisation du climat [Bojinski et al., 2014].
2. Le projet était nommé *Land Use Land Cover Change* (LULCC) avant 2015, et était conduit par l'*International Geosphere-Biosphere Program* (IGBP) et l'*International Human Dimensions Program on Global Environmental Change* (IHDP).
3. www.earthobservations.org/geoglam.php
4. Loi N° 2010-788 du 12 juillet 2010 portant engagement national pour l'environnement

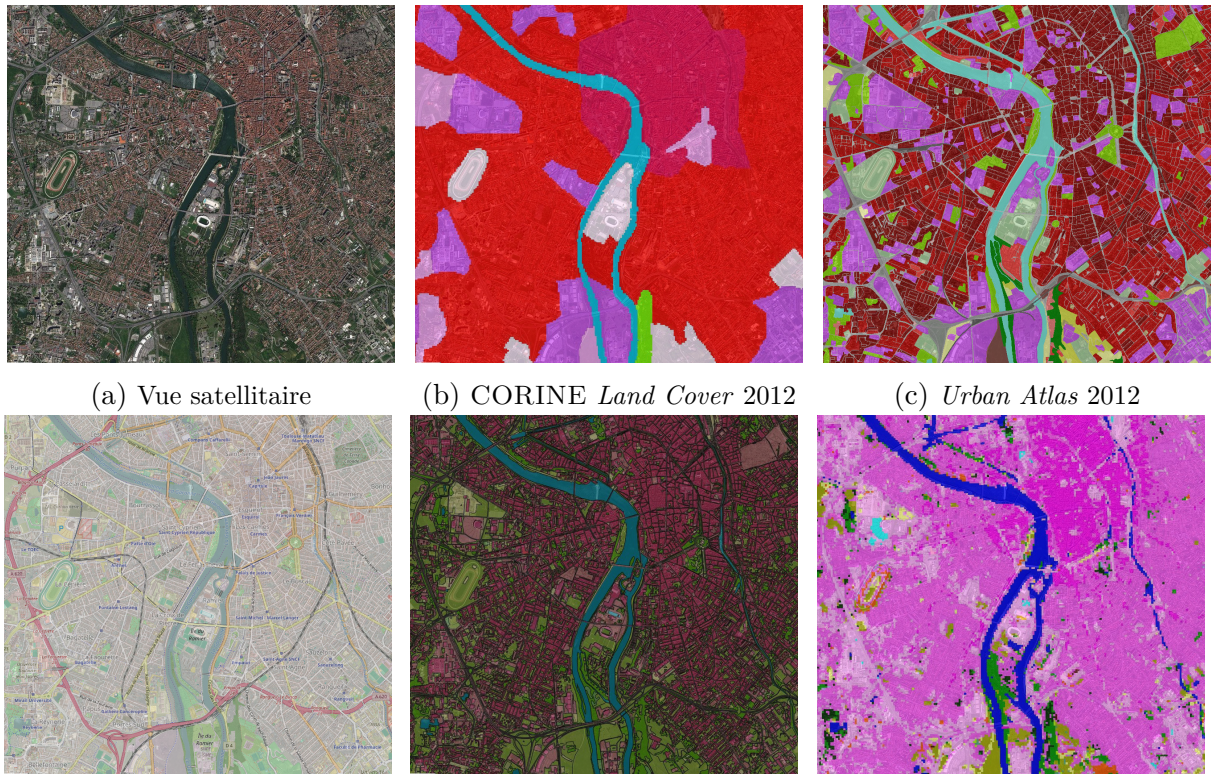


Source : osr-cesbio.ups-tlse.fr/~oso/posts/2017-04-13-carte-s2-2016-corse/
Légende : Annexe A, Figure A.3

FIGURE 1.1 – Exemple de carte d'occupation des sols.

1. une couverture qui représente la surface du territoire cartographié,
2. la résolution thématique définie par
   — une nomenclature donnée par la légende qui indique les classes d'occupation des sols cartographiées,
   — la précision sémantique représentant la conformité des occupations des sols attribuées aux objets par rapport à la réalité du terrain,
3. la résolution spatiale définie par
   — une Unité Minimale de Collecte (UMC) qui correspond à la taille minimale des objets représentés dans une carte sous format vectoriel [5] ou une résolution spatiale définie par la taille des pixels pour une carte sous format matriciel,
   — la précision géométrique qui définit le degré d'accord sur la position de l'objet dans la carte par rapport à sa position sur le terrain,
4. la temporalité qui inclut
   — l'année de référence qui correspond à l'année représentée par la carte (« millésime »),
   — la mise à jour qui correspond à la durée nécessaire pour actualiser l'ensemble de la carte.
   — le temps de production qui peut être assimilé à la date de diffusion.

5. La notion d'échelle est aussi utilisée pour une carte sous format papier. Elle est définie comme le rapport entre la taille des objets observés sur la carte par rapport à la taille réelle de ces objets. Une échelle de 1 : 5000 signifie qu'un centimètre sur la carte représente cinquante mètres en réalité.



(a) Vue satellitaire (b) CORINE *Land Cover* 2012 (c) *Urban Atlas* 2012

(d) *OpenStreetMap* (e) OCcupation des Sols à Grande Échelle 2013 (f) [Inglada et al., 2017]

Source : amcarto.alwaysdata.net/demo/test/compa_ocs.html
Légende : Annexe A, Figure A.3

FIGURE 1.2 – Exemple de cartes d'occupation des sols, Toulouse, France.

La Figure 1.2 montre des cartes d'occupation des sols présentant différentes caractéristiques sur la ville de Toulouse, dont une vue satellitaire est donnée par la Figure 1.2a.

La première différence visible concerne la nomenclature dont le code couleur est donné par légende fournie dans l'Annexe A, Figure A.3. La nomenclature de la carte CORINE *Land Cover* (CLC) (Figure 1.2b) est la moins détaillée, *i.e.* celle avec le moins de classes d'occupation des sols, tandis que celle de l'*Urban Atlas* (Figure 1.2c) est la plus détaillée en zone urbaine. Une nomenclature comme celle de CLC (Figure 1.2b) mélange les occupations des sols, *i.e.* caractéristiques biophysiques, avec les usages des sols (*land use* en anglais), *i.e.* la fonction ou le type d'usage que l'homme fait du territoire. Ainsi, les pelouses et les prairies naturelles sont décrites par une même occupation des sols « surfaces enherbées ». Pourtant, les usages sont différents : les pelouses peuvent servir pour le loisir dans le cadre de terrains sportifs, tandis que les prairies naturelles peuvent accueillir des animaux pour le pâturage et l'élevage.

Par ailleurs, les tailles des objets vus sont très différentes pour les cartes CLC, *Urban Atlas* et OCS-GE (Figures 1.2b, 1.2c et 1.2e respectivement). Par exemple, les routes ne sont pas visibles sur la carte CLC (Figure 1.2b), tandis que toute la structure de la ville est visible avec la carte *Urban Atlas* (Figure 1.2e). Ces différences s'expliquent par l'utilisation d'UMC différentes. Ainsi la carte CLC (Figure 1.2b) a la plus grande UMC à 25 hectares, tandis que la carte OCS-GE (Figure 1.2e) a une UMC de 0,05 hectare [6]. Au

6. 0,25 hectare en milieu rural

niveau intermédiaire, la carte d'*Urban Atlas* (Figure 1.2c) a une UMC de 0,25 hectare[7]. La carte CES OSO (Figure 1.2f) est ici sous format matriciel avec des pixels de 30 mètres × 30 mètres. Celle dérivée au format vectoriel a une UMC de 0,1 hectare.

De manière générale, un compromis est nécessaire entre la taille des objets cartographiés et les coûts de production : les cartes pour lesquelles les objets sont grossiers sont moins coûteuses. Autrement dit, elles nécessitent moins de temps à produire à budget identique. C'est pourquoi, les cartes à petites UMC sont produites généralement sur de petites étendues comme les communes.

En fonction des applications, il n'est pas toujours nécessaire d'avoir un détail très fin des objets qui couvrent de grandes étendues. Par exemple, un plan d'urbanisme dans une commune nécessitera une petite UMC pour observer individuellement chaque bâtiment, tandis que l'évaluation quantitative de l'étalement urbain en France nécessitera des cartes couvrant des surfaces plus grandes mais avec des éléments plus grossiers. Concernant les études sur le changement climatique, le GCOS[8] recommande la production de trois cartes [GCOS, 2016]. Deux cartes à moyenne et haute résolution spatiale décriront l'occupation des sols, tandis que la troisième cartographiera les changements d'occupation des sols. L'ensemble de ces cartes sera utile pour le suivi des changements d'occupation des sols, la gestion des ressources, ainsi que pour les études sur la modélisation du changement climatique. Le Tableau 1.1 détaille leurs caractéristiques en termes de mise à jour et de résolutions spatiale et temporelle.

Tableau 1.1 – Recommandation du *Global Climate Observing System* (GCOS) pour l'observation des occupations des sols dans le cadre des études sur le changement climatique.

Produits	OCS MR	OCS HR	Changements OCS
Mise à jour	annuelle	5 ans	1 à 10 ans
Résolution spatiale	250 m.	10 à 30 m.	250 m. à 1 km.
Incertitudes sémantiques autorisées (omission et commission par classe)	15 %	5 %	20 %
Incertitudes géométriques autorisées	83 m.	3 à 10 m.	83 à 333 m.

OCS : occupation des sols ; MR : moyenne résolution ; HR : haute résolution
Source : GCOS [2016]

1.2 Télédétection spatiale

Initialement utilisés pour la reconnaissance militaire, les satellites sont dorénavant indispensables dans de nombreux domaines comme la géodésie, la météorologie ou encore l'étude du climat. Les milliers de satellites gravitant autour de la Terre jouent un rôle majeur dans notre vie quotidienne : les prévisions météorologiques, la télévision ou encore la localisation sur terre avec le *Global Positioning System* (GPS).

Cette partie s'intéresse en particulier aux satellites imageurs qui permettent l'acquisition d'images décrivant les surfaces émergées. Une première partie introduit le concept de la télédétection, une deuxième partie retrace l'évolution des capteurs optiques permettant l'acquisition de ces séries temporelles, et une dernière partie décrit l'intérêt des séries temporelles d'images satellitaires pour la caractérisation de l'occupation des sols.

7. 1 hectare en milieu rural
8. *Global Climate Observing System* : programme qui identifie les Variables Climatiques Essentielles (VCE) (page 4)

1.2.1 Généralités sur les satellites imageurs

La télédétection est l'ensemble des techniques utilisées pour l'observation, l'analyse et l'interprétation de phénomènes ou d'objets, *e.g.* les surfaces terrestres. Plus précisément, l'énergie du rayonnement électromagnétique, émis ou réfléchi par les objets, est mesurée à distance sans contact matériel, puis est analysée.

En télédétection spatiale, deux grandes familles de capteurs existent : 1) les capteurs actifs, et 2) les capteurs passifs. Les capteurs actifs produisent leur propre source d'énergie pour illuminer la cible. Autrement dit, ils dégagent un rayonnement électromagnétique dirigé vers la cible, puis ils mesurent la réponse de la cible. Ils sont basés par exemple sur la technique du *Radio Detection And Ranging* (RADAR) ou encore du *Light Detection And Ranging* (LiDAR). Au contraire, les capteurs passifs utilisent les rayonnements naturels du Soleil comme source d'énergie. Une fois l'énergie diffusée par la cible, les capteurs passifs mesurent la radiation transmise. Dans ces travaux de thèse, seules les images issues de capteurs passifs sont étudiées.

Les images optiques sont composées de pixels dont les valeurs représentent la réflectance mesurée par le capteur du satellite. Cette dernière représente la quantité de lumière réfléchie par la surface observée. De manière générale, chaque image est caractérisée par :

— sa taille. Elle dépend de l'orbite et de l'ouverture angulaire de l'instrument, *i.e.* la fauchée représentant la surface imagée en une seule acquisition par le satellite.

— la taille du pixel appelée ici par abus de langage résolution spatiale. La résolution spatiale est en réalité définie par la distance minimale entre deux objets adjacents pouvant être distingués. Elle dépend des caractéristiques des détecteurs et de la lentille du capteur optique. Bien que pour une majorité des satellites la résolution spatiale peut être confondue avec la taille des pixels, ce n'est pas le cas pour certains satellites. Par exemple, les images des satellites Pléiades ont un résolution spatiale de 70 centimètres, mais elles sont distribuées ré-échantillonnées à 50 centimètres.

— son nombre de bandes spectrales appelé ici par abus de langage résolution spectrale. Plus précisément, la résolution spectrale est la capacité du capteur à discriminer les signaux de différentes longueurs d'ondes. Elle dépend donc du nombre de bandes, mais aussi de la largeur de ces bandes.

Outre les caractéristiques liées à chaque image – taille de l'image, des pixels et nombre de bandes –, le temps de revisite du satellite est aussi une caractéristique importante. Il représente la durée nécessaire au satellite pour imager à nouveau la même scène.

Le temps de revisite dépend principalement de l'orbite et de la fauchée du satellite. D'un côté, les satellites avec une orbite géostationnaire n'observent qu'une seule région. Les principaux satellites avec une telle orbite sont ceux de télécommunications, et aussi quelques satellites météorologiques. D'un autre côté, les satellites avec une orbite basse, et souvent héliosynchrone, peuvent cartographier l'ensemble des surfaces émergées en observant chaque région toujours à la même heure.

Pour les satellites qui acquièrent des images à angle constant, le temps de revisite est confondu avec le cycle orbital. Pour les autres satellites, le temps de revisite est généralement inférieur au cycle orbital.

Une série temporelle d'images satellitaires est obtenue lors de l'acquisition d'une même scène à différentes dates. Elle est caractérisée par sa résolution temporelle, *i.e.* le nombre d'images qui la compose et l'écart entre les différentes acquisitions.

Malheureusement, il est difficile pour les capteurs d'observer à la fois des scènes à hautes résolutions spatiale et spectrale sur de grandes étendues avec un fort temps de

revisite. En plus d'être contraints par les orbites des satellites et les capacités d'enregistrement, les capteurs sont soumis a plusieurs compromis entre la résolution spectrale et le rapport signal sur bruit, entre la résolution spatiale et le volume de données à stocker, entre la résolution spatiale et la résolution spectrale, et entre la résolution spatiale et la résolution temporelle. La partie suivante met en avant les capacités des capteurs satellitaires actuels.

1.2.2 Évolution des capteurs vers la haute résolution

Cette partie retrace l'évolution des capteurs satellitaires optiques permettant l'observation des surfaces émergées. Elle permet de souligner l'amélioration des différentes caractéristiques vues précédemment

Les premiers capteurs permettant l'acquisition de séries temporelles d'images satellitaires sur de grandes étendues étaient principalement dédiés à la météorologie. Par exemple, le capteur *Advanced Very High Resolution Radiometer* (AVHRR), qui équipe les satellites *National Oceanic and Atmospheric Administration* (NOAA) depuis les années 1970, permet l'acquisition d'images à une résolution kilométrique sur un champ d'observation très large (environ 3000 kilomètres) pour six longueurs d'ondes [9]. Toutes les surfaces sont vues au moins une fois par jour.

Au cours des années 2000, la résolution spatiale des capteurs à forte revisite temporelle atteint 250 à 300 mètres, comme pour les capteurs *Moderate Resolution Imaging Spectroradiometer* (MODIS) des satellites américains Terra et Aqua, et *Medium Resolution Imaging Spectrometer* (MERIS) du satellite européen *ENVIronment SATellite* (ENVISAT). Plus précisément, le capteur MODIS fournit une image par jour de l'ensemble des surfaces émergées dans trente-six bandes spectrales à des résolutions allant de 250 mètres au kilomètre. La haute répétitivité des acquisitions ainsi que la couverture globale de ces images font de MERIS et MODIS des capteurs de choix pour l'étude des surfaces émergées. Cependant, la basse résolution spatiale de ces capteurs ne permet pas des études sur l'occupation des sols avec des nomenclatures détaillées, par exemple les parcelles agricoles ne peuvent être discriminées.

Parallèlement, les capteurs à très haute résolution spatiale – SPOT-6 -7, Pléiades, Quickbird, Ikonos ou encore WorldView – et à très haute résolution spectrale (dit capteur hyper-spectral) se sont aussi développés. Par exemple, les capteurs à très haute résolution spatiale sont particulièrement adaptés pour la surveillance de sites sensibles, la cartographie en trois dimensions de zones urbaines ou encore la surveillance de zones vulnérables aux aléas géophysiques. Cependant, les petites fauchées ainsi que les temps de revisite espacés de ces capteurs ne permettent pas les études sur de grandes étendues.

Ainsi, des projets dédiés à l'observation de la Terre à travers des satellites à haute résolution spatiale et moyenne résolution temporelle se sont aussi développés. Le premier est le programme civil Landsat lancé par la *National Aeronautics and Space Administration* (NASA) en 1972. Initialement consacré à l'évaluation des récoltes céréalières aux États-Unis et dans l'ex-URSS (Union des Républiques Socialistes Soviétiques), ce programme permet désormais l'étude de l'ensemble des surfaces continentales. Au total huit satellites ont été lancés entre 1972 et 2013 dont deux sont encore en orbite – Landsat-7 et -8. Le neuvième satellite de la constellation devrait être lancé vers 2020. Le capteur *Operationbal Land Imager* (OLI), qui équipe Landsat-8, fournit des images à une résolution de 30 mètres pour huit bandes spectrales avec un temps de revisite de seize jours [10].

9. Capteur AVHRR/3.
10. Une bande panchromatique à 15 mètres de résolution est aussi disponible

La France, en collaboration avec la Belgique et la Suède, s'est aussi dotée en 1978 d'un programme d'observation de la Terre : Satellites Pour l'Observation de la Terre (SPOT). Les images des trois premiers satellites comptaient trois bandes spectrales à une résolution de 20 mètres, et couvraient des zones de 3600 km.^{2}. Les satellites SPOT-4 et -5 fournissaient des images pour quatre bandes spectrales à 10 et 20 mètres de résolution respectivement. Lors de leurs missions de fin de vie, les deux satellites SPOT-4 et -5 ont subi une légère baisse de leur orbite afin de permettre des acquisitions tous les cinq jours sur plus d'une centaine de sites : expérience *Take-5* [Hagolle et al., 2015b].

Au niveau européen, l'*European Space Agency* (ESA) est depuis 2008 chargée de développer et livrer les satellites Sentinel pour répondre à une partie des besoins du programme européen de surveillance de la Terre Copernicus [11] [Drusch et al., 2012; Torres et al., 2012]. Plus spécifiquement, les acquisitions récentes des satellites Sentinel-2 comptent treize bandes spectrales dans le visible et l'infra-rouge à une résolution de 10 ou 20 mètres sur une fauchée de 290 kilomètres.

La nouveauté apportée par Sentinel-2 repose sur la combinaison de ces hautes résolutions spectrale et spatiale avec une forte revisite temporelle – les deux satellites Sentinel-2 couvrent la totalité des terres émergées tous les cinq jours [12] [Malenovskỳ et al., 2012]. Ces données sont donc un atout pour la cartographie de l'occupation des sols de l'ensemble des terres émergées, notamment pour les classes de végétation qui nécessitent un suivi temporel régulier mais pour lesquelles la résolution spatiale supérieure à 100 mètres des capteurs comme MODIS était inadaptés [13].

Par ailleurs, l'usage des séries temporelles comme Landsat, SPOT (*World Heritage*) et Sentinel s'est démocratisé puisque l'ensemble de ces données sont désormais mises à disposition du public gratuitement par l'*United States Geological Survey* (USGS), le Centre National d'Études Spatiales (CNES) et l'ESA respectivement.

1.2.3 Séries temporelles d'images satellitaires

L'acquisition fréquente d'images satellitaires sur l'ensemble des surfaces continentales est essentielle dans de nombreux domaines comme la surveillance des océans et des glaciers [Berthier et al., 2010], le suivi et la gestion des territoires [Weber and Puissant, 2003] ou encore l'étude de la déforestation [Achard et al., 2002].

La Figure 1.3 montre un exemple de série temporelle d'images satellitaires et ses caractéristiques associées. La résolution spatiale couplée à la surface imagée définit la taille des images, tandis que le nombre de bandes indique la profondeur des images. Finalement, la résolution temporelle permet de visualiser plusieurs fois la même scène.

Cette partie décrit l'intérêt des séries temporelles d'images satellitaires à hautes résolutions. Afin de mieux comprendre l'importance de la combinaison des hautes résolutions spatiale et spectrale avec des acquisitions fréquentes, chacune des caractéristiques est illustrée dans la suite.

11. Le programme européen de surveillance de la Terre Copernicus est l'ex-programme *Global Monitoring for Environment and Security* (GMES).

12. Actuellement les deux satellites Sentinel-2 observent la totalité de l'Europe et de l'Afrique tous les cinq jours. Le reste du globe est vu tous les dix jours en attendant l'installation d'une nouvelle station de réception.

13. Le satellite taïwanais Formosat-2 fournissait jusqu'en 2016 des données similaires à Sentinel-2 mais sur une fauchée de 24 kilomètres. Avec un temps de revisite à un jour, les images acquises par Venμs depuis août 2017 devrait permettre de mesurer l'impact de la très haute revisite temporelle couplée à la haute résolution spatiale. Cependant, Venμs est une mission scientifique de deux ans qui ne permet pas d'assurer la couverture de l'ensemble des terres émergées avec sa fauchée de 27 kilomètres.

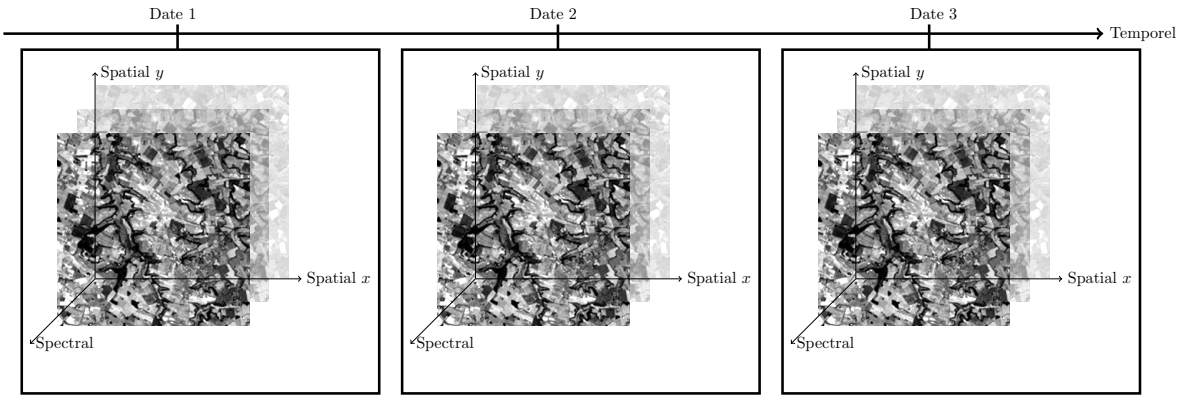
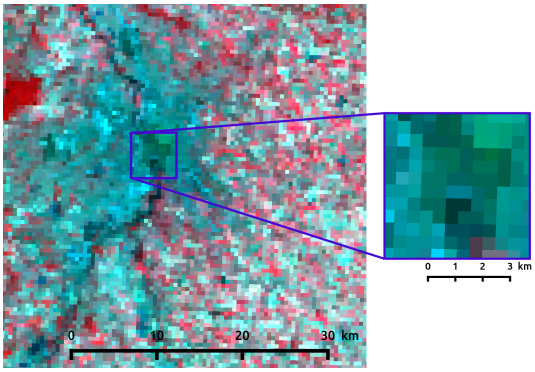
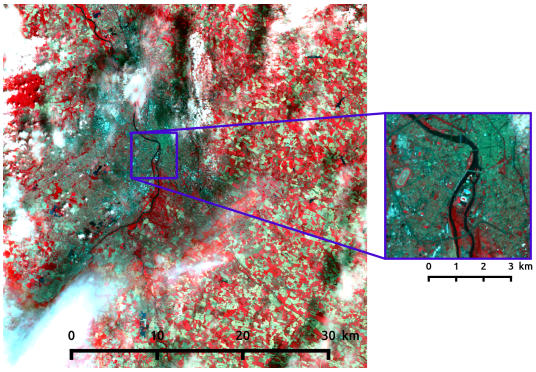


FIGURE 1.3 – Caractéristiques des séries temporelles d'images satellitaires.

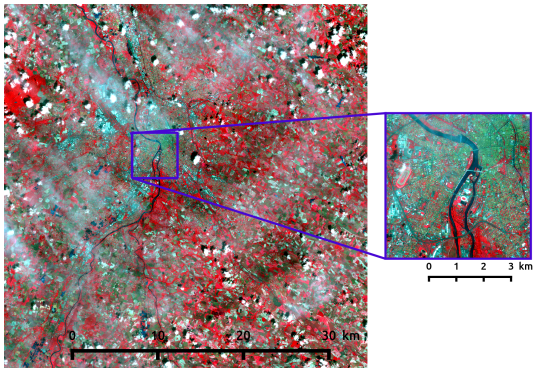
Tout d'abord, la Figure 1.4 montre l'importance de la résolution spatiale dans le contexte de la cartographie de l'occupation des sols. L'exemple montre trois acquisitions sur Toulouse en août 2016 par différents capteurs satellitaires. Il permet d'illustrer la notion de résolution spatiale. Les pixels d'une image à 500 mètres de résolution spatiale (Figure 1.4a) mélangent nécessairement plusieurs occupations des sols très différentes. Ces images sont donc inadaptées pour cartographier individuellement des objets de petites tailles comme les bâtiments, les routes ou des fleuves. Les images à 10 et 30 mètres (Figures 1.4b et 1.4c) de résolution spatiale permettent de visualiser des objets de taille intermédiaire comme des parcelles agricoles ou des milieux forestiers.



(a) 500 mètres (capteur MODIS)



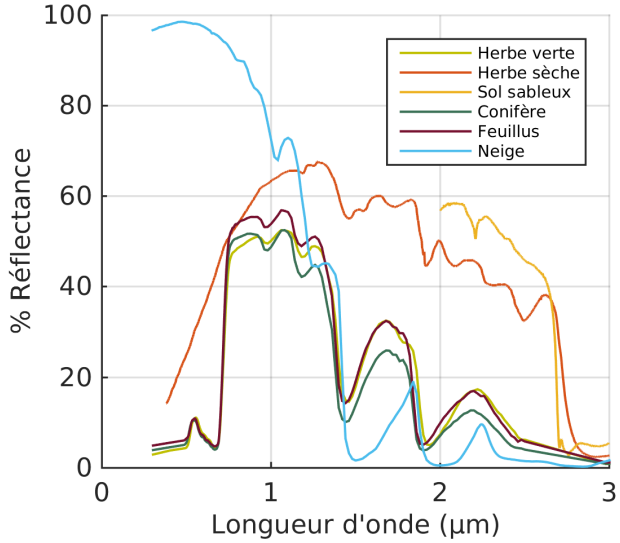
(b) 30 mètres (Landsat-8)



(c) 10 mètres (Sentinel-2)

FIGURE 1.4 – Illustration de la notion de résolution spatiale. Images en fausse couleur (proche infra-rouge, rouge et vert).

Concernant la résolution spectrale, les acquisitions dans différents domaines de longueurs d'onde facilitent la caractérisation des occupations des sols puisque chaque matériau a une réponse spectrale spécifique. La Figure 1.5 montre des profils spectraux pour six occupations des sols extraits de la base de données *Advanced Spaceborne Thermal Emission and Reflection Radiometer* (ASTER) [Baldridge et al., 2009]. Cet exemple montre notamment que les profils spectraux du conifère et de l'herbe verte sont très similaires pour des longueurs d'onde inférieures à 1.5 μm. En s'appuyant seulement sur cette information, les deux occupations des sols seraient donc confondues. Par contre, une acquisition dans le domaine du moyen infra-rouge (entre 1.5 μm. et 1.8 μm.) permettrait, pour cet exemple, de discriminer le conifère des autres occupations des sols.



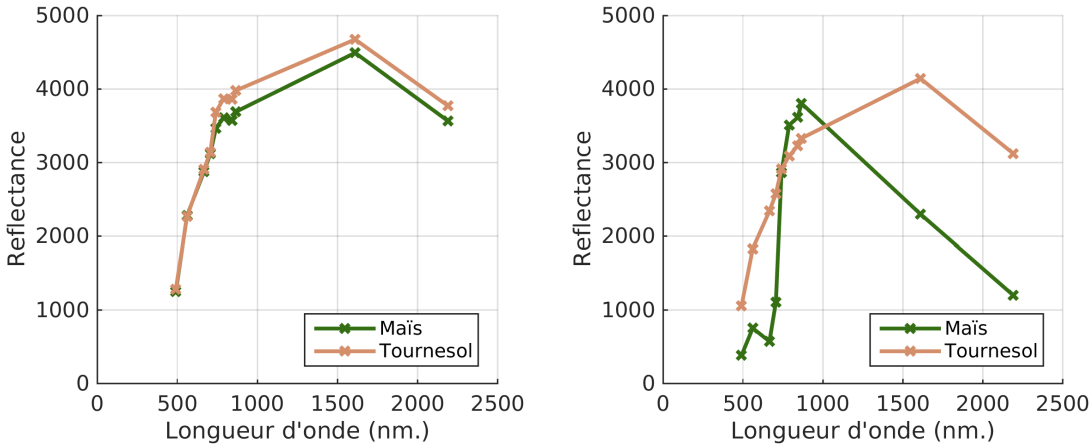
Source : ASTER Spectral Library – speclib.jpl.nasa.gov

FIGURE 1.5 – Exemple de profils spectraux pour six occupations des sols.

La Figure 1.5 montre que les profils de végétation – herbes sèches et vertes, conifères et feuillus – présentent une faible valeur de réflectance dans le domaine du visible suivi par un pic de réflectance dans le proche infra-rouge, aux alentours de 0.7 μm. En effet, les pigments dans les feuilles des plantes sont connus pour absorber la lumière du visible, tandis qu'une structure dense des plantes est connue pour refléter fortement la lumière infra-rouge. Plus la plante a de feuilles, plus le pic est visible. Cette propriété est régulièrement utilisée pour définir des indices de végétation, dont le plus connu est le *Normalized Difference Vegetation Index* (NDVI) [14], qui permettent d'accentuer les différences entre les classes de végétation et les autres occupations des sols.

L'information fournie par les bandes spectrales à une seule date peut être insuffisante pour caractériser les occupations des sols dont le comportement spectral évolue au cours du temps. Par exemple, la Figure 1.6 montre les profils spectraux d'échantillons de maïs et de tournesol acquis par un capteur à haute résolution spectrale (aussi dit multi-spectral) à sept bandes à deux dates différentes. Une acquisition de mi-septembre (Figure 1.6b) correspondrait à deux profils spectraux qui se distinguent clairement : la réponse du maïs est plus faible que celle du tournesol dans le moyen infra-rouge. Par contre, une acquisition fin mai (Figure 1.6a) montre deux profils spectraux quasiment superposables : les deux cultures sont en phase de levée.

14. L'indice *Normalized Difference Vegetation Index* (NDVI) est défini dans la Section 4.1.2.



(a) 27/05/2016 (b) 14/09/2016

FIGURE 1.6 – Profils spectraux de deux échantillons, maïs et tournesol, à deux dates différentes.

Ainsi, l'acquisition de séries temporelles d'images satellitaires, *i.e.* une même scène observée à différentes dates, permet de décrire le comportement temporel des différentes classes. Le cycle phénologique[15] des plantes étant différent, sa visualisation permet de discriminer différentes plantes. À titre d'exemple, les profils temporels de NDVI, *i.e.* profils pour lesquels l'indice de NDVI est calculé à chaque acquisition, sont affichés pour trois cultures d'hiver (blé, orge et colza) et deux d'été (maïs et tournesol) sur la Figure 1.7. Au début de la croissance de la plante, la valeur de NDVI augmente jusqu'à une valeur maximale, qui correspond à la floraison de la plante. Le profil de NDVI finit par décroître lorsque de la phase de sénescence des plantes. Ainsi, des acquisitions en mai (DoY $\sim$ 125) et en septembre (DoY $\sim$ 250) permettront de différencier les cultures d'hiver et d'été respectivement.

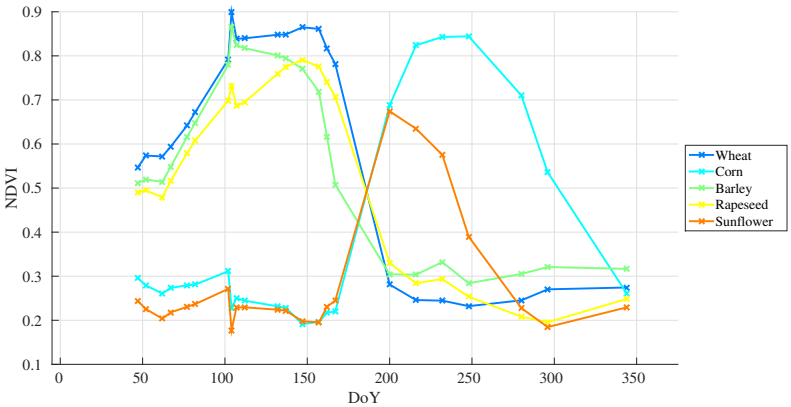
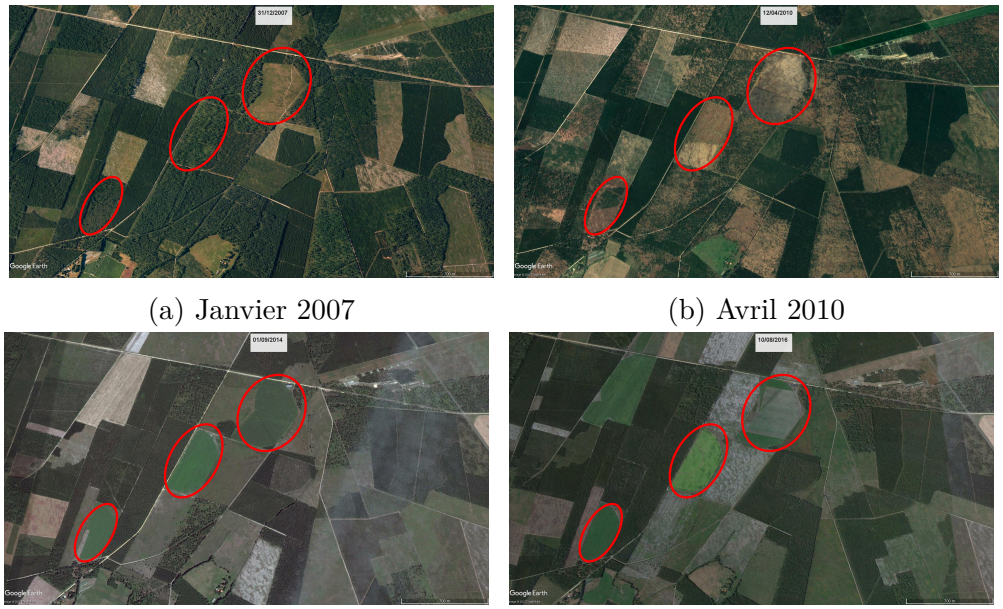


FIGURE 1.7 – Exemple de profils *Normalized Difference Vegetation Index* (NDVI) en fonction du jour de l'année (*Day of Year* (DoY) en anglais) pour cinq occupations des sols (légende en anglais).

Outre la caractérisation fine des cycles phénologiques, l'acquisition de séries temporelles sur le long terme permet de suivre l'évolution des territoires et de détecter les changements d'occupation du sol. La Figure 1.8 offre des vues satellitaires de la forêt landaise entre 2007 et 2016. Sur ces images, il est possible d'observer l'amenuisement de

15. La phénologie d'une plante est la description des événements majeurs qui décrivent l'évolution de la plante – la floraison, la feuillaison, la sénescence, *etc.*.



(a) Janvier 2007 (b) Avril 2010

(c) Septembre 2014 (d) Août 2016

Source : Google Earth www.google.com/earth

FIGURE 1.8 – Évolution de la forêt des Landes de Gasgogne, France, entre 2007 et 2016. Les zones en rouge correspondent à des changements d'occupation des sols.

zones forestières au profit d'exploitations agricoles. Dès 2010, des zones forestières ont disparu, principalement suite au passage de la tempête Xyntia en 2009. Puis à partir de septembre 2014, l'apparition de champs de maïs (zones semis-circulaires cerclées en rouge) est visible.

La partie suivante détaille comment les séries temporelles d'images satellitaires sont utilisées pour l'obtention des cartes d'occupation des sols.

1.3 Des images vers la carte

Comme vu précédemment, l'acquisition de séries temporelles sur de grandes étendues par les satellites imageurs permet l'étude de l'occupation des sols à une échelle globale. Dans cette partie, les méthodes mises en place pour produire des cartes d'occupation des sols à partir d'images satellitaires sont décrites. L'objectif n'est pas de proposer une liste exhaustive, mais plutôt de mettre en perspective les différentes approches de production.

Afin de produire une carte d'occupation des sols, deux grandes stratégies sont possibles. La première consiste à mettre à jour d'anciennes cartes d'occupation des sols en identifiant les zones de changement ou les zones incomplètes, tandis que la seconde consiste à générer une nouvelle carte. La stratégie de mise à jour est moins courante. En effet, elle est attrayante sur le plan opérationnel et économique lorsque seules les zones de changement sont à analyser. Cependant, la reconnaissance de ces zones à mettre à jour est une question complexe, particulièrement dans des régions très hétérogènes avec des dynamiques temporelles fortes. Actuellement, les méthodes de détection de changement identifient bien souvent trop de faux positifs, *i.e.* des zones à mettre à jour alors qu'aucun changement n'a eu lieu [Gressin, 2014]. La généralisation de telles méthodes pour la cartographie de grandes étendues est donc difficile, et a peu été exploitée.

À titre d'exemple la base de données OCS-GE produite par l'IGN prend le parti de fusionner des données existantes. Des traitements sont ensuite appliqués afin de respecter

les spécifications de l'OCS-GE. Une description détaillées de cette base de données sera présentée dans la Section 3.3.2.

Quelque soit la stratégie choisie (mise à jour ou nouvelle carte), la production peut être réalisée soit manuellement, soit automatiquement. Ces deux types d'approches sont détaillés dans la suite avec une définition plus précise, des exemples de cartes ainsi que les avantages et inconvénients. Les cartes présentées sont décrites en termes de nomenclature (*e.g.* nombre de classes d'occupation des sols), de résolution spatiale, de surface couverte, de temps de production et de mise à jour.

1.3.1 Approches manuelles

Les approches manuelles désignent ici toutes les approches nécessitant une forte intervention humaine pour la production des cartes. Trois sous-groupes d'approches sont distingués :

1. La photo-interprétation. Des opérateurs externes, généralement experts en Système d'Information Géographique (SIG) ou du terrain d'étude, identifient les occupations des sols à l'aide d'images aériennes et satellitaires.
2. Les enquêtes terrain. Des experts référencent les occupations des sols en allant sur le terrain.
3. La production collaborative (*crowd-sourcing* en anglais). Des personnes bénévoles – pas nécessairement expertes en géomatique, en cartographie ou du terrain d'étude – contribuent à la cartographie de toutes les surfaces émergées à travers notamment les plate-formes Internet SIG dédiées à la cartographie.

La photo-interprétation est entièrement basée sur l'utilisation des images satellitaires, tandis que les enquêtes terrain ou les approches collaboratives utilisent souvent l'image satellitaire en support avec d'autres technologies notamment le *web-mapping* et le GPS.

Photo-interprétation

Parmi les cartes produites uniquement par photo-interprétation, CLC est celle qui couvre le plus grand territoire : elle s'étend sur 39 pays européens. Cet inventaire est produit à partir d'images satellitaires d'une résolution de 20 à 25 mètres dans le cadre du programme Copernicus et de la directive européenne *Infrastructure for Spatial Information in the European Community* (INSPIRE), piloté par l'*European Environment Agency* (EEA). La nomenclature utilisée est hiérarchique et emboîtée [16] sur trois niveaux, et décrit un total de 44 classes. Au total quatre versions sont disponibles – 1990, 2000, 2006 et 2012 [Büttner, 2014]. La Figure 1.9 montre la dernière version 2012 pour l'ensemble de l'Europe, et la Figure 1.2b est un zoom sur la ville de Toulouse. Des cartes de changement d'occupation des sols entre versions sont aussi disponibles. Elles permettent des analyses simplifiées de l'évolution des territoires européens.

Le premier niveau de la nomenclature CLC est aussi utilisé pour la production des cinq couches thématiques hautes résolutions, *High Resolution Layers* (HRL) : surfaces imperméables, prairies, forêt, surfaces en eau et zones humides. Les HRL sont produites par une approche de mise à jour, *i.e.* par fusion d'anciennes bases de données, combinée avec la photo-interprétation et aussi une extraction automatique d'information (approches

16. Une nomenclature hiérarchique emboîtée est une nomenclature qui s'étend sur plusieurs niveaux. Par exemple, un classe surface en eau pourra être décrite plus finement en deux sous classes comme eaux continentales et eaux maritimes.

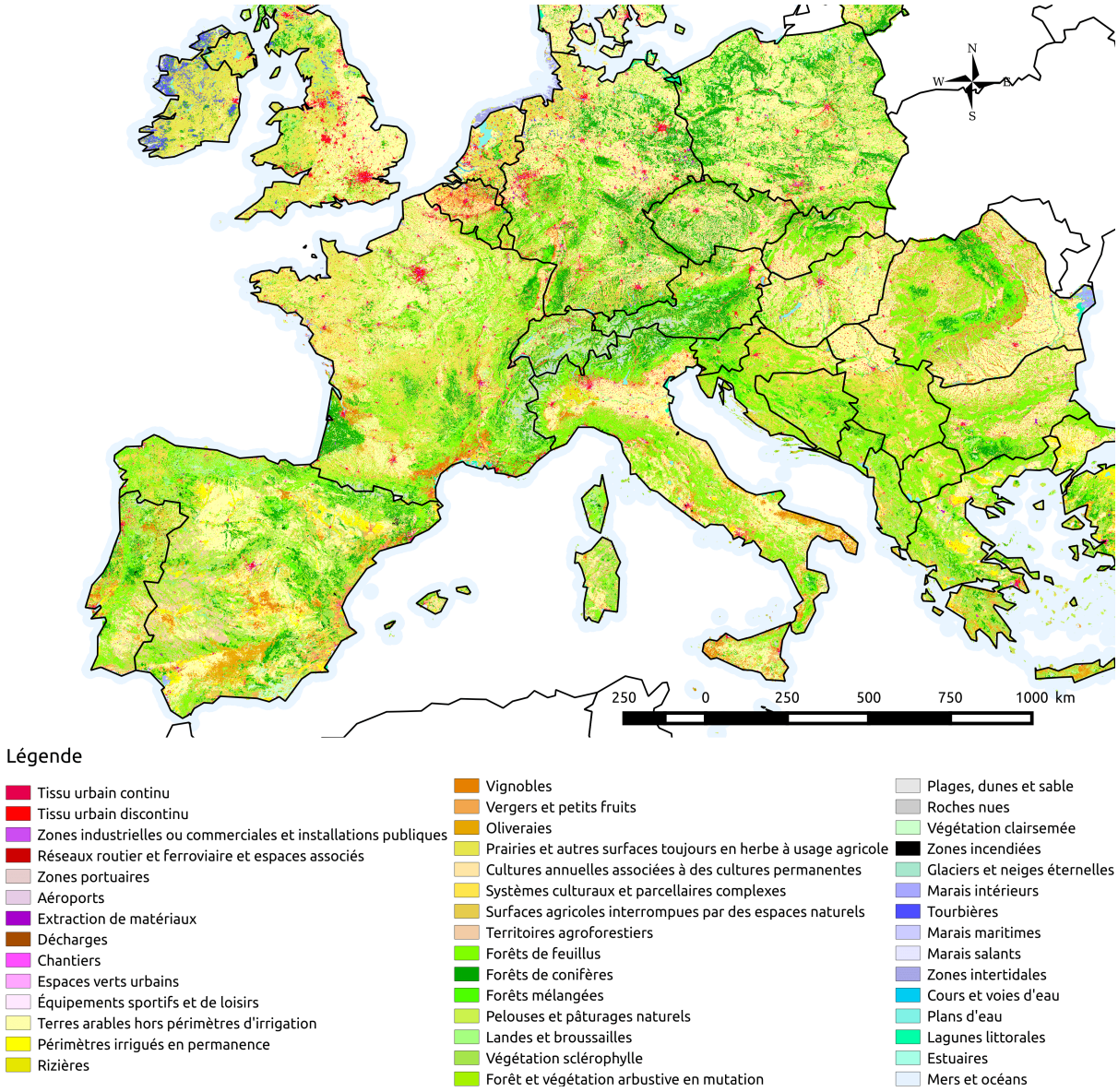


FIGURE 1.9 – CORINE *Land Cover* 2012

mixtes). La Figure 1.10 montre les couches de forêts (en vert) et de surfaces imperméables (en rouge) à l'ouest de l'Europe. Produite similairement, les données *Urban Atlas* fournies aussi par l'EEA décrivent l'occupation des sols pour 305 métropoles européennes de plus de 100 000 habitants et leurs alentours. Un exemple est donné par la Figure 1.2c.

L'UMC de 25 hectares fait de CLC une base de données [17] idéale pour des statistiques nationales, mais inadaptées à des échelles plus fines comme la commune. Ainsi, des organismes régionaux ont produit leurs propres cartes d'occupation des sols par photo-interprétation. Les projets CIGAL, SIGALE, OCSOL PACA et Mode d'Occupation des Sols (MOS) proposent des cartes d'occupation des sols régionales en Alsace, Nord-Pas-de-Calais, Provence-Alpes-Côte d'Azur (PACA) et en Île de France, respectivement. Par exemple, CIGAL se base sur des images SPOT-5 et des ortho-photographies de la BD Ortho de l'IGN pour obtenir une carte avec une UMC de 0.5 hectare décrivant 55 classes

17. Le terme base de données fait ici référence à la base de données géographiques pour laquelle il est possible de dériver une carte d'occupation des sols, *i.e.* une prise de vue de la base de données. Par exemple, la base de données BD Topo de l'IGN contient plusieurs couches thématiques pour lesquelles il est possible de construire différentes cartes d'occupation des sols en fonction de ces besoins.

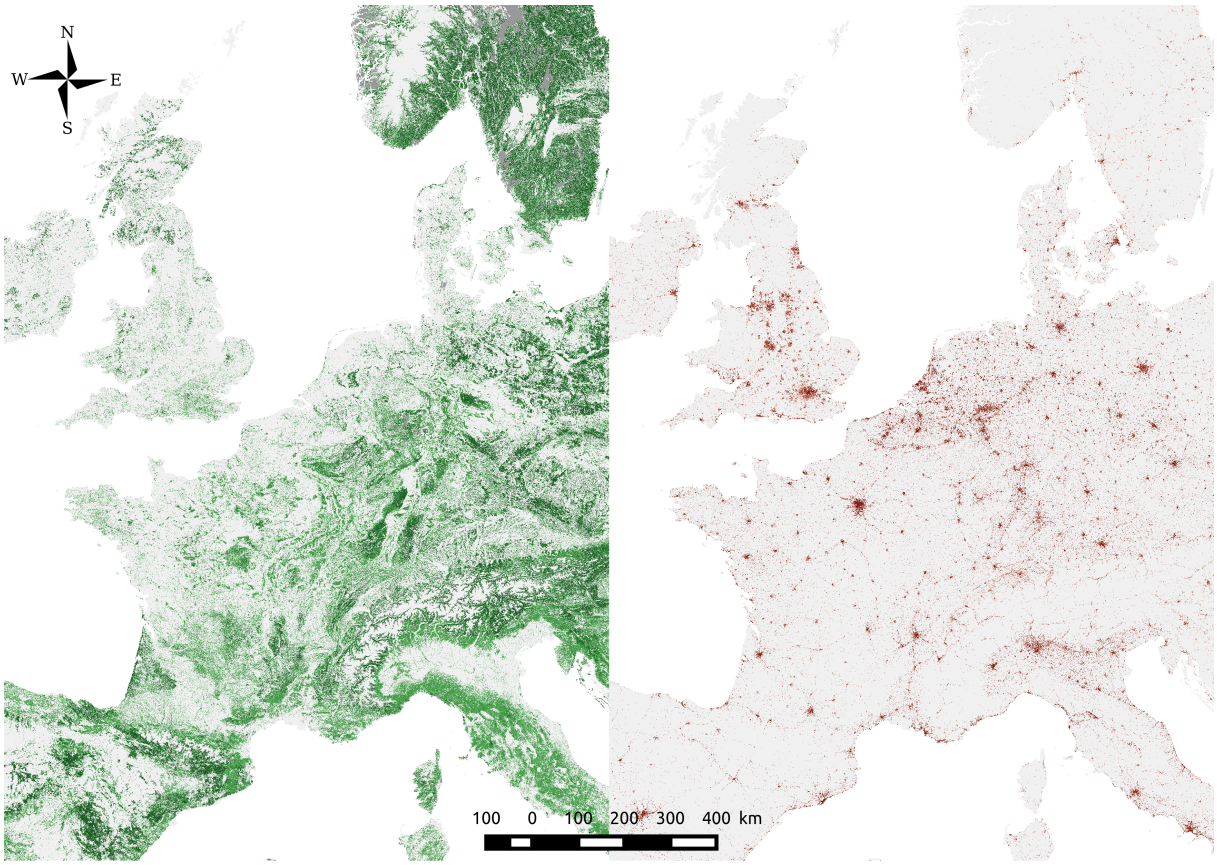


FIGURE 1.10 – *High Resolution Layers* 2012. Couche forêts à gauche, couche surfaces imperméables à droite.

thématiques.

La génération de ces cartes par photo-interprétation a pour principal inconvénient des temps de production et de mise à jour très longs. Par exemple, la version 2012 de CLC a été mise à disposition seulement en 2015, et le temps de mise à jour est de six années entre les dernières versions. De plus, ces données sont coûteuses à produire, et l'expérience inégale entre les opérateurs engendre des hétérogénéités spatiales dans le carte produite. Finalement, la qualité des cartes produites repose essentiellement sur les images satellitaires utilisées pour la photo-interprétation. Les opérateurs n'ayant pas le temps d'analyser l'ensemble des images acquises aux différentes dates, leur décision repose bien souvent sur seulement quelques images dont le contenu peut être insuffisant pour discriminer les classes d'occupation des sols comme les cultures. Par ailleurs, la nomenclature CLC mélange occupation et usage des sols. Ce qui peut poser problème à certains utilisateurs qui ont besoin d'une information seulement sur l'occupation du sol. Ces données sont tout de même de très bonne qualité grâce aux procédures rigoureuses d'identification des occupation des sols.

Enquête terrain

Une autre approche manuelle pour produire des cartes sur de grandes étendues est d'effectuer une interpolation sur des relevés ponctuels issus d'enquêtes terrain. Par exemple, pour Teruti-Lucas – géré en France par Agreste, le service statistique du ministère de l'Agriculture –, les enquêteurs identifient l'occupation des sols sur deux niveaux : 1) le segment, *i.e.* une portion du territoire homogène d'environ 1.5 km.^{2}, et 2) le point, *i.e.*

un cercle de 3 ou 40 mètres de diamètre. L'image satellitaire, plus spécifiquement des ortho-photographies, est seulement utilisée pour le géo-référencement.

Outre les problèmes d'exhaustivité, les relevés terrain nécessitent souvent plusieurs passages des enquêteurs notamment pour la végétation. Par exemple, les cultures d'été et les cultures d'hiver ne peuvent pas être discriminées à la même période. Ainsi l'enquêteur doit effectuer trois à quatre passages pour être rigoureux, ce qui augmente le coût et le temps de production.

Crowd-sourcing

Le dernier type d'approche manuelle consiste à faire intervenir des contributeurs non-experts en cartographie. En effet, la démocratisation des SIG en ligne et des GPS ont conduit à l'émergence de « la cartographie 2.0 ». Créé en 2004, *OpenStreetMap* (OSM) [18] a pour objectif de produire une carte mondiale librement modifiable et accessible. C'est le projet d'information géographique bénévole (*volunteered geographic information* en anglais) le plus abouti actuellement (Figure 1.2d). Tous les citoyens peuvent contribuer en utilisant des images aériennes – assez rares, car les images ne doivent pas être sous Copyright –, des traces GPS ou des panoramas à 360 degrés acquis par exemple par la camera *OpenStreetCam* [19] [Neis and Zielstra, 2014]. Sur un principe similaire, WikiMapia [20] propose aux internautes de cartographier les sols à partir de Google Maps depuis 2006. L'utilisation de plate-formes collaboratives permet aussi de remonter dans le temps. Par exemple, le projet national GeoHistoricalData [21] cartographie des sources historiques comme la carte de Cassini datant de la fin du XVIIIe siècle.

Le *crowd-sourcing* a aussi été utilisé afin d'améliorer des cartes existantes. La carte GeoWiki a été en partie réalisée à l'aide d'outils collaboratifs pour mettre à jour les cartes GLC 2000, MCD12Q1 2005 et GlobCover 2005 produites automatiquement (et décrites dans la partie suivante) [Fritz et al., 2009, 2012; See et al., 2015].

Les avantages de ces données collaboratives sont le faible coût de production [22] et la possibilité de remonter dans le temps. Cependant, les personnes non-expertes en SIG et du terrain d'étude sont généralement moins précises et rigoureuses dans leurs déclarations, ce qui conduit à de nombreuses incohérences et confusions dans les données répertoriées. Des contributeurs bénévoles en désaccord fournissent alors des informations contradictoires [Johnson and Iizuka, 2016]. Ainsi, les précisions des données peuvent être faibles : 76 % pour le Portugal [Estima and Painho, 2013] et 64 % pour Hambourg [Arsanjani et al., 2015] pour OSM. Par ailleurs, rien ne garantit la complétion de ces données puisque les contributeurs choisissent les zones qu'ils cartographient.

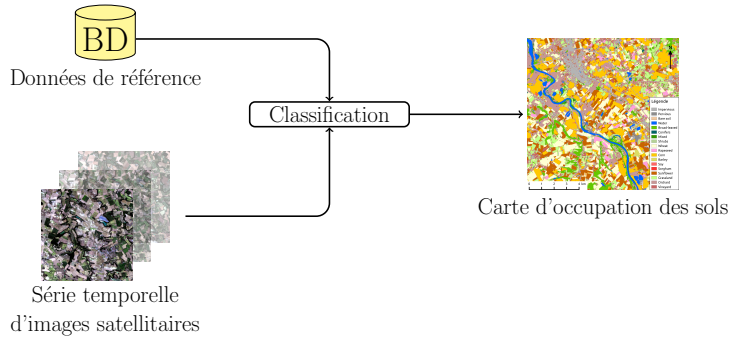
De manière générale, toutes les approches manuelles sont dépendantes des opérateurs, des experts et des contributeurs qui construisent la carte. Ainsi, la reproductibilité et la continuité des produits ne sont pas assurées, les coûts de production sont élevés, et surtout les temps de mise à jour sont longs.

18. openstreetmap.org
19. openstreetcam.org
20. wikimapia.org
21. geohistoricaldata.herokuapp.com
22. Le coût de production s'entend ici d'un point de vu économique. Il est bien entendu que le coût investi par les différents contributeurs bénévoles est important, même s'il ne fait pas l'objet d'une transaction financière.

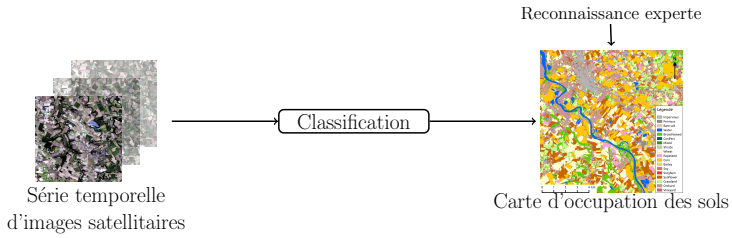
1.3.2 Approches automatiques

Contrairement aux approches manuelles, les approches automatiques visent à extraire l'occupation des sols à partir d'images principalement aéroportées et satellitaires en limitant l'intervention humaine. Pour ce faire, des méthodes de classification sont utilisées : elles consistent à associer chaque pixel de l'image à une classe d'occupation des sols. Ces méthodes sont basées sur le principe de l'apprentissage automatique (*machine learning* en anglais). Traditionnellement, deux approches sont distinguées : supervisée et non-supervisée.

La Figure 1.11 montre la différence entre ses deux approches qui réside principalement dans l'utilisation faite de la donnée de référence. Dans le cas d'un apprentissage supervisé, la donnée de référence donne une connaissance *a priori* sur l'occupation des sols. Elle permet d'obtenir des échantillons d'apprentissage pour lesquels l'occupation des sols est connue, et qui seront ensuite utilisés par l'algorithme de classification supervisée. Ainsi, les échantillons d'apprentissage jouent un rôle essentiel dans la précision du résultat final : ils doivent représenter précisément chacune des classes et caractériser la variabilité de la zone étudiée. De plus, la donnée de référence doit dater de la même année que les images à classifier pour éviter des erreurs liées aux changements d'occupation des sols entre les années, *e.g.* la rotation des cultures ou l'expansion de zones péri-urbaines. Dans le cas d'un apprentissage non-supervisé, les pixels de l'image sont regroupés par similarité (techniques de *clustering*), et les classes d'occupation des sols sont déterminées *a posteriori*.



(a) Apprentissage supervisé



(b) Apprentissage non-supervisé

FIGURE 1.11 – Schémas simplifiés des apprentissages supervisé et non-supervisé pour la cartographie de l'occupation des sols.

Une des premières bases de données mondiales d'occupation des sols produite automatiquement est la donnée *Data and Information System Cover* (DISCover) issue des données *Global Land Cover Characterization* (GLCC) 2.0 [23]. Elle décrit 17 classes définies par l'*International Geosphere-Biosphere Program* (IGBP) à une résolution spatiale

23. lta.cr.usgs.gov/glcc/globdoc2_0

kilométrique [Belward et al., 1999; Loveland et al., 2000, 2009; Scepan, 1999]. Elle est obtenue par un apprentissage non-supervisé sur des images NOAA acquises par le capteur AVHRR entre 1992 et 1993. La carte de l'Université du Maryland UMD est produite par un apprentissage supervisé pour les mêmes images mais avec une nomenclature IGBP simplifiée à 12 classes [Hansen et al., 2000] La carte *Global Land Cover* (GLC) 2000 de la *Food and Agriculture Organization of the United Nations* (FAO), aussi à un kilomètre de résolution, décrit 22 classes définies par le *Land Cover Classification System* (LCCS) [Bartholomé and Belward, 2005]. La carte est obtenue par apprentissage non-supervisé à partir d'images acquises par le capteur VEGETATION du satellite SPOT-4.

L'arrivée des capteurs MERIS et MODIS a conduit aux développements de cartes à une résolution spatiale entre 250 et 500 mètres. Par exemple, les cartes *Collection 5 MODIS Land Cover Type Product* (MCD12Q1), produites annuellement depuis 2001 à partir de méthodes supervisées, décrivent 17 classes IGBP à 500 mètres de résolution spatiale [Friedl et al., 2010]. Les cartes GlobCover, GLCNMO et GLC250 aussi produites dans les années 2000 présentent des caractéristiques similaires [Arino et al., 2007; Bontemps et al., 2015; Defourny et al., 2006; Tateishi et al., 2014; Wang et al., 2015]. Les produits *Climate Change Initiative - Land Cover* (CCI-LC) 2000, 2005 et 2010 de l'ESA à 300 mètres de résolution fournissent en plus des cartes de la couverture des sols, une estimation sur les changements d'occupation des sols à une résolution kilométrique en s'appuyant sur des images acquises par le capteur AVHRR.

Depuis la mise à disposition gratuite des images Landsat-7 et -8, des cartes d'occupation des sols mondiales à une résolution de 30 mètres ont été produites. Par exemple, les cartes FROM-GLC et GlobeLand30, produites par les organismes chinois *Centre for Earth System Science China* (CESSC) et *National Administration of Surveying, Mapping and Geoinformation* (NASG), décrivent une dizaine de classes [Chen et al., 2015; Gong et al., 2013; Yu et al., 2013a,b, 2014].

La mise en orbite des deux satellites Sentinel-2 doit permettre la production de cartes à 10 mètres de résolution. Pour le moment, une carte de la France 2016 a été produite au Centre d'Études Spatiales de la BIOsphère (CESBIO) et mise à disposition au premier trimestre 2017 [24] (Figure 1.1). L'approche entièrement automatique s'appuie sur l'apprentissage supervisé d'une série temporelle d'images Sentinel-2 acquise en 2016. Elle est équivalente à celle utilisée pour la création de la carte 2013 produite à 30 mètres à l'aide d'images Landsat-8 (Figure 1.2f) [Inglada et al., 2017].

La description de ces différentes cartes montre que les approches non-supervisées ont été délaissées au profit des méthodes supervisées [Franklin and Wulder, 2002; Khatami et al., 2016]. Les deux principales raisons sont la mise à disposition de données facilitant la collection d'échantillons d'apprentissage et la meilleure précision des algorithmes supervisés [Grekousis et al., 2015].

Dans un contexte opérationnel, la majorité des méthodes utilise de manière inadéquate ou sous-optimale l'information temporelle contenue dans les séries d'images satellitaires. Par exemple, certains travaux sélectionnent des dates clés représentant des stades phénologiques discriminatifs. Afin de discriminer au mieux les cultures d'hiver des cultures d'été, il est courant de sélectionner des images peu nuageuses acquises à deux saisons différentes [Rodríguez-Galiano et al., 2012; Rogan et al., 2002]. Ces dates clés fournissent généralement des images comprenant de fortes différences dans les signatures spectrales de la végétation. Cependant, cette sélection subjective peut conduire à une faible précision globale (64.89 % par exemple pour FROM-GLC [Gong et al., 2013]). En outre, cette sélection de dates clés, généralement coûteuse en temps de calcul, est une étape complexe.

24. osr-cesbio.ups-tlse.fr/~oso

D'une part, les acquisitions (non-nuageuses) d'images satellitaires ne sont pas assurées aux dates clés. D'autre part, le changement climatique ainsi que l'activité anthropique modifient les occupations des sols, et peuvent donc modifier les dates clés d'une année sur l'autre.

Par ailleurs, les traitements sur de grandes étendues sont réalisés pour des raisons pratiques par tuile, *i.e.* une zone d'étude est découpée en zones rectangulaires sur lesquelles une série de traitement identique est appliquée. Dans le cadre d'une sélection de date clés par tuile, les traitements peuvent conduire à la présence d'artefacts et de démarcations à la jonction entre les tuiles. Cet effet de bord diminue la qualité du résultat final. Pour toutes ces raisons, la sélection de dates clés n'est pas adaptée à la mise en place d'une chaîne de traitement automatique.

Une autre approche couramment utilisée consiste à combiner des images sur plusieurs années sans prendre en compte les évolutions du paysage. Par exemple, les cartes FROM-GLC et GlobeLand30 utilisent des images Landsat-5 -7 et -8 acquises sur une quinzaine d'années (1984 à 2011). Sur une telle période, l'urbanisation, l'expansion des zones agricoles ou encore la déforestation modifient fortement les paysages. D'une part, l'interprétation de la carte produite est difficile car il est impossible de déterminer à quelle année appartiennent les résultats. D'autre part, les échantillons d'apprentissage utilisés doivent être caractérisés par une seule occupation des sols qu'il est difficile de définir si les occupations des sols changent d'une année sur l'autre.

Pour obtenir des données de référence de qualité, les collectes sont généralement effectuées par photo-interprétation. Comme cette collecte est longue, la mise à jour fréquente des cartes est difficile. Seule la donnée MCD12Q1 fournit des cartes annuellement en photo-interprétant régulièrement des images MODIS pour 1860 sites stratifiés par régions éco-climatiques sur l'ensemble du globe [Friedl et al., 2010]. La basse résolution spatiale de 250 mètres des images MODIS permet d'avoir un faible nombre de zones à photo-interpréter, et donc une mise à jour annuelle qu'il ne serait pas possible de maintenir avec des images à une résolution spatiale de 10 mètres. Cependant, plusieurs occupations des sols sont présentes dans des pixels de 250 mètres $\times$ 250 mètres.

Pour les cartes produites à 30 mètres, comme FROM-GLC et GlobaLand30, des profils temporels MODIS à 500 mètres en complément des images Landsat-8 sont aussi utilisés pour la photo-interprétation des échantillons d'apprentissage. L'utilisation de profils temporels à basse résolution pour l'étape de photo-interprétation conduit à diminuer la résolution effective de la carte finale [Broxton et al., 2014].

De plus, l'utilisation de la photo-interprétation limite bien souvent la quantité de données de référence extraite. Or, la classification sur de grandes étendues nécessite un nombre conséquent d'échantillons bien réparti spatialement pour assurer une bonne représentation de la zone à cartographier. Ainsi, Inglada et al. [2017] utilisent d'anciennes cartes comme données de référence. Cependant, l'utilisation de données obsolètes conduit à des incertitudes sur l'information fournie à l'algorithme de classification, notamment pour les classes de végétation qui évoluent annuellement, qui potentiellement diminue la précision de la carte produite.

1.4 Position du problème

La Section 1.3 a montré le potentiel des méthodes de classification supervisée pour l'obtention de cartes d'occupation des sols à partir de séries temporelles d'images satellitaires à hautes résolutions. Cependant, les méthodes proposées sont loin d'être opérationnelles et de nombreux défis restent à relever. Ce chapitre décrit les problématiques et les ob-

jectifs spécifiquement abordés dans cette thèse. L'organisation du manuscrit y est aussi détaillée.

1.4.1 Défis

Afin d'obtenir des cartes d'occupation des sols précises avec des temps de production et de mise à jour réduits – par exemple avec une mise à jour annuelle comme recommandée par le GCOS (Tableau 1.1) –, l'utilisation de méthodes d'apprentissage supervisé est incontournable (Section 1.3.2). Cependant, le traitement des nouvelles séries temporelles d'images satellitaires à hautes résolutions, comme celles fournies par Sentinel-2, implique la gestion de grands volumes de données jamais étudiés auparavant. Ainsi, la mise en place d'une chaîne de traitement automatique est très complexe, et nécessite de répondre a plusieurs défis.

La Figure 1.12 détaille le processus de classification supervisée de la Figure 1.11a. L'étape d'apprentissage supervisé s'appuie sur des échantillons d'apprentissage décrits par l'information extraite des images satellitaires et l'occupation des sols extraite des données de référence. L'objectif de cette étape est de construire un modèle de classification à partir des échantillons d'apprentissage. Le modèle sera ensuite utilisé pour déterminer la classe d'occupation des sols de chaque pixel des images satellitaires. Ainsi, la qualité de la carte produite dépend principalement de la qualité des données en entrée (satellitaires et de référence), et du choix de l'algorithme de classification.

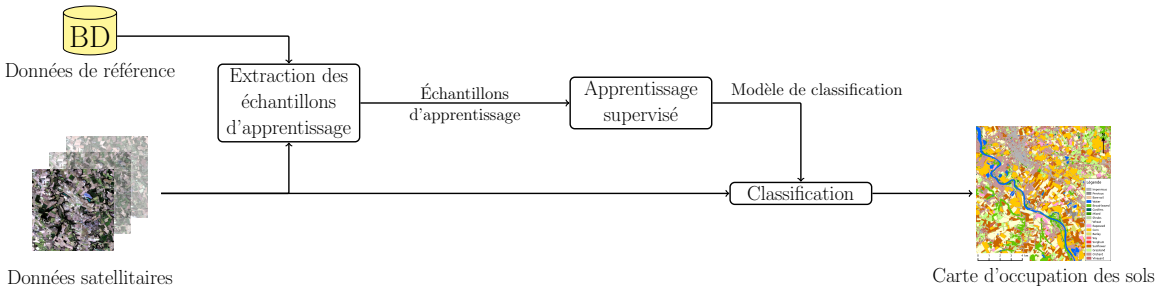


FIGURE 1.12 – Processus de classification supervisée.

Concernant les algorithmes de classification, le développement de l'intelligence artificielle a permis plusieurs évolutions : les méthodes sont rapides à mettre en œuvre, adaptées pour prendre en compte un grand nombre d'échantillons et efficaces pour travailler dans des espaces de grande dimension [25]. Ces méthodes de classification ont été appliquées avec succès pour la cartographie de l'occupation des sols [Khatami et al., 2016], mais rarement en utilisant la totalité de l'information fournie par les nouvelles séries temporelles d'images satellitaires. Les algorithmes de classification actuels peuvent donc souffrir des variabilités spatiales et temporelles de ces données.

Ces nouvelles données satellitaires questionnent alors les choix :

1. de l'algorithme de classification,
2. des données à fournir en entrée du système de classification pour
   — exploiter correctement l'information spectro-temporelle des séries temporelles,
   — prendre en compte la variabilité des paysages sur de grandes étendues.

25. La dimension d'un problème de classification est donnée par le nombre de variables qui décrit les échantillons.

En effet, le choix de l'algorithme de classification ainsi que son paramétrage sont essentiels. De plus, les performances des algorithmes vont fortement dépendre des données satellitaires utilisées. Comme vu dans la Section 1.2.3, la combinaison des hautes résolutions spatiale, spectrale et temporelle induit une description fine des occupations des sols, particulièrement pour la végétation. Plus spécifiquement, la présence de plusieurs bandes spectrales et de plusieurs dates d'acquisition interroge le choix des données satellitaires fournies en entrée du système de classification. De plus, la stabilité des algorithmes peut se détériorer à cause de la variabilité des paysages induite par la taille des superficies analysées lors de la classification sur de grandes étendues. Si les échantillons d'apprentissage sont spatialement restreints, l'ajout d'une information supplémentaire aux bandes spectrales pourrait aider à limiter les effets climatiques, topographiques et atmosphériques.

Outre le choix des données satellitaires à fournir en entrée du système de classification, l'importance des données de référence pour la production automatique des cartes le plus tôt possible dans l'année a été montrée à la Section 1.3.2. Cependant, la collecte de ces données sur de grandes étendues est difficile.

Une première possibilité pour obtenir un jeu de données de qualité est de réaliser des enquêtes terrain ou de la photo-interprétation sur des images aériennes ou à très haute résolution spatiale acquises sur la même période temporelle que les images à classifier. Ces données coûteuses sont en plus longues à produire, alors que les données satellitaires sont traitées en quasi-temps réel [26]. Et même si les procédures de vérification et de production assurent des données de référence de grande qualité, ces données sont susceptibles de contenir des erreurs dues à l'humain ou l'informatique. Par ailleurs, ces procédures sont difficilement applicables sur de grandes étendues.

Une seconde stratégie consiste à utiliser des bases de données déjà existantes et produites sur de grandes étendues. Par exemple, l'information de cartes d'occupation des sols des années précédentes, de données institutionnelles, gouvernementales ou encore de données collaboratives peut être utilisée. Ces données couvrant généralement de grandes surfaces permettent l'extraction d'un grand nombre d'échantillons d'apprentissage. Cependant, l'évolution constante des paysages – étalement urbain, rotation des cultures, ou encore déforestation – conduit à la présence d'erreurs lorsque ces données plus ou moins datées sont utilisées pour la classification de séries temporelles de l'année en cours.

Bien que ces problèmes soient connus, l'utilisation d'échantillons imparfaits a rarement été pris en compte dans le domaine de la télédétection. Ainsi, l'influence de ces imperfections sur les performances de classification est inconnue. Et aucune stratégie adaptée à la grande dimension des séries temporelles d'images satellitaires, visant à prendre en compte ces données imparfaites, n'existe.

1.4.2 Objectifs

L'objectif général de la thèse vise à améliorer la production des cartes d'occupation des sols à partir des nouvelles séries temporelles d'images satellitaires comme celles fournies par les capteurs Sentinel-2.

Le premier objectif consiste à étudier le choix du classifieur en lien avec les données satellitaires, les données de référence et la surface de la zone d'étude. Plus spécifiquement, le choix du classifieur est discuté ainsi que les données à fournir en entrée du système de classification. De plus, ces choix sont testés lors de la classification sur de grandes étendues.

26. Une image Sentinel-2 corrigée des effets atmosphériques (Section 3.2) est par exemple produite en moins de 48 heures par le pôle de données et de services surfaces continentales Theia.

Finalement, l'influence des échantillons d'apprentissage erronés sur les performances de l'algorithme de classification est quantifiée.

Le second objectif consiste à proposer un cadre méthodologique pour prendre en compte les données de référence imparfaites dans le processus de classification. En particulier, ces travaux s'intéressent aux cas où la donnée de référence est soit ancienne par rapport à l'acquisition des données satellitaires soit contaminée par des échantillons erronés. Dans ce contexte, une méthodologie est proposée pour filtrer les données erronées avant l'étape d'apprentissage supervisé. Les données imparfaites sont tout d'abord identifiées, puis traitées afin de réduire leur impact sur le processus de classification.

1.4.3 Organisation du manuscrit

Selon les objectifs présentés dans la Section 1.4.2, ce manuscrit est divisé en cinq parties :

Partie I. Cette première partie décrit donc le contexte des travaux de thèse. Après avoir présenté brièvement les enjeux autour de la cartographie de l'occupation des sols, l'apport des séries temporelles pour la classification de l'occupation des sols est présenté. Puis, les approches de production des cartes d'occupation des sols existantes sont décrites. Enfin, les défis, les objectifs et l'organisation du manuscrit sont détaillés.

Partie II. La deuxième partie présente les méthodes et les données utilisées. Le Chapitre 2 est dédié aux méthodes de classification, notamment celles utilisées pour la classification de séries temporelles. Une description technique des algorithmes de classification utilisés dans ce manuscrit est aussi fournie. Le Chapitre 3 présente les données satellitaires et les données de référence utilisées au cours de ces travaux. Les prétraitements appliqués sur l'ensemble des données sont aussi décrits.

Partie III. Dans la troisième partie, la stabilité et la robustesse des algorithmes de classification pour la cartographie de l'occupation des sols sur de grandes étendues sont étudiées. Le Chapitre 4 est dédié aux problématiques concernant le choix de l'algorithme de classification et des données à fournir en entrée du système de classification. Au cours de ce chapitre, différentes expérimentations sont réalisées notamment pour attester de la stabilité des algorithmes de classification lors de la cartographie sur de grandes étendues. Le Chapitre 5 s'intéresse quant à lui à la robustesse des algorithmes de classification lorsque des données de référence imparfaites sont utilisées. Plusieurs configurations de classification – nombre de classes, type de données, nombre d'échantillons – sont testées afin d'évaluer quantitativement l'influence de ces données imparfaites sur les performances de classification.

Partie IV. La quatrième partie est consacrée à la prise en compte des données de référence imparfaites dans le processus de classification. Le Chapitre 6 s'intéresse en particulier à la détection de ces données imparfaites. Une méthode de détection est proposée et comparée avec les méthodes de l'état-de-l'art. Puis, le Chapitre 7 propose un cadre méthodologique afin de prendre en compte les données imparfaites dans le processus de classification. L'impact du processus proposé sur les performances de la classification est évalué.

Partie V. La dernière partie contient le Chapitre 8 présentant la conclusion générale. Ce dernier permet de résumer les principaux résultats du manuscrit, et de souligner les conclusions les plus importantes. Les perspectives méthodologiques et applicatives sont aussi discutées.

Deuxième partie

Méthodes et données

Chapitre 2

Classification supervisée de séries temporelles d'images satellitaires

Sommaire

2.1 Introduction à l'apprentissage supervisé 28
2.2 Classification supervisée de séries temporelles 30
2.3 Choix des algorithmes de classification 32
2.4 Algorithmes de classification supervisée 34
2.4.1 *Support Vector Machine* 34
2.4.2 *Random Forest* 40
2.5 Évaluation des performances des algorithmes de classification 46
2.5.1 Évaluation d'une classification multi-classes 47
2.5.2 Évaluation statistique 49

La production automatique de cartes d'occupation des sols à partir d'images satellitaires est principalement basée sur des méthodes de classification. L'objectif d'une méthode de classificiation est de construire un modèle capable de prédire pour chaque pixel de l'image une classe, aussi appelée étiquette. Issus du domaine de l'apprentissage automatique, ces algorithmes sont traditionnellement divisés en deux catégories dans la littérature : supervisée et non-supervisée [27].

Les approches non-supervisées (aussi appelées *clustering*) cherchent à regrouper les échantillons similaires au sein d'une même classe, *e.g. k-Means*, *Self-Organizing Map* (SOM) ou encore *Iterative Self-Organizing Data Analysis Technique yAy* (ISODATA). Les groupes, aussi appelés *clusters*, sont alors constitués d'échantillons similaires qui sont dissemblables des échantillons appartenant à d'autres *clusters*. Une classe est ensuite associée *a posteriori* à chaque *cluster*.

Ces approches sont favorisées quand peu de connaissances sur les types d'occupation des sols sont disponibles [Eva et al., 2004; Gong et al., 2013]. Cependant, la reconnaissance de *clusters* est une question complexe et fastidieuse qui ne peut être réalisée que par un expert de la zone d'étude. Bien souvent, des post-traitements – fusion ou division de *clusters* – sont nécessaires avant de pouvoir étiqueter les *clusters*, et les faire coïncider avec la nomenclature [Loveland et al., 2000]. Par ailleurs, des pré-traitements sont aussi nécessaires afin d'éviter que les classes avec de fortes variances dominent les *clusters*. Par

27. Le cas semi-supervisé n'est pas abordé dans ce manuscrit.

exemple, une classe de culture à forte variabilité (induite par les pratiques agricoles, la qualité des sols ou encore les conditions climatiques) peut présenter différentes apparences (plusieurs comportements), et donc être présente dans plusieurs *clusters*. Une même classe aura alors plusieurs étiquettes. De plus, les méthodes de *clustering* sont coûteuses en temps et en ressources informatiques lorsque la taille des images – nombre de pixels et de bandes – augmente. Pour toutes ces raisons, les approches supervisées sont généralement favorisées dans le contexte de la cartographie sur de grandes étendues [Khatami et al., 2016].

Ce chapitre se focalise sur les approches supervisées. Une première partie introduit quelques généralités sur la classification supervisée. Une deuxième partie s'intéresse à l'utilisation des approches supervisées pour la classification de séries temporelles d'images satellitaires dans la littérature. Puis, une troisième partie discute le choix des algorithmes de classification toujours dans le contexte de la cartographie de l'occupation des sols. Une troisième partie détaille alors le principe de fonctionnement des algorithmes utilisés pendant ces travaux. Finalement, une dernière partie décrit le processus d'évaluation des méthodes de classification, et donc des cartes produites.

2.1 Introduction à l'apprentissage supervisé

L'objectif d'un apprentissage supervisé est d'apprendre automatiquement des règles pour prédire les étiquettes de nouveaux échantillons. L'ensemble des règles est appris à partir d'exemples fournis par une donnée de référence.

Plus précisément, la Figure 2.1 détaille l'ensemble du processus de classification supervisée. Les échantillons contenus dans la donnée de référence sont divisés au cours de l'étape d'échantillonnage en deux sous-ensembles. D'une part, les échantillons d'apprentissage sont utilisés comme connaissance *a priori* sur l'occupation des sols. D'autre part, les échantillons test servent dans la phase d'évaluation, décrite dans la Section 2.5.

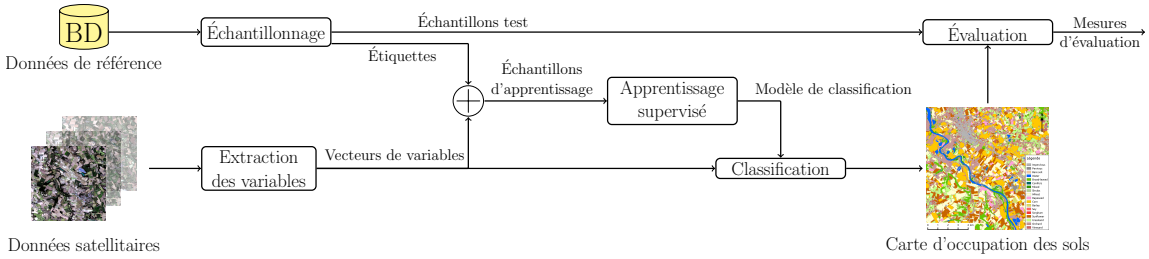


FIGURE 2.1 – Processus de classification supervisée.

L'étape centrale du processus est l'apprentissage supervisé. À partir des échantillons d'apprentissage, l'algorithme de classification apprend un modèle de classification. La règle de décision définie par le modèle permet de prédire les classes d'occupations des sols pour de nouveaux échantillons. Idéalement, le modèle de classification doit être capable de généraliser ce qu'il a appris à de nouveaux échantillons.

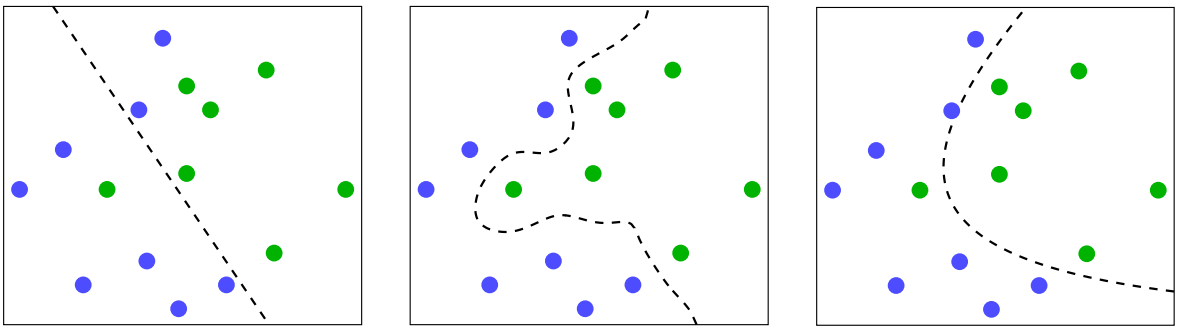
Une des principales difficultés de la phase d'apprentissage est de trouver le compromis entre un modèle trop simple et un modèle trop spécifique aux échantillons d'apprentissage. Le modèle peut décrire parfaitement les échantillons d'apprentissage, mais être incapable de prédire correctement les classes de nouveaux échantillons qui n'ont pas été utilisés pour construire le modèle.

Un modèle peu stable qui apprend par cœur les données d'apprentissage sans avoir aucune capacité de généralisation fait du sur-apprentissage (*over-fitting* en anglais). Au

contraire, un modèle trop simple est incapable de saisir les relations pertinentes entre les échantillons d'apprentissage. Il commet alors de nombreuses erreurs sur les échantillons d'apprentissage, et fait du sous-apprentissage (*under-fitting* en anglais). Dans les deux cas, le modèle construit est incapable de généraliser et de prédire l'étiquette de nouveaux échantillons.

Ce problème est aussi connu sous le nom de compromis biais-variance. Le biais d'un algorithme est caractérisé par son erreur sur l'ensemble des données d'apprentissage, tandis que la variance d'un algorithme correspond à l'écart entre l'erreur faite sur les données d'apprentissage et l'erreur faite sur les données test. Le compromis biais-variance consiste donc à trouver un équilibre entre la complexité du modèle et sa capacité à généraliser.

La Figure 2.2 représente trois fois le même problème de classification à deux classes (bleue et verte) pour lequel différentes frontières de décision, en pointillé noir, sont dessinées. La Figure 2.2a montre un modèle trop simple, pour lequel le biais est élevé, *i.e.* trois échantillons sont mal classés. Le modèle de la Figure 2.2b montre un exemple de sur-apprentissage où l'ensemble des échantillons sont parfaitement classés. Ce modèle est caractérisé par une forte variance car l'ajout ou la suppression d'un échantillon modifierait totalement la frontière de décision. Enfin, la Figure 2.2c montre un bon compromis entre le biais de l'algorithme et la variance. La frontière de décision devrait permettre une bonne généralisation pour prédire l'étiquette de nouveaux échantillons.



(a) Sous-apprentissage (b) Sur-apprentissage (c) Bon compromis

FIGURE 2.2 – Illustration des problèmes de sur-apprentissage et sous-apprentissage.

Les échantillons d'apprentissage jouent un rôle essentiel pour obtenir un bon compromis biais-variance. Comme montré à la Figure 2.1, les échantillons d'apprentissage sont décrits par les valeurs des vecteurs de variables extraits des données satellitaires. Le terme primitive est aussi utilisé pour parler de variables calculées à partir des données satellitaires. Les algorithmes de classification utilisent alors ces échantillons afin de construire leur règle de décision.

Les échantillons d'apprentissage doivent être représentatifs de la population sur laquelle le modèle sera appliqué. Ils doivent entre autre décrire les apparences multiples des classes. Par exemple, une classe surface urbaine doit idéalement décrire l'ensemble des toits apparents.

Deux paramètres importants du problème de la classification supervisée sont alors 1) n le nombre d'échantillons d'apprentissage, et 2) p la dimension du vecteur de variables. Augmenter les valeurs de n et p permet généralement d'apprendre un modèle plus complexe tout en contrôlant la variance.

Cependant, augmenter la dimension du vecteur p n'est pas toujours une bonne solution. Certaines méthodes traditionnelles, notamment les méthodes statistiques, sont mises en défaut lorsque la dimension du problème p devient trop élevée, éventuellement plus

grande que n. Ce phénomène connu sous le nom de malédiction de la dimension (*curse of dimensionality* ou *Hughes phenomenon* en anglais) fait diminuer les performances des algorithmes de classification [Hughes, 1968]. En effet, les calculs de distance, utilisés par de nombreux algorithmes de classification pour mesurer la similarité entre échantillons, sont difficiles à mener dans des espaces de grande dimension. Si le nombre d'échantillons est trop peu important devant le nombre de variables, les échantillons vont être isolés dans un espace de grande dimension, compliquant les regroupements d'échantillons similaires.

De manière générale, l'augmentation du nombre d'échantillons d'apprentissage assure donc une meilleure description des apparences des classes et limite la malédiction de la dimension. Ainsi, les performances de la classification augmentent généralement avec le nombre d'échantillons. Plusieurs études dans la littérature ont tenté de définir le nombre optimal d'échantillons nécessaires pour l'apprentissage, notamment en fonction de p. La règle des $30p$, *i.e.* utilisé au moins $30p$ échantillons d'apprentissage par classe, fait souvent référence [Foody and Mathur, 2004a].

Cependant, des études montrent qu'un nombre inférieur d'échantillons bien informatifs, *e.g.* les échantillons à la frontière de classes [Foody and Mathur, 2004b], peut aussi permettre de bien séparer les classes pour certains types d'algorithmes [Piper, 1992; Van Niel et al., 2005]. Par ailleurs, les récents travaux de Li et al. [2014] concluent qu'une soixantaine d'échantillons d'apprentissage par classe suffit à obtenir une précision importante pour une quinzaine d'algorithmes de classification. Cependant, ces études sur le nombre optimal d'échantillons nécessaires sont souvent réalisées dans des espaces de petites dimensions, *e.g.* $p = 3$ [Foody and Mathur, 2004a], bien inférieures au nombre de variables qu'il est possible d'extraire des séries temporelles d'images satellitaires.

2.2 Classification supervisée de séries temporelles

Dans le contexte de la cartographie de l'occupation des sols, de nombreuses méthodes ont été proposées pour la classification des données satellitaires. Cependant, peu d'études se focalisent sur la classification de séries temporelles d'images satellitaires optiques. Ce manque d'études s'explique par l'absence de données de référence de qualité, et par la récente disponibilité des séries temporelles optiques à haute résolution spatiale.

Pour la classification de séries temporelles, les premiers travaux ont été menés en utilisant principalement des approches paramétriques comme le maximum de vraisemblance, le *Gaussian Mixture Model* (GMM) ou encore l'analyse discriminante. Ces méthodes font une hypothèse sur la nature du modèle dont les paramètres sont estimés à l'aide des échantillons d'apprentissage. Elles ont été utilisées plusieurs fois pour la classification de séries temporelles à basse résolution spatiale (MODIS, MERIS, et AVHRR) [DeFries and Townshend, 1994; Jia et al., 2014b; Radoux et al., 2014].

La majorité des méthodes paramétriques suppose que la distribution des variables décrivant les échantillons appartenant à une même classe suit une loi normale, ce qui est rarement le cas dans le contexte de la classification de séries temporelles. Ainsi, ces approches échouent à prendre en compte les différentes apparences de certaines classes, et les variations spectro-temporelles présentes dans les séries temporelles. Les frontières obtenues entre les classes sont alors imprécises [Hubert-Moy et al., 2001].

Pour ces raisons, les méthodes non-paramétriques sont plus efficaces que les méthodes paramétriques lorsque les distributions des classes d'occupation des sols sont inconnues [Foody and Mathur, 2006]. Parmi ces approches, l'*Artificial Neural Network* (ANN) est une méthode précise [Gopal et al., 1999; Mas and Flores, 2007], mais qui requiert la configuration de nombreux paramètres (nombre de neurones et de couches). L'optimisation de ces

paramètres a tendance à conduire au sur-apprentissage (Section 2.1) [Rodríguez-Galiano et al., 2012]. Par ailleurs, le modèle construit par l'ANN est une boîte noire puisque les règles de décision apprises par le modèle sont difficiles à analyser et interpréter.

Une autre approche non-paramétrique couramment utilisée est le *Support Vector Machine* (SVM) [Mountrakis et al., 2011]. Il a été appliqué de nombreuses fois avec succès pour la classification de séries temporelles optiques [Carrão et al., 2008; Dash et al., 2007; Huang et al., 2002a; Jia et al., 2014a]. De plus, il présente de nombreux avantages dont une faible sensibilité au phénomène de Hughes [Hughes, 1968] et de bonnes performances lorsque le nombre de variables est grand devant le nombre d'échantillons d'apprentissage [Pal, 2005]. Afin de traiter des problèmes complexes, le SVM utilise souvent une fonction noyau qui nécessitera la configuration non triviale de paramètres supplémentaires [Müller et al., 2001].

Ainsi, les arbres de décision binaire plus faciles à paramétrer ont été utilisés pour la cartographie de séries temporelles Landsat [Dash et al., 2007; Gebhardt et al., 2014], et pour les cartes MCD12Q1 à partir d'images MODIS [Friedl et al., 2010]. Rapides à apprendre, ils permettent en plus d'utiliser des variables de différentes natures (réflectance, indices normalisés, *etc.*) et de prendre en compte les données manquantes (*e.g.* nuageuses) [Friedl et al., 2002]. Cependant, leurs performances sont en-deçà de celles des ANN et SVM dans les espaces à grande dimension [Hansen, 2012; Pal and Mather, 2003]. Et comme les modèles ANN, ils ont tendance à faire du sur-apprentissage [Ghimire et al., 2012].

Chaque algorithme décrit précédemment a ses propres avantages et inconvénients qui conduit à des résultats différents sur les mêmes données. Les méthodes d'ensemble cherchent alors à combiner différents algorithmes de classification pour tirer bénéfice de chacun d'entre-eux [Dietterich, 2000a]. Ces méthodes ont une variance moins importante que la variance individuelle des algorithmes de classification puisque la sensibilité aux données d'apprentissage est diluée entre tous les algorithmes. Pour autant, le biais n'est pas augmenté par rapport à celui de chaque algorithme. Ainsi, ces méthodes sont moins sensibles aux données d'entrée et plus robustes au bruit [Foody et al., 2016; Johnson and Iizuka, 2016]. De plus, les performances de classification sont améliorées [Bauer and Kohavi, 1999], notamment pour la classification de données multi-capteurs et multi-temporelles [Briem et al., 2002; Du et al., 2012; Miao et al., 2012].

Parmi les méthodes d'ensemble, le *Random Forest* (RF) construit un ensemble d'arbres de décision binaire [Breiman, 2001]. Dans le cadre de la cartographie de l'occupation des sols, le nombre de travaux utilisant le RF ne cesse d'augmenter [Gislason et al., 2006; Inglada et al., 2017; Pal, 2005; Rodríguez-Galiano et al., 2012]. Outre ses performances comparables avec d'autres algorithmes de classification [Belgiu and Drăguţ, 2016], le RF présente de nombreux avantages : un temps de calcul réduit dû à la possibilité de construire les arbres en parallèle [Lassalle et al., 2015], la possibilité de prendre en entrée des gros volumes de données (*i.e.* un nombre élevé de variables) [28], et une interprétation facilité offerte par la visualisation des arbres [29].

28. R. Genuer, J-M. Poggi, C. Tuleau-Malot, N. Villa-Vialaneix. Random Forests for Big Data. *Submitted to Big Data Research on March 2017*

29. Cette propriété est parfois remise en cause dans la littérature à cause du côté aléatoire introduit par les échantillons *bootstrap* et le principe de *random feature selection* (Section 2.4.2). De plus, un nombre élevé d'arbres complique l'analyse des règles de décision pour l'ensemble des échantillons. Il n'en reste pas moins qu'il est possible de visualiser chaque arbre de la forêt, et donc de suivre chaque échantillon.

2.3 Choix des algorithmes de classification

La section précédente a mentionné quelques algorithmes de classification utilisés dans le cadre de la classification supervisée de séries temporelles. Il existe dans la littérature plusieurs centaines d'algorithmes de classification développés pour de nombreux domaines d'application [Fernández-Delgado et al., 2014]. Chaque algorithme de classification a ses propres avantages et inconvénients, et leur performance dépend des jeux de données étudiés [Caruana and Niculescu-Mizil, 2006].

Le contexte de ces travaux est lié à la mise en place d'une chaîne de traitement automatique pour la production de cartes d'occupation des sols sur de grandes étendues à partir de séries temporelles d'images satellitaires comme celles fournies par Sentinel-2. La Section 2.2 a identifié plusieurs classifieurs déjà utilisés pour la classification de séries temporelles. Afin de sélectionner les plus appropriés, les critères suivants sont importants :

— la précision,

— le temps de calcul,

— le paramétrage,

— la stabilité,

— la robustesse.

Un des principaux inconvénients identifiés au Chapitre 1 est le temps de production des méthodes actuelles. Bien que la carte produite doit être aussi précise que possible, un compromis est nécessaire avec les temps de calcul. Ces derniers dépendent des volumes de données à traiter, mais aussi de la complexité du modèle appris par les algorithmes de classification. Il est alors important de bien connaître le comportement des algorithmes de classification afin de déterminer le point de fonctionnement optimisant le compromis entre la précision et la complexité du modèle appris pour une application donnée, ici la classification de séries temporelles d'images satellitaires sur de grandes étendues.

Dans un contexte d'automatisation des chaînes de traitement, il est aussi intéressant que les performances de l'algorithme de classification ne soient pas dépendantes d'un réglage très fin de ses paramètres. Cette propriété aura un double avantage. Le premier sera de réduire le temps dédié à l'optimisation des paramètres, et le second sera de pouvoir garder le même paramétrage pour analyser différentes zones d'études ou la même zone d'étude à différentes dates.

Un autre critère concerne la stabilité de l'algorithme de classification. L'algorithme ne doit pas trop s'adapter aux échantillons d'apprentissage, et être capable de prédire correctement l'étiquette de nouveaux échantillons (Section 2.1). Dans le cas d'études sur de grandes étendues, l'algorithme de classification doit être donc capable de gérer d'une part les fortes variabilités intra-classes, *i.e.* les classes avec différentes apparences, et d'autre part les faibles variabilités inter-classes, *i.e.* les classes très similaires.

Finalement, il est intéressant que l'algorithme de classification soit robuste à la présence de données imparfaites dans les échantillons d'apprentissage. Dans le cas où un grand nombre de données de référence est nécessaire, cela permettrait d'autoriser l'utilisation de données de référence pour lesquelles certaines étiquettes sont incertaines. Ce critère de robustesse est spécifiquement étudié au Chapitre 5.

Concernant la classification de séries temporelles d'images satellitaires, les revues récentes de Khatami et al. [2016] et Gómez et al. [2016] soulignent les bonnes performances du SVM et du RF. Ces résultats sont en accord avec les travaux de Gong et al. [2013] dont les cartes FROM-GLC les plus précises sont obtenues avec le SVM et le RF. Par

ailleurs, des études préliminaires réalisées au CESBIO montrent aussi le potentiel des deux algorithmes pour la classification de séries temporelles [Inglada et al., 2015].

Depuis quelques années, les méthodes d'apprentissage profond (*deep learning* en anglais) basées sur des réseaux neuronaux sont de plus en plus utilisées. Peu de travaux existent dans le contexte de la cartographie de l'occupation des sols à partir de séries temporelles. Néanmoins, l'algorithme très utilisé du *Convolutional Neural Network* (CNN) a été appliqué avec succès sur des images satellitaires à très haute résolution spatiale [Maggiori et al., 2017; Postadijan et al., 2017]. Un des principaux avantages du CNN est de s'affranchir de l'extraction des variables pertinentes grâce aux différentes couches de convolution du réseau. Cependant, plusieurs limitations persistent pour la classification de séries temporelles à haute résolution spatiale : 1) le paramétrage (les nombres de couches et de neurones) des réseaux est complexe, 2) un grand nombre d'échantillons d'apprentissage est requis, 3) des ressources informatiques importantes sont nécessaires, et 4) l'interprétation physique des résultats est impossible (boîte noire) à cause de la complexité des modèles produits (jusqu'à plusieurs millions de neurones).

En outre, le CNN est issu du domaine multimédia pour lequel la phase d'apprentissage est réalisée à partir de vignettes RGB (problème à trois dimensions). Une vignette correspond à une petite image carrée (de l'ordre de la dizaine de pixels de côté) pour laquelle la classe d'intérêt occupe une place majoritaire. En télédétection, il est difficile d'identifier des régions carrées pour toutes les occupation des sols. Par exemple, les routes et les rivières sont des linéaires qu'il est impossible de représenter par des images carrées. De plus, des vignettes peuvent en télédétection contenir plusieurs occupations des sols, particulièrement lors de l'utilisation d'images à une résolution spatiale de l'ordre de la dizaine de mètre. Typiquement, il sera difficile de prélever une vignette homogène autour d'un bâtiment isolé ; la route et le jardin seront aussi visibles. Lors de la prédiction, des approches dites de segmentation sémantique – *e.g.* les architectures SegNet et UNet – peuvent prédire plusieurs classes au sein d'une même vignette. En revanche, les régions attribuées à une classe ont tendance à déborder sur les autres classes dues à la présence des couches de déconvolution.

Afin de remédier aux temps d'apprentissage très longs des CNN, des réseaux de neurones pré-entraînés (*frameworks* en anglais) ont été mis à disposition. Ils ont permis de démocratiser l'utilisation du CNN. Dans ce cas là, les principaux paramètres du réseau sont déjà appris, et une étape peu coûteuse permet d'adapter le réseau à son problème de classification. Malheureusement, ces réseaux sont pré-entraînés sur des images RGB, et ne peuvent donc pas être directement appliqués sur des séries temporelles. L'atout des résolutions spectrales et temporelles ne peut donc pas être pleinement exploité. Finalement, la robustesse du CNN en présence de données mal étiquetées est inconnue puisque dans le domaine multimédia les imagettes utilisées pour l'apprentissage sont toutes correctement étiquetées.

Le potentiel des réseaux de neurones profonds, comme le CNN, est indéniable mais nécessite encore des adaptations pour être appliqué de manière opérationnelle au problème de la cartographie de l'occupation des sols [30]. Pour toutes ces raisons, les réseaux de neurones profonds n'ont pas été abordés dans ces travaux. La suite de ce chapitre se focalise donc sur la description technique du SVM et du RF.

30. Bien que les réseaux de neurones profonds ont remporté de nombreux concours dans le domaine du *machine learning* ces dernières années, on notera que la première place du *Data Fusion Contest* 2017 a été gagnée par une variante de l'algorithme du RF sur un problème de classification en milieu urbain avec des séries temporelles d'images satellitaires.

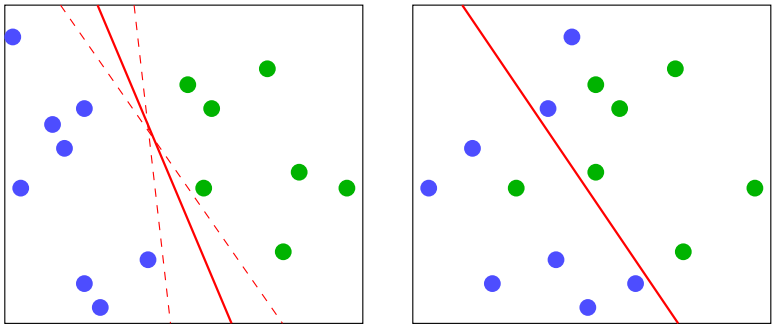
2.4 Algorithmes de classification supervisée

L'objectif de cette partie est de présenter en détail les deux algorithmes de classification utilisés dans le cadre de ces travaux : le SVM et le RF.

2.4.1 *Support Vector Machine*

Le problème abordé par le SVM est celui de la discrimination binaire pour lequel l'objectif est de définir une règle de décision associant à chaque observation sa classe. Il a été initialement posé pour la classification de deux ensembles linéairement séparables.

La Figure 2.3 montre deux exemples de discrimination où il s'agit de séparer les échantillons bleus des échantillons verts. La Figure 2.3a montre un cas où les deux ensembles d'échantillons sont linéairement séparables. Dans ce cas là, il est possible de trouver une frontière de décision linéaire, aussi appelée séparateur linéaire, qui sépare correctement tous les échantillons de l'ensemble d'apprentissage. Cependant, l'exemple de la Figure 2.3b montre un cas où les données sont non-linéairement séparables. Dans ce cas là, il est impossible de définir une frontière linéaire qui permette de séparer correctement les deux ensembles.



(a) Cas linéaire (b) Cas non-linéaire

FIGURE 2.3 – Exemple de discrimination binaire où il s'agit de séparer les points bleus des points verts. La frontière de décision est représentée en rouge.

Le cas linéaire est peu présent dans des problèmes réels de classification. Il permet cependant d'introduire simplement la théorie du SVM avec plusieurs notions clés. C'est pourquoi, la résolution du problème linéaire est d'abord présentée avant de décrire la résolution du problème non-linéaire. Puis, la généralisation aux problèmes de classification multi-classes est abordée. Enfin, le calcul d'un vecteur de probabilité d'appartenance aux classes est décrit.

Données linéairement séparables

L'objectif est de déterminer une frontière de décision linéaire qui sépare correctement deux ensembles d'échantillons (Figure 2.3a). Cette frontière de décision est une droite affine dans un espace à deux dimensions ($p = 2$), un plan dans un espace à trois dimensions ($p = 3$), et de manière générale un hyperplan dans un espace à p dimensions.

Comme le montre les droites en pointillé rouge dans la Figure 2.3a, il existe généralement une infinité de frontières de décision qui séparent correctement les deux ensembles dans un cas linéaire. Parmi toutes les solutions possibles, celle maximisant la distance

entre l'ensemble des échantillons et la frontière de décision est considérée comme optimale. Cette frontière de décision permet de maximiser la confiance lors de la prise de décision, et donc en principe de diminuer la probabilité d'erreur.

La Figure 2.4 montre un nouveau problème de classification à deux classes (bleue et verte). Affectons l'étiquette $+1$ aux échantillons verts et -1 aux échantillons bleus. La ligne en rouge représente la frontière de décision optimale qui est définie par le vecteur $\boldsymbol{w}$ et le biais b tel que $\boldsymbol{w}^T\boldsymbol{x} + b = 0$. Comme expliqué précédemment, la frontière de décision choisie est celle qui maximise la distance entre les échantillons et la frontière de décision. Cette distance est appelée la marge. Par exemple, la marge de l'échantillon vert p_1 est la distance entre p_1 et la frontière de décision en rouge dans la Figure 2.4.

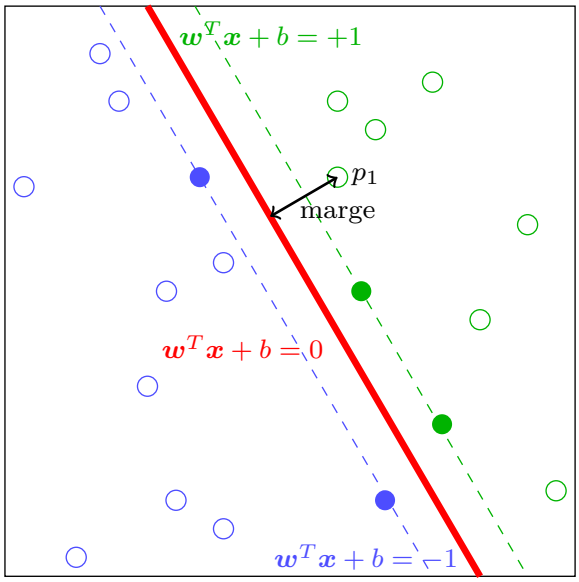


FIGURE 2.4 – Illustration de la notion de marge pour l'échantillon vert p_1 dans le cas de données linéairement séparables. La droite en rouge représente la frontière de décision.

Trouver la frontière optimale revient alors à maximiser le minimum des marges. Par exemple, cette distance minimale est illustrée par les courbes en pointillé bleu et vert. Pour simplifier, la marge d'un ensemble d'échantillons appartenant à une même classe correspondra alors au minimum des marges dans la suite. Dans l'exemple de la Figure 2.4, la marge est alors la distance entre la frontière de décision en rouge et les lignes en pointillé vert et bleu.

La résolution du problème posé par le SVM consiste à trouver le vecteur $\boldsymbol{w}$ et le biais b qui maximise la marge. Elle repose sur la théorie statistique de l'apprentissage [Vapnik, 1995, 1998]. Soit un ensemble de n échantillons d'apprentissage $\{(\boldsymbol{x}_i, y_i), i = 1...n\}$, avec $\boldsymbol{x}_i \in \mathbb{R}^p$ les valeurs des p variables pour le i-ème échantillon, et y_i l'étiquette associée. Formellement, le vecteur $\boldsymbol{w}$ et le biais b définissant la frontière de décision sont donnés par la résolution du problème suivant :

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \text{ sous contraintes } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1,\ i = 1, ..., n. \tag{2.1}$$

La fonction de décision D, qui représente la règle de classification, s'écrit formellement de la manière suivante :

$$D(\boldsymbol{x}) = sgn(\boldsymbol{w}^T\boldsymbol{x} + b), \tag{2.2}$$

avec sgn la fonction signe. Dans l'exemple de la Figure 2.4, si $\boldsymbol{w}^T\boldsymbol{x}+b > 0$ alors $D(\boldsymbol{x}) = +1$, et donc la classe verte est associée à l'échantillon $\boldsymbol{x}$. Au contraire, si $\boldsymbol{w}^T\boldsymbol{x} + b < 0$, la classe bleue sera associée à l'échantillon $\boldsymbol{x}$.

Le problème d'optimisation sous contraintes défini par l'équation (2.1) est un problème quadratique [31]. Comme par hypothèse les ensembles sont linéairement séparables, le problème est convexe et n'admet qu'une seule solution optimale. L'absence d'optimum local garantit la convergence de l'algorithme du SVM vers la solution optimale. Comme pour tout problème d'optimisation, une formulation duale équivalente est aussi possible, où la solution optimale peut être déterminée en recherchant le point selle du lagrangien $\mathcal{L}$. Dans le cas du SVM, elle s'exprime sous la forme :

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) - 1\right), \tag{2.3}$$

avec $\alpha_i \geq 0$ les multiplicateurs de Lagrange associés aux contraintes. Les conditions d'optimalité de Karush, Kuhn et Tucker (stationnarité, complémentarité, admissibilité primale, et admissibilité duale) permettent de caractériser la solution du problème [32]. Les échantillons pour lesquels les multiplicateurs de Lagrange α_i ne sont pas nuls sont appelés les vecteurs support, ensemble noté $\mathcal{V}$. Sur l'exemple de la Figure 2.4, les vecteurs support sont les échantillons appartenant aux droites en pointillé vert et bleu (points entièrement colorés). Les autres échantillons d'apprentissage ont un multiplicateur de Lagrange nul ($\alpha_i = 0$), et ne sont donc pas utilisés pour la résolution du SVM. La fonction de décision peut alors s'écrire sous la forme :

$$D(\boldsymbol{x}) = sgn\left(\sum_{i\in\mathcal{V}} \alpha_i y_i \left(\boldsymbol{x}^T\boldsymbol{x_i}\right) + b\right). \tag{2.4}$$

Données non-linéairement séparables

Dans un problème réel, les données sont rarement linéairement séparables (Figure 2.3b). Ainsi, le concept de marge souple a été proposé par Cortes and Vapnik [1995] afin d'accepter des erreurs de classification lors de la recherche de l'hyperplan optimal.

La Figure 2.5 illustre un nouvel exemple de classification à deux classes. Une frontière de décision linéaire peut presque être trouvée, mais quelques échantillons seraient alors mal classés par la frontière de décision. L'objectif est donc de définir l'hyperplan qui sépare les échantillons en faisant le moins d'erreur possible.

Pour ce faire, les variables ressort ξ_i (*slack variables* en anglais) sont introduites pour relâcher les contraintes sur les données d'apprentissage :

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i,\ i = 1, ..., n. \tag{2.5}$$

L'objectif de ce nouveau problème est toujours de maximiser la marge, mais en minimisant la somme des erreurs permises. Si l'échantillon $\boldsymbol{x}_i$ est du bon côté de la marge, alors la variable ressort ξ_i associée est nulle. Au contraire, si l'échantillon $\boldsymbol{x}_i$ est du mauvais côté de la marge, alors la variable ressort ξ_i correspond à la distance entre $\boldsymbol{x}_i$ et la marge.

31. Une fonction quadratique est une fonction polynomiale du second degré qu'il est possible de représenter sous la forme d'une parabole.
32. Non explicitées dans le manuscrit.

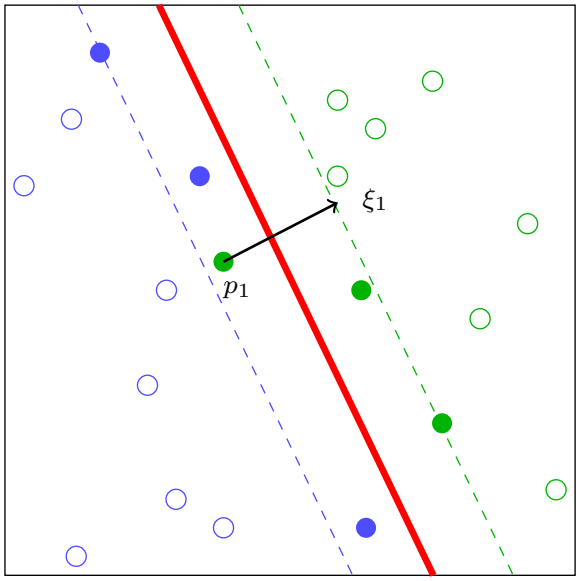


FIGURE 2.5 – Illustration de la notion de variables ressort dans le cas de données non-linéairement séparables. La variable ressort ξ_1 associée à l'échantillon vert p_1 représente la distance entre p_1 et la ligne en pointillé vert.

La Figure 2.5 illustre le principe de variables ressort pour l'échantillon vert p_1 qui est entre les deux droites en pointillé bleu et vert. Sa variable ressort ξ_1 est donc non nulle, et correspond à la distance entre p_1 et la ligne en pointillé vert.

Formellement, les valeurs de $\boldsymbol{w}$ et b sont données par la résolution du problème suivant :

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \begin{cases} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \sum_{i=1}^{n} \xi_i \end{cases} \quad \text{sous contraintes} \quad \begin{cases} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \text{ et} \\ \xi_i \geq 0,\ i = 1, ..., n. \end{cases} \tag{2.6}$$

Afin de résoudre ce nouveau problème, un paramètre de régularisation C est introduit. Configuré par l'utilisateur, il permet de déterminer le compromis entre l'importance des variables ressort et la largeur de la marge. Le vecteur $\boldsymbol{w}$ et le biais b sont alors donnés par la résolution du problème suivant :

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i \text{ avec } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \text{ et } \xi_i \geq 0,\ i = 1, ..., n. \tag{2.7}$$

Si toutes les variables ξ_i sont nulles, le problème est identique au cas linéaire. Comme précédemment, la résolution peut se faire avec le lagrangien :

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(y_i \left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) - 1\right) - \sum_{i=1}^{n} \beta_i \xi_i, \tag{2.8}$$

avec $\alpha_i \geq 0$ et $\beta_i \geq 0$ les multiplicateurs de Lagrange.

Toujours dans le cas non-linéaire, une autre solution a été proposée. Elle consiste à projeter les données dans un espace de dimension supérieure où une séparation linéaire pourrait être trouvée. La projection consiste à appliquer une transformation non-linéaire aux données d'entrée en utilisant une fonction noyau [Müller et al., 2001; Schölkopf and Smola, 2002]. L'utilisation d'une fonction noyau (*kernel trick* en anglais) permet de remplacer le calcul du produit scalaire par une autre mesure de similarité sans complication

algorithmique. Ceci est possible grâce à la formulation duale du problème du SVM qui permet d'introduire la non-linéarité à travers les noyaux.

Plus spécifiquement, le noyau est une fonction K qui associe à tout couple d'observations $\boldsymbol{x_1}$ et $\boldsymbol{x_2}$ une mesure de leur influence réciproque à travers leur distance, leur similarité ou leur corrélation. Une fonction noyau doit respecter les conditions du théorème de Mercer. Le noyau K doit notamment être positif, *i.e.* la matrice de Gram [33] associée à K doit être une matrice symétrique définie positive. La fonction de décision du SVM lors de l'utilisation d'un noyau K s'exprime alors de la manière suivante :

$$D(\boldsymbol{x}) = sgn\left(\sum_{i\in\mathcal{V}} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right). \tag{2.9}$$

De nombreuses fonctions noyau ont été proposées [Schölkopf and Smola, 2002]. Dans la littérature, le noyau gaussien K_{RBF} – *Radial Basis Function* (RBF) en anglais – reste le plus utilisé dans la littérature. Il s'exprime sous la forme :

$$K_{RBF}(\boldsymbol{x_1}, \boldsymbol{x_2}) = exp\left(-\gamma\|\boldsymbol{x_1} - \boldsymbol{x_2}\|^2\right), \tag{2.10}$$

avec $\gamma > 0$ le paramètre qui permet de contrôler l'écart-type de la gaussienne du noyau. Par abus de langage, on parle de noyau linéaire dans le cas où aucun noyau est utilisé. Le produit scalaire est alors calculé dans l'espace initial sans effectuer aucune projection comme dans l'équation (2.2).

Validation croisée

L'utilisation du SVM-linéaire ou du SVM-RBF nécessite donc la configuration d'un ou plusieurs hyper-paramètres (C et γ). Ce paramétrage est une étape cruciale dans le processus de classification car un mauvais paramétrage peut conduire à une très mauvaise prédiction. Il est donc important de tester les différents paramétrages possibles et de sélectionner le meilleur. La stratégie la plus longue, mais aussi la plus complète consiste à utiliser une grille de recherche. Dans cette approche, un modèle est construit pour chaque configuration de paramètres, *e.g.* le couple (C, γ) pour l'algorithme du SVM-RBF. Celui donnant le meilleur résultat est alors sélectionné. Idéalement, les échantillons utilisés pour évaluer la qualité des modèles doivent être indépendants des échantillons d'apprentissage et des échantillons test : ce sont les échantillons de validation. Cependant, partitionner en trois sous-ensembles – apprentissage, validation et test – les données de référence n'est pas toujours possible dû au manque de données de référence de qualité. Il est alors courant d'utiliser une stratégie d'échantillonnage des échantillons d'apprentissage appelée validation croisée [Hsu et al., 2003], même si d'autres méthodes d'optimisation existent [Fauvel, 2012]..

L'approche la plus connue est la validation croisée sur k partitions (*k-fold cross-validation* en anglais). Dans cette configuration, les échantillons d'apprentissage sont divisés en k partitions indépendantes. $k-1$ partitions sont réunies pour former les échantillons d'apprentissage, et donc construire le modèle, tandis que la dernière partition est utilisée pour calculer une mesure d'évaluation M comme l'*Overall Accuracy* (OA).

Ce principe est illustré par la Figure 2.6 pour cinq partitions ($k = 5$) : les partitions utilisées pour l'apprentissage sont en grises, tandis que celles pour la validation sont en vertes. Par exemple, au Tour 1, les quatre premières partitions sont utilisées pour

33. La matrice de Gram pour l'ensemble des observations $\{\boldsymbol{x_1}, ...\boldsymbol{x_i}, ..., \boldsymbol{x_n}\}$ est la matrice carrée G de taille n dont les coefficients sont définis par $G_{ij} = K(\boldsymbol{x_i}, \boldsymbol{x_j})$ pour $1 \leq i, j \leq n$.

apprendre le modèle, et la cinquième partition est utilisée pour valider le modèle construit en estimant $M1$.

Cette opération est répétée en laissant chacune des partitions en validation. Finalement, les mesures d'évaluation obtenues pour chacune des partitions sont moyennées pour obtenir une estimation de l'erreur de prédiction.

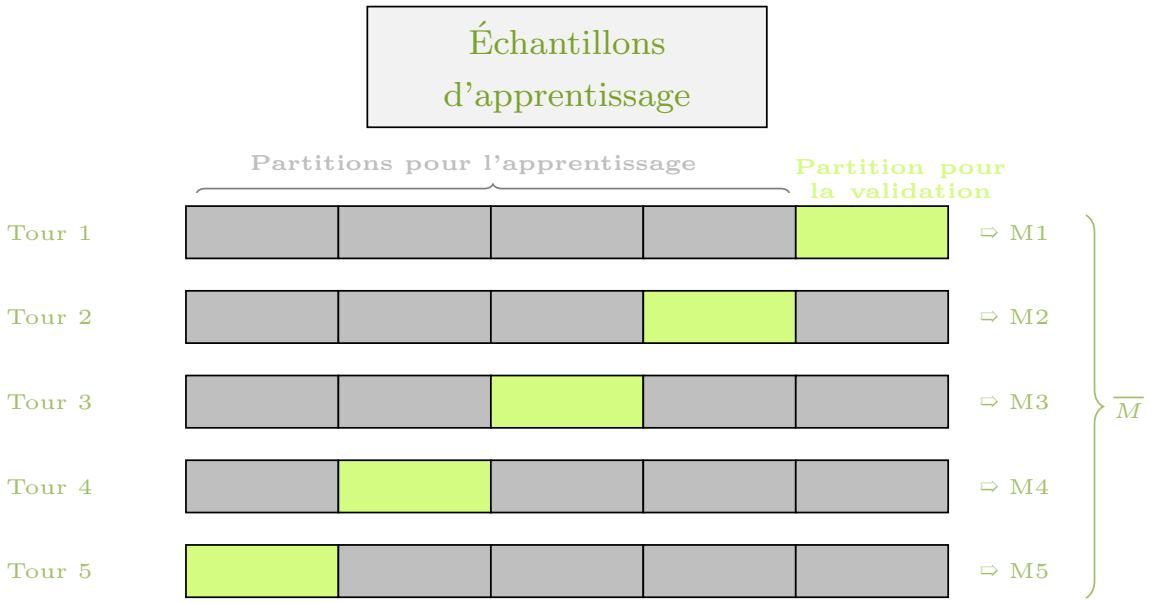


FIGURE 2.6 – Validation croisée sur cinq partitions.

À noter que la méthode de validation croisée *leave-one-out* est un cas particulier où le nombre de partitions k est égal au nombre d'échantillons d'apprentissage N.

Cas multi-classes

Afin de pouvoir adapter l'algorithme du SVM aux problèmes multi-classes, deux stratégies ont été mises en place dans la littérature [Hsu and Lin, 2002]. La première « un-contre-tous » (*one-against-all* en anglais) consiste à apprendre un modèle pour chaque classe : les échantillons positifs sont ceux de la classe considérée, tandis que les échantillons négatifs sont ceux de toutes les autres classes. La seconde approche « un-contre-un » (*one-against-one* en anglais) consiste à apprendre un modèle pour chaque paire de classes. Quelque soit la stratégie choisie, la phase de prédiction peut se faire soit par un vote majoritaire, soit par l'estimation des probabilités *a posteriori*.

La comparaison des deux stratégies dans la littérature donne parfois des résultats contradictoires [Milgram et al., 2006]. Il est cependant admis que l'approche « un-contre-un » requièrt l'apprentissage de plus de modèles, mais que la convergence de chacun de ces modèles est plus rapide. Ainsi, Milgram et al. [2006] conseillent de préférer une approche « un-contre-tous » en présence de peu de classes et peu d'échantillons, et une approche « un-contre-un » lorsque le nombre d'échantillons d'apprentissage est grand [34].

Dans ces travaux, les deux implémentations du SVM utilisées (LibSVM [Chang and Lin, 2011] et OpenCV basé aussi sur la LIbSVM) utilisent une approche « un-contre-un ».

34. Une approche globale, où le problème multi-classes est traité en une seule fois, est aussi possible, mais plus coûteuse en terme de calcul.

Probabilité d'appartenance aux classes

Comme pour de nombreux algorithmes de classification, il est possible de dériver une probabilité d'appartenance aux classes pour chaque échantillon à partir du modèle du SVM. Ces probabilités permettent par exemple d'estimer la confiance du classifieur pour chaque échantillon.

Dans le cas d'un problème de classification à deux classes, les distances des observations à l'hyperplan sont utilisées comme proxy pour calculer ces probabilités. La méthode la plus utilisée est la calibration de Platt. Elle consiste à estimer les probabilités en ajustant un modèle de régression logistique à la variable proxy [Lin et al., 2007; Platt, 1999]. Pour le cas multi-classes, Wu et al. [2004] proposent de généraliser le calcul des probabilités lors de l'utilisation d'une approche de résolution de type « un-contre-un ».

2.4.2 *Random Forest*

Le RF est un algorithme d'apprentissage supervisé basé sur la technique de l'arbre de décision binaire. La particularité du RF est de combiner un ensemble d'arbres de décision binaire afin de construire sa règle de décision.

Dans un premier temps, le principe des méthodes d'ensemble est décrit. Dans un deuxième temps, l'induction des arbres de décision binaire est expliquée. Finalement, une dernière partie est dédiée au principe de fonctionnement du RF.

Méthodes d'ensemble

L'idée des méthodes d'ensembles est de combiner les prédictions de différents algorithmes de classification afin d'obtenir un classifieur plus performant [Dietterich, 2000a]. L'utilisation des prédictions de plusieurs algorithmes de classification au lieu d'un seul permet notamment :

— d'améliorer la décision finale en s'appuyant sur les prédictions de plusieurs classifieurs ayant des comportements différents,

— d'obtenir un classifieur plus générique et moins sujet à proposer une solution sous-optimale,

— de traiter des problèmes complexes qui ne peuvent être résolus de manière optimale avec un seul algorithme de classification.

Deux stratégies existent pour la construction de ces ensembles : 1) combiner différents algorithmes de classification, et 2) combiner différentes variantes d'un même algorithme de classification. Les sorties de chaque classifieur sont ensuite fusionnées, généralement par un vote majoritaire [Rokach, 2010].

L'intérêt des méthodes d'ensemble réside donc dans l'utilisation de classifieurs qui ont des comportements considérablement différents. Les différences de comportement entre classifieurs peuvent être quantifiées à travers la notion de diversité. Bien qu'il n'y ait pas de consensus dans la littérature sur la définition de la diversité, il est admis qu'une mesure élevée de diversité correspond à une méthode d'ensemble qui a de bonnes performances en prédiction [Brown et al., 2005; Kuncheva and Whitaker, 2003; Mellor and Boukir, 2017]. Cependant, l'utilisation de classifieurs qui ont des prédictions trop contradictoires, et donc une très grande mesure de diversité, résultera en un classifieur final de faible qualité [Kapp et al., 2007].

Un autre concept important des méthodes d'ensemble est la marge introduite par Schapire et al. [1998]. La marge est définie comme la différence entre le pourcentage de

classifieurs votant correctement et le pourcentage de ceux votant incorrectement. Elle permet de fournir une mesure de confiance dans les méthodes d'ensemble [Guo, 2011; Mellor et al., 2015]. À noter que les concepts de diversité et de marge sont très liés puisque qu'une majorité des définitions proposées pour la diversité sont fonction de la marge [Stapenhurst, 2012].

Dans la suite, trois techniques pour construire des méthodes d'ensembles à partir d'un même algorithme de classification sont décrites.

La première technique présentée est celle du *random subspace* [Bryll et al., 2003; Ho, 1998]. Dans cette approche, la diversité est ajoutée en utilisant un sous-ensemble de variables tirées aléatoirement sans remise pour décrire les échantillons. En revanche, tous les échantillons sont utilisés pour apprendre l'ensemble des classifieurs. Ainsi, chaque classifieur est spécialisé pour un groupe de variables spécifiques. La combinaison des classifieurs permet alors d'obtenir un algorithme fiable sur tout l'espace des caractéristiques.

Les deux autres techniques présentées sont le *bagging* et le *boosting*. La diversité est ajoutée en jouant sur l'utilisation des échantillons d'apprentissage pour construire chaque classifieur.

Plus spécifiquement, le *bagging* (de *bootstrap agreggating*) consiste à construire plusieurs classifieurs en s'appuyant sur des sous-ensembles d'échantillons d'apprentissage différents [Breiman, 1996]. Chaque sous-ensemble d'échantillons d'apprentissage, dit échantillons *bootstrap*, est obtenu par un tirage au sort de N échantillons[35] avec remise parmi les N échantillons d'apprentissage [Efron and Tibshirani, 1994]. La combinaison des prédictions des classifieurs appris sur les différents échantillons *bootstrap* permet d'améliorer la capacité de généralisation des algorithmes sujets au sur-apprentissage en diminuant la variance individuelle de chaque classifieur (Section 2.1). Par ailleurs, la construction des différents classifieurs se parallélise facilement.

Concernant le *boosting*, il repose sur la construction de classifieurs faibles (*weak classifiers* en anglais), *i.e.* des classifieurs qui font mieux que le hasard. Dans le cas d'une classification binaire, un classifieur est dit faible s'il se trompe moins d'une fois sur deux.

Dans ce cas là, l'ensemble des classifieurs est construit de manière récursive : chaque classifieur est une version adaptative du précédant où les échantillons mal prédits sont sur-pondérés. Ainsi, le classifieur de l'étape t se focalisera sur les échantillons mal prédits à l'étape $t-1$. Pour prédire l'étiquette de nouveaux échantillons, le vote de chaque classifieur est pondéré en fonction de sa précision. L'algorithme le plus connu reposant sur ce principe est *AdaBoost* (*Adaptive Boosting*) [Freund and Schapire, 1996]. Contrairement au *bagging*, l'étape d'apprentissage du *boosting* ne peut pas être parallélisée puisque les classifieurs sont dépendants les uns des autres.

Arbres de décision binaire

Dans le contexte de la classification, les arbres de décision permettent de résumer un ensemble de règles dans une structure d'arbre hiérarchique. Ils ont pour principal avantage de fournir une représentation graphique et intuitive de la règle de décision qui permettra de déterminer l'étiquette de nouveaux échantillons.

La Figure 2.7 montre un exemple d'arbre de décision binaire pour la classification de cinq classes de culture en fonction de différentes informations caractérisant leur cycle phénologique. À l'état initial, l'arbre est constitué de la racine qui teste la date du début de croissance. Si cette dernière a eu lieu avant le 1er avril, alors l'échantillon emprunte la

35. Une autre valeur est possible, mais généralement le nombre d'échantillons des sous-ensembles est identique à celui de l'ensemble de départ.

branche de gauche. Sinon, l'échantillon suit la branche de droite. Chaque nœud est ainsi défini par le choix conjoint d'une variable et d'un test qui va induire une partition en deux sous-ensembles. Les nœuds terminaux en orange, appelés aussi feuilles, sont des nœuds qui ne possèdent pas de nœuds fils. Ils contiennent la décision de classement final. La notion de niveau correspond à la profondeur des différents nœuds. Par défaut, la racine est au niveau 0.

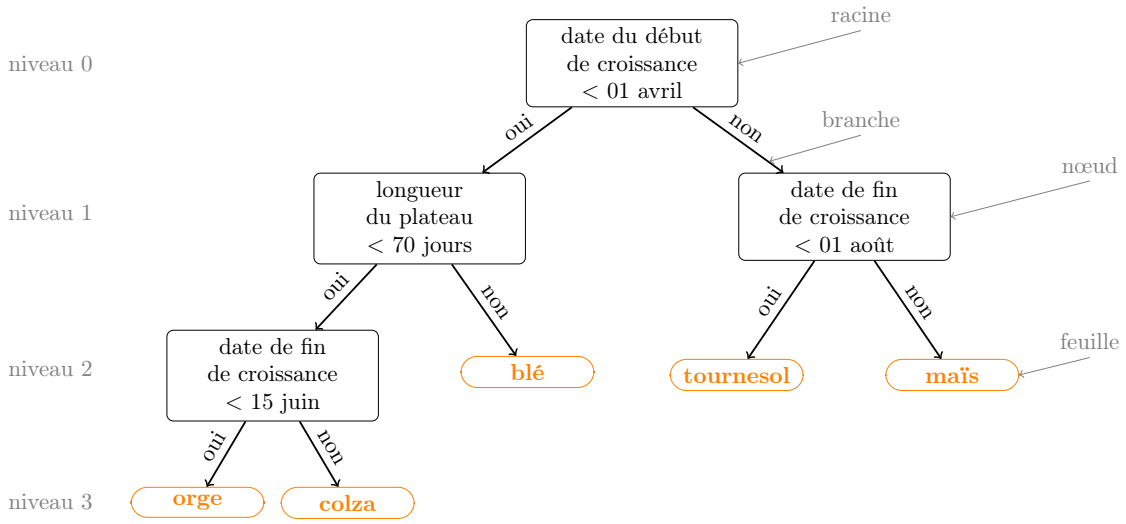


FIGURE 2.7 – Exemple d'arbre de décision binaire pour la classification des cultures en fonction de différentes informations caractérisant leur cycle phénologique.

La construction d'un arbre de décision binaire débute avec la création de la racine qui contient l'ensemble des échantillons d'apprentissage. L'objectif est d'ajouter de nouveaux nœuds qui permettent de diviser les échantillons en sous-ensembles plus homogènes. Idéalement, un ensemble d'échantillons est homogène si les échantillons ont des comportements similaires, c'est-à-dire qu'ils appartiennent à la même classe pour un problème de classification. Comme le montre la Figure 2.7, la construction se poursuit sur plusieurs niveaux jusqu'à l'obtention de nœuds terminaux. Pour résumer, la construction d'un arbre de décision binaire nécessite :

1. La définition d'une règle de partitionnement qui permet de diviser les échantillons en sous-ensembles plus homogènes.
2. Une règle permettant de décider qu'un nœud est terminal.
3. Une règle permettant l'affectation de chaque feuille à l'une des classes. Généralement, la classe attribuée à une feuille correspond à celle la plus représentée parmi les échantillons d'apprentissage qui y appartiennent.

La règle de partitionnement est associée à chaque nœud afin de répartir les échantillons dans deux nœuds fils. Cette règle est déterminée de la manière suivante : une variable et un test associé à cette variable sont sélectionnés dans l'ensemble des variables qui décrivent les échantillons. Les échantillons sont ensuite répartis en fonction du résultat du test : si la réponse au test est positive, alors les échantillons vont dans le nœud fils de gauche, sinon ils vont dans le nœud fils de droite.

Le point crucial est le choix de la variable et du test associé à cette variable. La majorité des méthodes d'arbres de décision repose sur la même stratégie. À chaque nœud, un critère d'évaluation est évalué pour chacune des variables et pour les tests possibles sur ces variables. La variable associée à un test qui maximise le critère d'évaluation est

alors choisie. Le critère d'évaluation est généralement basé sur une mesure d'impureté, qui dépend du degré d'homogénéité des échantillons appartenant au nœud. L'impureté est minimale lorsque les échantillons appartiennent tous à la même classe, on parle alors de nœud pur. Au contraire, l'impureté est maximale si les échantillons sont répartis équiprobablement entre toutes les classes.

Formellement, le critère ΔI cherche à maximiser la différence d'impureté entre la population P du nœud et celles des populations P_g et P_d des deux nœuds fils.

$$\Delta I(P, P_g, P_d) = I(P) - (I(P_g) + I(P_r)), \tag{2.11}$$

avec I une mesure d'impureté. Comme de nombreux travaux montrent que le choix de I n'est pas décisif [Murthy, 1998; Robnik-Sikonja, 2004], seules les deux mesures d'impureté les plus couramment utilisées sont présentées ici. Ces mesures sont calculées pour une population P composée de m échantillons d'apprentissage appartenant à K classes. Le nombre d'échantillons appartenant à la k-ième classe est noté m_k.

La première mesure est le gain d'information utilisée pour l'induction des arbres ID3 et C4.5 [Quinlan, 1986, 1993]. Elle consiste à mesurer la quantité d'information nécessaire pour déterminer la classe d'un échantillon. Elle est calculée en se basant sur l'expression de l'entropie de Shannon :

$$I_H(P) = -\sum_{k=1}^{K} \frac{m_k}{m} \times \log_2\left(\frac{m_k}{m}\right). \tag{2.12}$$

La seconde mesure est l'indice de Gini utilisé pour l'induction des arbres *Classification And Regression Trees* (CART) [Breiman et al., 1984]. Cette mesure représente l'erreur attendue si la classe d'un échantillon était choisie aléatoirement en suivant la distribution des échantillons de la population P. L'indice de Gini s'exprime alors de la manière suivante :

$$I_{Gini}(P) = 1 - \sum_{k=1}^{K} \left(\frac{m_k}{m}\right)^2. \tag{2.13}$$

Concernant le choix d'une règle permettant de décider qu'un nœud est terminal, la stratégie la plus simple consiste à décider qu'un nœud devienne une feuille s'il est pur, *i.e.* tous les échantillons qui le composent appartiennent à la même classe.

Cependant, la construction d'un arbre jusqu'à sa profondeur maximale, *i.e.* sans l'utilisation d'un critère d'arrêt, conduit généralement à l'obtention d'un modèle trop complexe qui s'adapte parfaitement aux échantillons d'apprentissage. Afin d'améliorer la capacité de généralisation du modèle construit et d'éviter le sur-apprentissage, il est possible d'arrêter la construction de l'arbre prématurément. Par exemple, si :

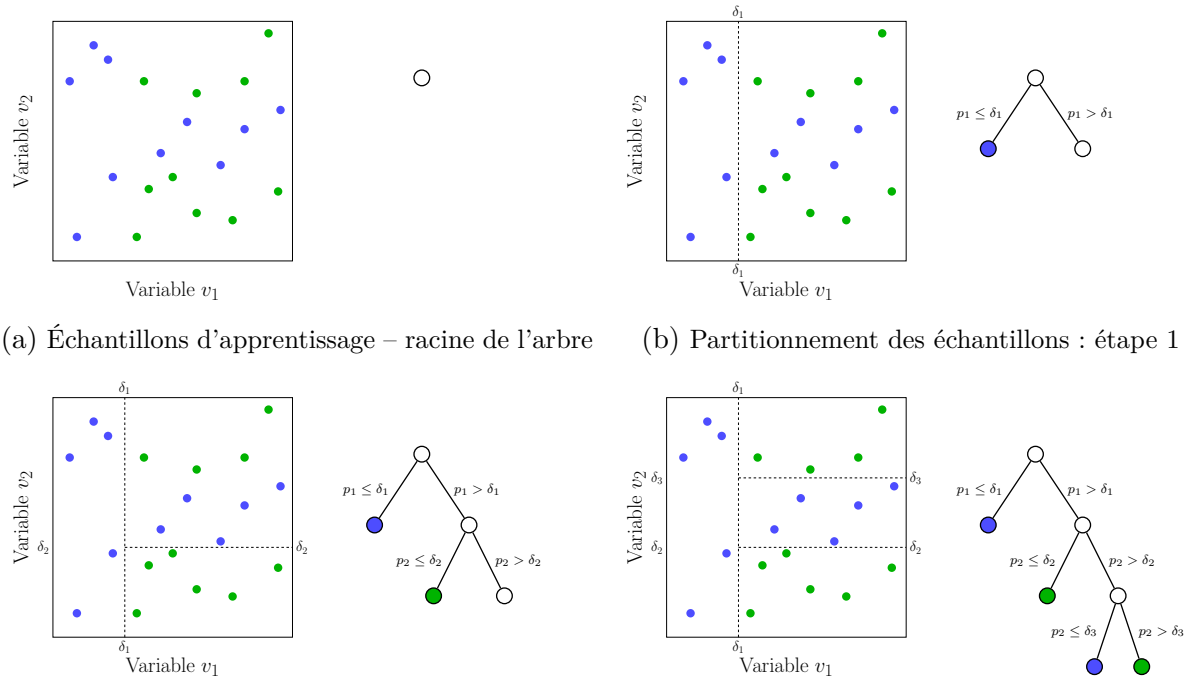
— une profondeur *max_depth* définie par l'utilisateur est atteinte,

— le nombre d'échantillons qui compose une feuille est inférieur à un paramètre *min_samples* défini par l'utilisateur,

— si la variance au sein des nœuds ne décroît pas au-delà d'un certain seuil.

Une autre solution est la méthode d'élagage (ou *pruning* en anglais). Elle consiste à construire l'arbre jusqu'à sa profondeur maximale, puis à supprimer les nœuds inintéressants.

Afin d'illustrer les différentes étapes de la construction d'un arbre de décision binaire, la Figure 2.8a montre les échantillons d'apprentissage d'un problème de classification à deux classes (verte et bleue) dans un espace à deux dimensions (variables v_1 et v_2).



(a) Échantillons d'apprentissage – racine de l'arbre

(b) Partitionnement des échantillons : étape 1

(c) Partitionnement des échantillons : étape 2

(d) Partitionnement des échantillons : étape 3

FIGURE 2.8 – Principe de construction d'un arbre de décision binaire pour un problème de classification binaire à deux classes (verte et bleue).

L'objectif est de construire un arbre de décision binaire (jusqu'à sa profondeur maximale) à partir de ces données d'apprentissage.

La Figure 2.8a montre l'étape initiale où l'ensemble des échantillons appartiennent à la racine. Ensuite, la figure 2.8b montre la première étape de partitionnement. La variable sélectionnée est v_1 pour une valeur de seuil δ_1. La règle de partitionnement dépend alors du test suivant : $v_1 < \delta_1$. À la suite de cette partition, le nœud fils de gauche est pur, *i.e.* il contient uniquement des échantillons appartenant à la classe bleue. Il devient une feuille qui votera pour la classe bleue. Comme le montre la Figure 2.8c, la procédure de séparation des échantillons est de nouveau appliquée pour le nœud fils de droite qui n'est pas homogène. Une étape est encore nécessaire avant que la construction de l'arbre s'arrête automatiquement, *i.e.* que toutes les feuilles soient des nœud purs (Figure 2.8d).

De l'arbre à la forêt

Les arbres de décision binaire décrits précédemment sont connus pour être très sensibles au sur-apprentissage, et avoir une faible capacité de généralisation. Cependant, leur phase d'apprentissage rapide et la lisibilité de leur règle de décision les rendent attractifs. Ainsi, il a été proposé de construire des ensembles d'arbres de décision binaire en tirant profit des bonnes propriétés des méthodes d'ensemble, notamment l'amélioration de la capacité de généralisation.

Dans la littérature, plusieurs méthodes ont été proposées pour ajouter de la diversité à un ensemble d'arbres de décision binaire. La diversité est apportée en modifiant le processus de construction des arbres, *e.g.* en changeant les échantillons utilisés pour l'apprentissage, le critère de partitionnement des nœuds ou encore le critère d'arrêt de construction d'un arbre.

Pour obtenir de bonnes performances, un ensemble d'arbres de décision binaire a besoin

que chaque arbre soit performant (faible biais, mais une forte variance est autorisée), et que les arbres du modèle soient faiblement corrélés. La corrélation entre les arbres est une mesure de diversité qui correspond au degré d'accord des prévisions des arbres. Deux arbres sont faiblement corrélés si leur prévision sur un même ensemble d'échantillons sont dissimilaires.

La méthode d'ensemble la plus connue utilisant un ensemble d'arbres de décision binaire est le *Random Forest - Random Input* proposée par Breiman [2001], qui est souvent appelée *Random Forest*. De manière classique, l'étiquette d'une nouvelle observation est obtenue par un vote majoritaire sur l'ensemble des résultats des arbres construits.

Le RF a pour spécificité :

— d'utiliser des échantillons *bootstrap* pour la construction des K arbres de décision qui constituent le modèle final ;

— d'utiliser le principe du *random feature selection*, *i.e.* à chaque nœud, le critère de partitionnement est évalué seulement pour un sous-ensemble de m variables tirées aléatoirement sans remise avec le critère de Gini donné par l'équation (2.13) ;

— de construire les arbres jusqu'à leur profondeur maximale.

L'utilisation d'échantillons *bootstrap* et du principe du *random feature selection* permet de diversifier les arbres construits, et donc de les décorréler. Par ailleurs, l'utilisation d'un faible nombre de variables pour la construction de chaque nœud permet de réduire la complexité algorithmique du RF.

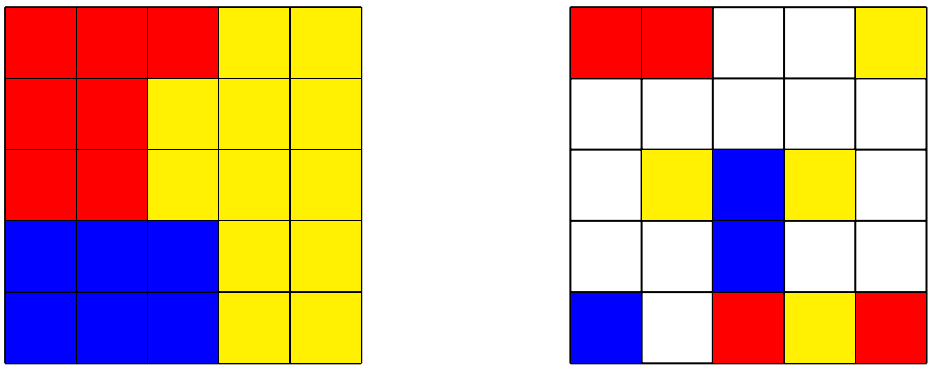
Dans l'implémentation utilisée (OpenCV), la construction des arbres est arrêtée prématurément si une profondeur maximale *max_depth* pré-définie est atteinte ou si le nombre d'échantillons du nœud est inférieur à un paramètre *min_samples*. Cette variante de la méthode initiale permet d'une part de réduire le sur-apprentissage des arbres, et d'autre part de réduire la complexité algorithmique. Chaque nœud terminal vote alors pour la classe présente en majorité parmi les échantillons d'apprentissage.

De plus, le fonctionnement du RF permet le calcul des trois métriques particulièrement intéressantes : 1) l'erreur *Out Of Bag* (OOB), 2) l'importance des variables, et 3) le vecteur de probabilités d'appartenance aux classes pour tous les échantillons.

Pour un arbre donné, la construction avec des échantillons *bootstrap* implique qu'une partie des échantillons d'apprentissage ne soit pas utilisée pour la construction de chaque arbre. Ces échantillons non-utilisés pour la construction sont appelés échantillons OOB. En moyenne, environ un tiers des échantillons sont OOB lorsque le nombre d'échantillons d'apprentissage est grand (Annexe B.1.1). Pour chaque arbre, les échantillons OOB permettent d'estimer l'erreur OOB en comptabilisant le nombre de fois où ces échantillons sont mal prédits [Breiman, 2001].

Le RF est aussi capable de donner une indication sur les variables les plus importantes. Cette information donne une connaissance des variables qui permettent d'expliquer le résultat de classification, et de celles superflues. De plus, les variables les plus importantes peuvent être utilisées afin de construire un classifeur de meilleure précision [Genuer et al., 2010].

De manière intuitive, les variables les plus importantes sont celles qui sont les plus utilisées dans la construction des arbres, notamment dans les premières divisions. Cependant, cette vision est trop simpliste à cause de l'aléatoire introduit par le *random feature selection*. Dans la littérature, deux stratégies plus complexes sont proposées pour évaluer quantitativement l'importance des variables. La première, la plus utilisée, est le *Mean Decrease Accuracy* [Genuer, 2010]. Pour une variable donnée, elle consiste à évaluer l'erreur OOB avant et après une permutation aléatoire des valeurs de la variable. Si



(a) Résultat de classification (b) Échantillons test

FIGURE 2.9 – Exemple d'une carte d'occupation des sols (trois classes) à évaluer à partir d'échantillons test extraits des données de référence.

une forte diminution de l'erreur OOB est observée, alors la variable est importante. La deuxième mesure est le *Mean Decrease Gini* [Breiman et al., 1984]. Elle consiste à mesurer la diminution d'impureté de chaque nœud faisant intervenir la variable considérée. Si une variable fait fortement diminuer l'impureté des nœuds, alors elle est considérée comme importante.

L'ensemble des arbres du RF permet aussi le calcul des probabilités d'appartenance aux classes pour chaque échantillon. Afin d'obtenir l'étiquette de nouvelles observations, chaque arbre vote pour une classe, et la classe majoritaire est attribuée. Le vote de tous les arbres permet, après division par le nombre d'arbres du modèle, d'obtenir un vecteur d'appartenance à chacune des classes.

À noter qu'un grand nombre de variantes autour des RF ont été proposées dans la littérature [Rokach, 2016], comme les *Extremely Randomized Trees*, une version totalement aléatoire sans utilisation du *bagging* où la variable et le test sont sélectionnés aléatoirement à chaque nœud [Geurts et al., 2006]. Cette version ne permet donc pas d'estimer l'importance des variables, ni l'erreur OOB sans un jeu test indépendant. D'autres variantes modifient les règles de construction du RF : 1) changement de la règle de partitionnement [Breiman, 2001; Robnik-Sikonja, 2004], 2) pondération des échantillons en fonction de leur utilité pour la décision finale [Sasikala et al., 2015], 3) sélection ou pondération des arbres en fonction de leur précision [Bernard et al., 2009; Robnik-Sikonja, 2004; Tsymbal et al., 2006]. L'ensemble de ces variantes est décrit dans Bernard [2009].

2.5 Évaluation des performances des algorithmes de classification

L'étape d'évaluation consiste à quantifier la précision de l'algorithme de classification, et donc de la carte produite par le système de classification. Cette étape est effectuée en comparant les classes de la donnée de référence avec celles prédites par le classifieur sur les échantillons test.

La Figure 2.9a montre un exemple d'image synthétique pour un problème de classification à trois classes (bleue, jaune et rouge). La Figure 2.9b montre les échantillons test disponibles (pixels non-blancs) pour cette évaluation.

Afin d'effectuer cette évaluation, un tableau à double entrées appelé matrice de confu-

sion ou tableau de contingence, est généralement utilisé. Chaque ligne représente le nombre d'occurrences d'une classe de la donnée de référence (réelle), tandis que chaque colonne représente le nombre d'occurrences d'une classe prédite par le système de classification. La matrice de confusion associée à l'évaluation de la carte de la Figure 2.9 est la suivante :

Réelle \ Prédite			
	2	1	0
	0	3	1
	1	1	2

Les échantillons sur la diagonale représentent donc le nombre d'échantillons test correctement prédit par l'algorithme de classification, sept dans l'exemple précédent. De cette matrice de confusion, il est possible de calculer un ensemble de métriques caractérisant les performances de l'algorithme de classification utilisé. Les échantillons test utilisés doivent référencer la classe sur le terrain afin de ne pas biaiser les résultats [Congalton and Green, 2008]. En pratique, les échantillons test sont considérés comme parfaits (*gold standard*) [Foody, 2002].

Dans une première partie, les métriques d'évaluation dans le cas d'une classification multi-classes sont décrites. Une seconde partie présente un moyen d'estimer l'erreur réelle commise par le système de classification en s'appuyant sur l'intervalle de confiance.

2.5.1 Évaluation d'une classification multi-classes

Dans le cadre d'un problème de classification à K classes, la matrice de confusion C est définie de la manière suivante :

$$\begin{bmatrix} c_{11} & \cdots & \cdots & c_{1l} & \cdots & \cdots & c_{1K} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{l1} & \cdots & \cdots & c_{ll} & \cdots & \cdots & c_{lK} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{K1} & \cdots & \cdots & c_{Kl} & \cdots & \cdots & c_{KK} \end{bmatrix}$$

Le coefficient c_{ij} donne le nombre d'occurrences d'échantillons test appartenant à la classe i et prédit classe j par le classifieur. Les échantillons test correctement prédits par l'algorithme de classification sont sur la diagonale de la matrice de confusion (coefficients c_{ii}).

La métrique la plus simple à dériver de la matrice de confusion est le taux de bonnes classifications (OA en anglais) qui est calculé comme le nombre d'échantillons test correctement prédits (trace de la matrice de confusion) divisé par le nombre total d'échantillons test :

$$\mathrm{OA} = \frac{1}{N} \sum_{i=1}^{K} c_{ii}, \tag{2.14}$$

avec $N = \sum_{i=1}^{K} \sum_{j=1}^{K} (c_{ij})$, *i.e.* le nombre d'échantillons test.

En plus de la valeur d'OA, il est courant de calculer le coefficient Kappa qui doit en théorie s'affranchir du taux de bonnes classifications dû à l'aléatoire :

$$\mathrm{Kappa} = \frac{\mathrm{OA} - p_h}{1 - p_h}, \tag{2.15}$$

avec $p_h = \frac{1}{N^2} \sum_{i=1}^{K} \left(\sum_{j=1}^{K} c_{ij} \right) \left(\sum_{j=1}^{K} c_{ji} \right)$, le pourcentage de bonnes classifications attribué au hasard. Le référentiel de Landis and Koch [1977] propose une interprétation du coefficient Kappa en fonction de sa valeur, présentée par le Tableau 2.1.

Tableau 2.1 – Référentiel pour interpréter la valeur de Kappa.

Interprétation	**Valeur de Kappa**
Excellente	1.00 – 0.81
Bonne	0.80 – 0.61
Faible	0.60 – 0.41
Négligeable	0.20 – 0.00
Mauvaise	< 0.00

Source : Landis and Koch [1977]

Bien que le coefficient Kappa soit encore très souvent utilisé, il fait polémique dans la communauté de la télédétection. La première critique concerne la formulation de p_h qui ne permettrait pas de représenter le taux de bonnes classification dû au hasard [Foody, 1992]. Une deuxième critique concerne la difficulté à réaliser des comparaisons simples entre des valeurs de coefficient Kappa obtenues par différents systèmes de classification [Pontius Jr and Millones, 2011]. Enfin, la dernière critique concerne la très forte corrélation entre la valeur du coefficient Kappa et de l'OA [Liu et al., 2007]. Pour ces travaux, la recommandation d'utiliser seulement l'OA de Stehman [1997] a été suivie [36].

Les métriques globales comme l'OA et le Kappa sont souvent insuffisantes pour mesurer la qualité de la classification, en particulier dans le cas où le nombre d'échantillons test par classe est très déséquilibré. En effet, ces mesures ne tiennent pas compte de la distribution des classes et des coûts de classification. Considérons le cas réel d'une cartographie d'arbres « pins *versus* chênes » où les pins sont bien plus présents que les chênes. Par exemple, 95 arbres sont des pins et 5 arbres sont des chênes. Si l'algorithme de classification décide que tous les arbres sont des pins, l'OA sera alors de 95 %. Cependant, la classe d'intérêt dans ce type de problème déséquilibré est souvent la classe minoritaire, *i.e.* les chênes. Dans ce cas, la valeur d'OA ne permet pas de mettre en évidence qu'aucun chêne n'est détecté par l'algorithme de classification utilisé.

Afin de prendre en compte des différences de performance entre classes, des métriques par classe sont utilisées. Les mesures de précision UA_i (*user's accuracy* en anglais) et rappel PA_i (*producer's accuracy* en anglais) définies pour la i-ème classe sont couramment utilisées :

$$\mathrm{UA}_i = \frac{c_{ii}}{\sum_{j=1}^{K} c_{ji}}, \tag{2.16}$$

$$\mathrm{PA}_i = \frac{c_{ii}}{\sum_{j=1}^{K} c_{ij}}. \tag{2.17}$$

La précision d'une classe correspond donc au pourcentage d'échantillons correctement prédits dans cette classe par rapport à l'ensemble des prédictions faites pour cette classe,

36. Comme les valeurs du coefficient Kappa et de l'OA ont toujours indiqué des résultats cohérents dans les études réalisées, il a été décidé de ne montrer que les résultats obtenus pour l'OA dans ce manuscrit.

tandis que le rappel représente le pourcentage d'échantillons de la donnée de référence correctement prédits pour cette classe. En fonction de l'application, seulement l'une des deux mesures peut être d'intérêt.

Dans l'exemple précédent, s'il est important de détecter tous les chênes, une valeur de rappel forte va être favorisée, quitte à détecter trop de pins comme étant des chênes (précision faible). Au contraire, si l'objectif est d'être certain que les arbres détectés comme des chênes soient bien des chênes, la précision devra être maximisée quitte à en manquer certain (rappel faible).

Pour la majorité des applications, un compromis entre la précision et le rappel est généralement souhaité. Il est alors possible de combiner les deux mesures en une seule nommée le F-Score (ou encore F-1 et F-Mesure). Cette dernière correspond à la moyenne harmonique[37] de la précision et du rappel :

$$\text{F-Score}_i = 2\frac{UA_i \times PA_i}{UA_i + PA_i}. \tag{2.18}$$

Toutes les mesures proposées ci-dessus ne prennent pas en compte le contexte spatial induit par la classification d'images satellitaires. Comber et al. [2012] proposent une méthode pour évaluer spatialement les résultats de classification. Bien que très intéressante, cette approche nécessite un grand nombre d'échantillons test répartis équitablement spatialement.

2.5.2 Évaluation statistique

Toutes les métriques précédentes sont évaluées sur un ensemble d'échantillons test qui représente seulement une sous-partie de tous les échantillons. De plus, les algorithmes de classification utilisent parfois des procédures aléatoires, par exemple la technique du *bagging* pour la sélection des échantillons dans des méthodes d'ensemble. La non-exhaustivité des échantillons test et l'aléatoire présent dans les algorithmes de classification conduisent à des incertitudes sur les mesures de performances évaluées.

Dans ces travaux, l'intervalle de confiance est calculé afin de connaître la confiance dans les mesures de performances estimées. Il sera aussi utilisé pour donner une indication sur les différences de performances entre plusieurs algorithmes de classification [Foody, 2009; Labatut and Cherifi, 2012]. Plus précisément, l'intervalle de confiance représente l'intervalle dans lequel la valeur réelle de la mesure de performance a x % de chance d'être comprise.

L'intervalle de confiance est ici mesuré pour $nturns$ tirages aléatoires des échantillons test. L'estimateur e d'une mesure de performance est calculé pour chaque tirage aléatoire. Ainsi, la moyenne $\bar{e}$ et l'écart-type σ_e de cet estimateur e peuvent être calculés pour les $nturns$ tirages aléatoires. L'intervalle de confiance s'exprime alors de la manière suivante :

$$\bar{e} \pm t_{x\%}\frac{\sigma_e}{\sqrt{nturns}}, \tag{2.19}$$

avec $t_{x\%}$ le fractile. La valeur du fractile $t_{x\%}$ dépend de la valeur de x, mais aussi de $nturns$. Dans ces travaux, une loi de Student est utilisée pour déterminer la valeur de $t_{x\%}$ car le nombre de tirages aléatoires $nturns$ est petit (5 ou 10 tirages).

37. Manning et al. [2008] recommandent l'utilisation de la moyenne harmonique, plutôt qu'une moyenne arithmétique ou géométrique. Avec la moyenne harmonique, si la précision ou le rappel est faible, le F-Score sera aussi faible.

Chapitre 3

Données utilisées

Sommaire

3.1 Données satellitaires utilisées 51
3.1.1 Caractéristiques des capteurs utilisés 51
3.1.2 Zones d'études 54
3.2 Pré-traitements des données satellitaires 55
3.2.1 Corrections géométriques et radiométriques 55
3.2.2 Données manquantes 57
3.2.3 Traitements appliqués 60
3.3 Données de référence utilisées 61
3.3.1 Registre Parcellaire Graphique 62
3.3.2 OCcupation des Sols à Grande Échelle 62
3.3.3 Données terrain 63

Les cartes d'occupation des sols sont produites à partir de méthodes d'apprentissage automatique. Comme montré au Chapitre 2, ces méthodes s'appuient sur les informations contenues dans les données satellitaires et les données de référence.

L'objectif de ce chapitre est de présenter l'ensemble des données utilisé – images satellitaires et données de référence – au cours de ces travaux de thèse. Une première partie est dédiée à la présentation des données satellitaires utilisées ainsi qu'à la description des zones d'études. Une deuxième partie décrit les pré-traitements appliqués aux données satellitaires. Enfin, une dernière partie présente les données de référence utilisées.

3.1 Données satellitaires utilisées

Dans cette partie, les données satellitaires utilisées au cours de ces travaux sont décrites. Dans un premier temps, les caractéristiques des trois satellites imageurs utilisés sont détaillées. Dans un second temps, les zones d'études sont présentées.

3.1.1 Caractéristiques des capteurs utilisés

Les travaux de thèse se focalisent sur l'étude des séries temporelles d'images satellitaires sur de grandes étendues, comme celles fournies par les satellites Sentinel-2. Cependant, les deux satellites Sentinel-2 ont été lancés au cours de la thèse – 23 juin 2015 et

7 mars 2017. Ainsi, différentes études de ces travaux de thèse sont réalisées avec les capteurs satellitaires SPOT-4 et Landsat-8 dont les caractéristiques se rapprochent de celles de Sentinel-2. Par exemple, Sentinel-2 et Landsat ont pour caractéristique commune de réaliser des acquisitions à angles constants. Ainsi, les scènes sont observées avec les mêmes angles solaires et d'acquisition. Les effets directionnels, qui font fortement varier les réflectances mesurées par le capteur, sont ainsi limités. La suite de cette partie détaille les caractéristiques de chaque capteur utilisé.

Satellite SPOT-4

Le satellite SPOT-4 a été lancé en 1998 et désorbité au cours de l'été 2013. Il permettait l'acquisition de données tous les 26 jours et sa charge utile était composée de trois instruments :

1. deux instruments Haute Résolution Visible et Infra-Rouge (HRVIR) identiques avec une fauchée de 60 kilomètres dont les caractéristiques sont données par le Tableau 3.1,
2. l'instrument VEGETATION à une résolution spatiale d'environ un kilomètre qui couvrait l'ensemble des surfaces continentales quasi-quotidiennement.

Tableau 3.1 – Caractéristiques de l'instrument Haute Résolution Visible et Infra-Rouge (HRVIR) embarqué sur le satellite SPOT-4.

	Canal	**Résolution spatiale**	**Bande spectrale**
Multispectral (XS)	Vert	20 m	$0.50 - 0.59\ \mu m$
	Rouge		$0.61 - 0.68\ \mu m$
	Proche infra-rouge (PIR)		$0.78 - 0.89\ \mu m$
	Moyen infra-rouge (MIR)		$1.58 - 1.75\ \mu m$
Panchromatique	PAN	10 m	$0.61 - 0.68\ \mu m$

Pour ses travaux de thèse, seules les quatre acquisitions multi-spectrales à 20 mètres sont utilisées.

L'intérêt des données SPOT-4 est d'obtenir une série temporelle d'images satellitaires dont les caractéristiques sont proches des données Sentinel-2. En particulier, la mission de fin de vie du satellite SPOT-4 a permis d'augmenter la répétitivité des acquisitions à cinq jours [Hagolle et al., 2015b]. Cependant, cette expérience, nommée *Take-5*, n'a duré que six mois de février à juillet 2013. Afin d'obtenir une série temporelle décrivant le cycle de végétation sur une année culturale, la série temporelle de 2013 doit être complétée avec les acquisitions d'un autre capteur. Dans ces travaux, ce sont les acquisitions faites par le satellite Landsat-8 qui sont utilisées.

Landsat-8

Le satellite Landsat-8 a été lancé courant 2013 par la NASA, et est encore en orbite. Il permet l'acquisition d'images tous les seize jours avec une fauchée de 185 kilomètres. Il embarque deux capteurs :

— l'instrument OLI dont les caractéristiques sont décrites dans le Tableau 3.2,

— l'instrument infrarouge thermique *Thermal Infra-Red Sensor* (TIRS) pour des acquisitions à une résolution spatiale de 100 mètres.

Dans ses travaux, les bandes multi-spectrales, exceptée celle dédiée à la détection des cirrus, sont utilisées. L'intérêt des données Landsat-8 réside dans ses caractéristiques

Tableau 3.2 – Caractéristiques de l'instrument *Operational Land Imager* (OLI) embarqué sur le satellite Landsat-8.

	Canal	Résolution spatiale	Bande spectrale
Multispectral (XS)	Aérosol	30 m	$0.44 - 0.45\ \mu m$
	Bleu		$0.45 - 0.51\ \mu m$
	Vert		$0.53 - 0.59\ \mu m$
	Rouge		$0.64 - 0.67\ \mu m$
	Proche infra-rouge (PIR)		$0.85 - 0.88\ \mu m$
	Moyen infra-rouge 1 (MIR1)		$1.57 - 1.65\ \mu m$
	Moyen infra-rouge 2 (MIR2)		$2.11 - 2.29\ \mu m$
	Cirrus		$1.36 - 1.38\ \mu m$
Panchromatique	PAN	15 m	$0.52 - 0.90\ \mu m$

similaires à celles de Sentinel-2 : haute résolution spectrale avec sept bandes spectrales utilisées, et surtout haute répétitivité temporelle à 16 jours. Cependant, la résolution spatiale de 30 mètres est moins bonne que celles des données Sentinel-2.

La date du lancement de Landsat-8, février 2013, coïncide avec les dates de l'expérience SPOT-4-*Take-5*. La combinaison des deux capteurs permet alors pour 2013 d'obtenir une série temporelle d'images satellitaires couvrant quasiment une année et dont les caractéristiques sont proches de celles de Sentinel-2.

Sentinel-2

Les deux satellites Sentinel-2 ont été développés par l'ESA dans le cadre du projet européen Copernicus. À eux deux, ils permettent l'acquisition d'images sur l'ensemble des terres émergées tous les cinq jours. Le premier satellite Sentinel-2A a été lancé en juin 2015, tandis que son jumeau Sentinel-2B a été lancé en mars 2017.

La charge utile des deux satellites est composée de l'instrument *Multi-Spectral Instrument* (MSI) dont les caractéristiques sont détaillées dans le Tableau 3.3. L'imageur MSI a une fauchée de 290 kilomètres.

Tableau 3.3 – Caractéristiques de l'instrument *Multi-Spectral Instrument* (MSI) embarqué sur les deux satellites Sentinel-2.

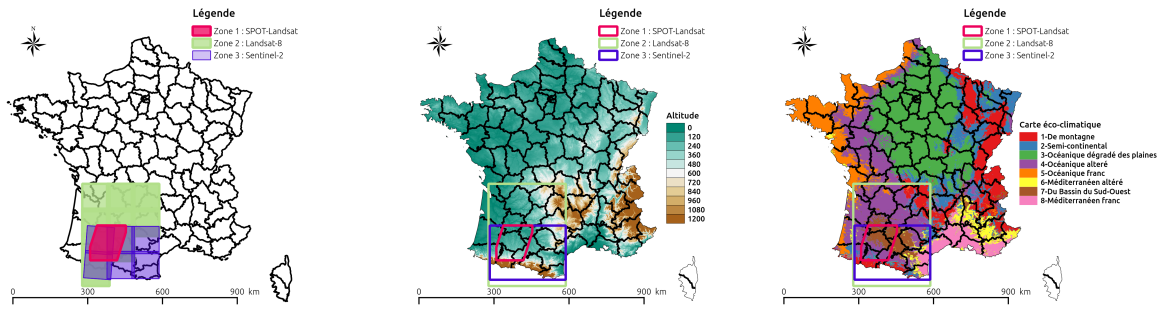
	Canal	Résolution spatiale	Bande spectrale
Multispectral (XS)	Aérosol	60 m	$0.43 - 0.45\ \mu m$
	Bleu	10 m	$0.46 - 0.52\ \mu m$
	Vert	10 m	$0.54 - 0.58\ \mu m$
	Rouge	10 m	$0.65 - 0.68\ \mu m$
	Red-edge 1	20 m	$0.70 - 0.71\ \mu m$
	Red-edge 2	20 m	$0.73 - 0.74\ \mu m$
	Red-edge 3	20 m	$0.77 - 0.79\ \mu m$
	Proche infra-rouge (PIR1)	10 m	$0.78 - 0.90\ \mu m$
	Proche infra-rouge (PIR2)	20 m	$0.85 - 0.87\ \mu m$
	Vapeur d'eau	60 m	$0.93 - 0.95\ \mu m$
	Moyen infra-rouge 1 (MIR1)	20 m	$1.57 - 1.66\ \mu m$
	Moyen infra-rouge 2 (MIR2)	20 m	$2.10 - 2.28\ \mu m$
	Cirrus	60 m	$1.37 - 1.39\ \mu m$

Dans le cadre de ces travaux, seulement les bandes à 10 et 20 mètres ont été utilisées. Les bandes à 60 mètres sont principalement utilisées en amont de la chaîne de classification, notamment pour réaliser les pré-traitements des images satellitaires.

3.1.2 Zones d'études

Au cours de ces travaux, trois zones d'études sont étudiées. À chaque zone est associée une série temporelle d'images satellitaires acquises par un ou plusieurs capteurs satellitaires. Ces trois zones sont situées dans le Sud-Ouest de la France.

La Figure 3.1 donne des informations sur ces trois zones d'études. En particulier, la Figure 3.1a permet de visualiser leur localisation avec les trois zones colorées (rouge, bleue, verte). La taille de ces zones d'étude est induite par l'emprise des données satellitaires. La Figure 3.1b montre l'altitude extraite d'un Modèle Numérique de Terrain (MNT) à une résolution spatiale de 25 mètres sur l'ensemble de la France. Les zones montagneuses sont représentées en marron, et les plaines en vert. La Figure 3.1b permet de mettre en évidence les fortes variations d'altitude pour les trois zones d'étude, notamment avec la présence des Pyrénées dans le Sud Ouest de la France. Par ailleurs, la Figure 3.1c montre les régions éco-climatiques telles que définies par Joly et al. [2010]. Les zones d'étude sont donc caractérisées principalement par deux zones éco-climatiques – océanique altéré (en violet) et du Bassin du Sud-Ouest (en marron) – avec des présences minoritaires de d'autres régions éco-climatiques.



(a) Localisation des trois zones d'étude (b) Altitude (Modèle Numérique de Terrain, 25 mètres) (c) Régions éco-climatiques

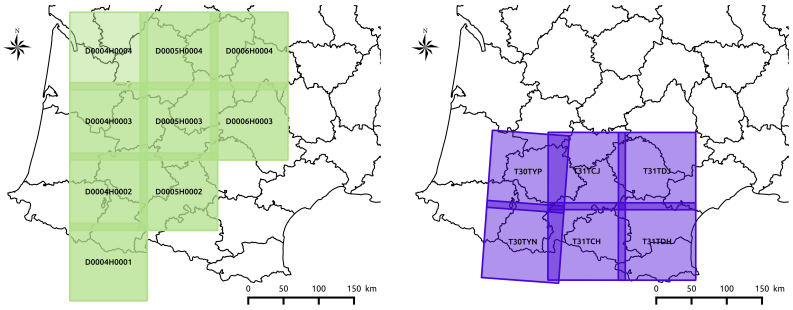
FIGURE 3.1 – Visualisation et caractéristiques des trois zones d'étude.

La première zone d'étude représentée par le carré rouge sur la Figure 3.1a est composée d'images SPOT-4 issues de l'expérience *Take-5*, et d'images Landsat-8 pour l'année 2013.

La deuxième zone d'étude, la plus grande, est représentée en vert sur la Figure 3.1a. Elle est composée uniquement d'images Landsat-8 acquises en 2013. Pour des raisons pratiques, les images sont distribuées par tuile. Comme le montre la Figure 3.2a, huit tuiles sont ici utilisées. Le nom des tuiles est donné selon la nomenclature de l'USGS. Cette zone d'étude est caractérisée par de fortes variations d'altitude avec la présence des Pyrénées, mais aussi du Massif Central comme montré par la Figure 3.1b. Elle est aussi caractérisée par une diversité de régions éco-climatiques montrée par la Figure 3.1c. Outre les climats du Bassin du Sud-Ouest (en marron) et océanique altéré (en violet), les climats suivants sont aussi bien présents : méditerranéen franc (en rose), méditerranéen altéré (en jaune), de montagne (en rouge) et semi-continental (en bleu).

La troisième zone d'étude représentée en bleu est composée d'images Sentinel-2 acquises fin 2015, et sur l'année 2016. Comme le montre la Figure 3.2b, six tuiles décrites selon la nomenclature de l'ESA sont utilisées.

Pour l'ensemble des trois zones d'étude, le détail des dates des images satellitaires utilisées est donné dans l'Annexe A.1.



(a) Tuiles Landsat-8 (nomenclature USGS)

(b) Tuiles Sentinel-2 (nomenclature ESA)

FIGURE 3.2 – Tuiles utilisées pour les données Landsat-8 et Sentinel-2.

3.2 Pré-traitements des données satellitaires

Lors de l'acquisition d'une image satellitaire optique, le capteur imageur mesure la quantité d'énergie réfléchie par la surface de la Terre à l'aide d'un radiomètre. Malheureusement, l'information reçue par le radiomètre n'est pas une mesure précise et uniforme. En effet, le signal reçu est perturbé par la présence d'erreurs, d'artefacts, de déformations due au processus d'acquisition.

Les pré-traitements des données satellitaires permettent alors d'obtenir des valeurs de réflectance corrigées de ces effets perturbateurs. Ils sont généralement divisés en trois catégories : 1) les corrections géométriques pour rendre superposables les images de la série, 2) les corrections radiométriques pour obtenir des valeurs de réflectance comparables entre les images de la série, et 3) la reconstruction des données manquantes pour faciliter l'utilisation des algorithmes de classification.

Dans le cas d'études de séries temporelles, les pré-traitements sont un enjeu important afin de pouvoir suivre l'évolution des réflectances d'une date à l'autre et de comparer les valeurs pixel à pixel. Par ailleurs, l'utilisation de différentes tuiles lors de traitements sur de grandes étendues requiert des valeurs de réflectance précises afin d'éviter des effets en bordure de tuile. Enfin, l'utilisation d'algorithmes de classification nécessite des vecteurs de variable qui ne contiennent pas de données manquantes. Les données manquantes apparaissent lors de la présence de nuages, de données saturées et des traitements sur de grandes étendues.

Dans un premier temps, les corrections géométriques et radiométriques sont décrites. Dans un deuxième temps, la problématique liée aux données manquantes est abordée. Une dernière partie détaille le processus de correction appliqué à l'ensemble des images utilisées.

3.2.1 Corrections géométriques et radiométriques

Les distributeurs d'images satellitaires appliquent généralement d'abord les corrections géométriques, puis les corrections radiométriques. En suivant cet ordre, les deux types de correction sont détaillés dans la suite. À titre d'exemple, l'Annexe A.4 donne les différents niveaux de pré-traitements appliqués aux images SPOT-5 et Sentinel-2.

Corrections géométriques

Les images issues du satellite ne sont pas directement superposables à une carte. En effet, des déformations géométriques sont introduites, dues au positionnement du satellite

sur son orbite, aux effets du relief terrestre, à la rotondité de la Terre et à sa rotation pendant la prise de vue. Des distorsions peuvent aussi être ajoutées en fonction des conditions d'acquisition, *e.g.* le mouvement du satellite et l'angle d'acquisition.

Les corrections géométriques sont appliquées afin de réduire l'ensemble de ces déformations introduites lors de l'enregistrement des images. Elles servent également à orthorectifier les images, *i.e.* les rendre superposables, et à les géo-référencer, *i.e.* leur donner une localisation en latitude et longitude précise.

Afin d'appliquer les corrections géométriques, un modèle géométrique est élaboré en fonction de la position du satellite, de ses lois d'attitude ou encore de l'angle de visée. Il permet de rendre superposables les images à une carte. Afin d'augmenter sa précision, des points d'appuis sont utilisés pour corriger les erreurs dues au positionnement du satellite sur son orbite et les effets du relief terrestre. Ils sont localisés en longitude et latitude grâce à une image de référence elle-même bien localisée au sol, et en altitude grâce à un MNT.

Corrections radiométriques

Le signal reçu par le radiomètre contient la lumière renvoyée par le paysage, mais aussi la lumière du soleil, les différentes réflexions atmosphériques et du bruit ajouté par le capteur. Les capteurs passifs ne peuvent donc pas mesurer de manière précise et uniforme la quantité d'énergie réfléchie par les objets situés à la surface de la Terre. D'une part, la lumière solaire qui éclaire ces objets est perturbée par la traversée de l'atmosphère et n'éclaire pas tous les objets sous un même angle [38]. D'autre part, la lumière réfléchie par les paysages est perturbée par la traversée de l'atmosphère.

Les corrections radiométriques consistent alors à transformer le signal reçu par le capteur en une valeur radiométrique la plus proche possible de celle mesurée au sol. L'objectif de ces corrections est donc de convertir les valeurs brutes des images, appelées comptes numériques, en réflectance *Top of Canopy* (TOC). La Figure 3.3 montre les différentes étapes des corrections radiométriques.

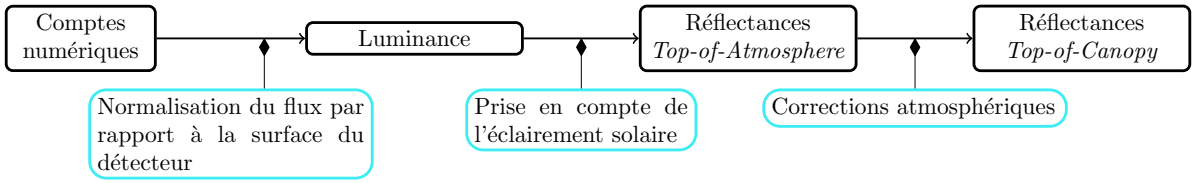


FIGURE 3.3 – Corrections radiométriques : des comptes numériques aux réflectances *Top-of-Canopy* (TOC).

.

La première étape consiste à convertir les comptes numériques en luminance. Cette dernière représente le flux lumineux normalisé par rapport à la surface du détecteur. La luminance est donc une mesure physique indépendante des caractéristiques du radiomètre utilisé.

La deuxième étape consiste à normaliser la luminance en réflectance *Top of Atmosphere* (TOA). Pour ce faire, l'angle d'acquisition du satellite et l'éclairement solaire incluant l'irradiance solaire et l'angle solaire zénithal est pris en compte.

38. Comme les données utilisées dans ces travaux sont acquises à angle constant et aux mêmes heures de passage, cette problématique est atténuée. Les différences radiométriques observées sont alors dues à l'acquisition de la série temporelle à différentes saisons.

La troisième et dernière étape consiste à transformer les réflectances TOA en réflectance TOC en corrigeant les effets du passage de la lumière réfléchie dans l'atmosphère [Hagolle et al., 2008]. Cette étape est connue comme celle des corrections atmosphériques. Elle nécessite de caractériser le plus précisément possible les conditions atmosphériques, comme la pression atmosphérique, le taux d'humidité, la quantité de vapeur d'eau ou encore l'épaisseur optique des aérosols. Comme les caractéristiques de l'atmosphère dépendent fortement des longueurs d'onde, et qu'elles varient dans le temps et l'espace, ces informations doivent être acquises par des capteurs spécifiques, pas toujours présents sur les zones étudiées. Ainsi, les corrections atmosphériques ne peuvent pas toujours être appliquées.

Finalement, une dernière étape, non-indiquée dans la Figure 3.3, consiste à corriger les variations d'éclairement dues au relief : les effets de pente. Ces corrections cherchent à réduire les différences de réflectance observées par exemple sur des versants opposés. Elles sont particulièrement intéressantes lors de traitements en zones montagneuses.

3.2.2 Données manquantes

Outre les déformations géométriques et les erreurs dues au processus d'acquisition, les données optiques sont perturbées par la présence de données manquantes. Ces données manquantes correspondent à des pixels pour lesquels les valeurs de réflectance de la surface au sol ne sont pas disponibles à cause soit de la couverture nuageuse et des problèmes de saturation, soit des dates d'acquisitions irrégulières.

Reconstruction des données manquantes

Concernant les données nuageuses et saturées, une première solution consiste à travailler avec toutes les images sans prendre de précaution par rapport aux données manquantes. Les vecteurs de variables sont alors erronés, *i.e.* certaines valeurs de réflectance ne représentent pas la surface au sol.

Une deuxième solution consiste à considérer seulement les images non-nuageuses et non-saturées en entrée du système de classification. Cette solution est sous-optimale : si l'image n'est pas entièrement nuageuse et saturée, alors l'information apportée par les données non-nuageuses et non-saturées est perdue. De plus, la série temporelle serait composée d'un faible nombre d'images car la majorité des images contient des nuages ou des pixels saturés.

Une troisième solution consiste alors à reconstruire les données manquantes en retrouvant les valeurs de réflectance : c'est le *gap-filling*. Cette reconstruction est réalisée en amont de la phase d'apprentissage. Pour ce faire, les données manquantes sont tout d'abord détectées.

De nombreuses méthodes de détection des données nuageuses existent dans la littérature. La méthode utilisée dans ces travaux tire bénéfice de la haute résolution temporelle des images acquises en repérant les variations de réflectance entre les acquisitions consécutives [Hagolle et al., 2010]. En effet, la présence de nuages dans les images modifie fortement les valeurs de réflectance dans la bande bleue, alors que les valeurs de réflectance sont normalement stables pour des acquisitions rapprochées. Cette approche est adaptée pour les images provenant d'acquisition à angle de vue constant comme Landsat-8 et Sentinel-2. La géométrie identique des images et l'heure de passage constante du satellite renforce la stabilité des valeurs de réflectance pour les données non-nuageuses et non-saturées entre deux acquisitions successives.

Une fois les données manquantes détectées, la reconstruction est réalisée dans ces travaux de thèse en appliquant une interpolation temporelle linéaire. Pour chaque donnée manquante, les valeurs de réflectance aux dates valides avant et après la date à reconstruire sont utilisées pour appliquer l'interpolation linéaire. Parfois une seule date est valide dans le profil. Dans ce cas là, la valeur de réflectance de cette date valide est répétée pour toutes les autres dates. Le choix de l'interpolation linéaire permet de gérer plus facilement les gros volumes de données induit par le traitement des séries temporelles d'images satellitaires à hautes résolutions [39]. Par ailleurs, l'utilisation de méthodes d'interpolation plus complexes comme les *splines* ou le filtre de Savitzy-Golay a montré peu de différence sur les performances obtenues en classification [Arnau Rodes, 2016].

La Figure 3.4 illustre le principe de reconstruction par interpolation temporelle linéaire pour un profil composé de huit dates représentant les valeurs de réflectance de la bande rouge. Le profil en bleu représente le profil réel pour lequel aucune donnée manquante n'est présente. Dans cet exemple, ce profil est perturbé par deux acquisitions à des dates nuageuses représentées par les lignes verticales en pointillé rouge (jour de l'an 200 et 248). Ainsi, le profil acquis par le capteur imageur serait celui représenté par les points rouges. La courbe en magenta représente alors le profil obtenu lors de la reconstruction des données manquantes par interpolation linéaire temporelle.

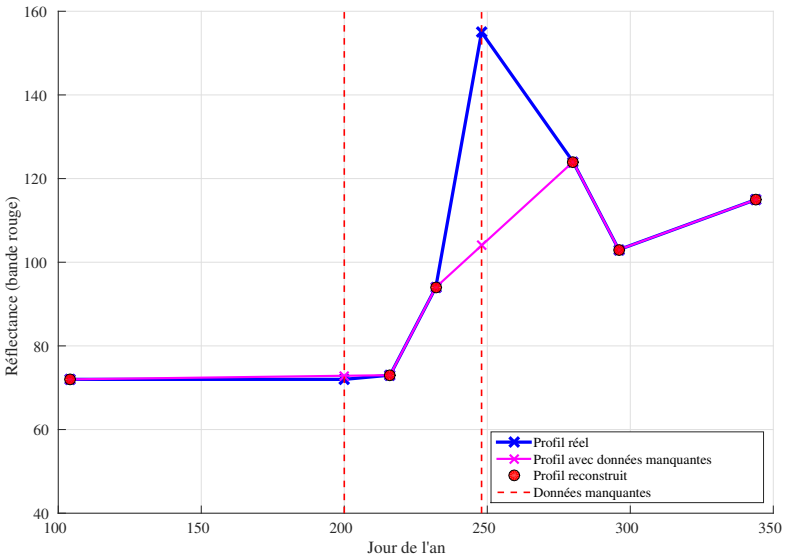


FIGURE 3.4 – Reconstruction de données manquantes par interpolation temporelle linéaire.

La Figure 3.4 montre que le jour de l'an 200 est bien reconstruit : faible écart entre le profil reconstruit en magenta et le profil réel en bleu. Dans ce cas, l'interpolation linéaire temporelle fonctionne bien puisque les valeurs de réflectance évoluent peu autour du jour de l'an 200. En revanche, la valeur de réflectance au jour de l'an 248 pour le profil reconstruit est d'environ 100, alors qu'en réalité la valeur de réflectance est presque de 160. Dans ce cas, les valeurs de réflectance du profil réel (courbe en bleu) changent rapidement (*e.g.* croissance de végétation). L'interpolation linéaire temporelle ne semble alors pas adaptée. Dans ces travaux, nous considérons que la forte revisite temporelle

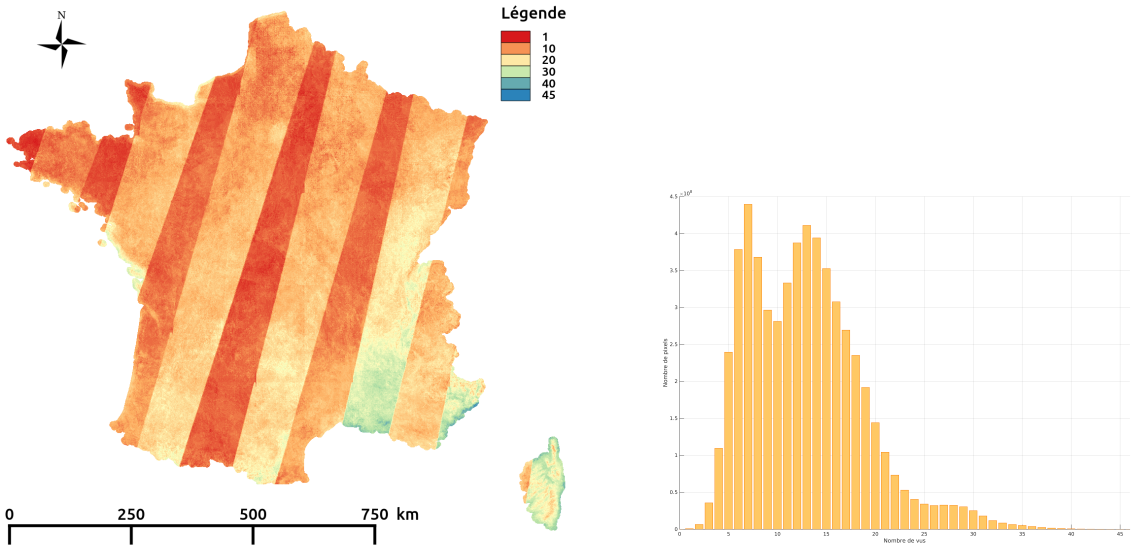
39. À titre d'exemple, une image Sentinel-2 téléchargée sur la Plateforme d'Exploitation des Produits Sentinel (PEPS) (`peps.cnes.fr`) – treize bandes spectrales corrigées géométriquement, en réflectance TOC et corrigées des effets de pente – couvrant 100×100 km.^{2} occupe 6.6 Go. d'espace disque. Et une série temporelle Sentinel-2 de la France pour une année, composée en moyenne de 35 images avec dix bandes spectrales échantillonnées à 10 mètres, occupent 19 To. sur l'espace disque.

offerte par les deux satellites Sentinel-2 permettra de minimiser le nombre de fois où ce cas se produit.

Harmonisation temporelle

En plus des données manquantes dues à la présence des nuages ou aux problèmes de saturation, une deuxième difficulté survient lors de traitements sur de grandes étendues. En effet, différents passages du satellite sont nécessaires pour cartographier de grandes étendues. Ainsi, toutes les zones ne sont pas acquises aux mêmes dates.

Afin d'utiliser les algorithmes de classification, les vecteurs de variables doivent avoir la même taille et chaque variable doit représenter la même information. Dans la réalité, tous les pixels ne sont pas vus au même moment et le même nombre de fois. Un pixel est vu lorsqu'il est ni nuageux ni saturé lors d'une acquisition. Afin d'illustrer les différences entre les pixels, la Figure 3.5 montre le nombre de vues de chaque pixel sur une série temporelle Sentinel-2A de 2016 (fin novembre 2015 à fin octobre 2016).



(a) Nombre de vues pour chaque pixel sur l'ensemble de la France.

(b) Histogramme associé.

Source : http://osr-cesbio.ups-tlse.fr/~oso/posts/2017-04-13-carte-s2-2016-corse/

FIGURE 3.5 – Visibilité des pixels pour une série temporelle d'images Sentinel-2 de fin novembre 2015 à fin octobre 2016.

Sur la Figure 3.5a, la succession des bandes claires et foncées s'explique par la trace du satellite Sentinel-2. Théoriquement, la résolution temporelles de Sentinel-2 est de dix jours. Comme les acquisitions se chevauchent, les zones représentées par les bandes jaunes orangées sont vues plus d'une fois tous les dix jours. Il existe aussi des différences entre régions dues aux conditions climatiques plus ou moins favorables lors des acquisitions. Les pixels du bassin méditerranéen sont les plus vus, tandis que ceux de Bretagne sont les moins vus.

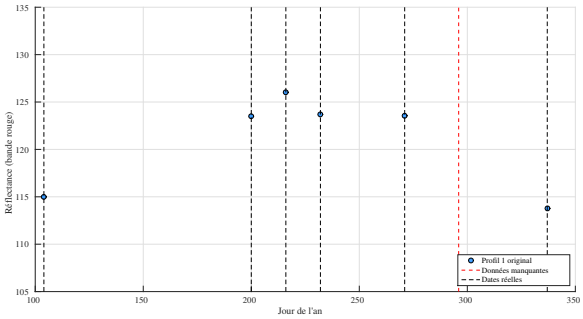
L'histogramme associé, Figure 3.5b, permet aussi de visualiser la répartition des pixels en fonction de leur visibilité. En moyenne, les pixels sont vus seulement une dizaine de fois pour 46 passages du satellite.

Même si les données nuageuses et saturées sont reconstruites par la procédure décrite précédemment, les vecteurs de variables auront des tailles différentes lors de l'utilisation

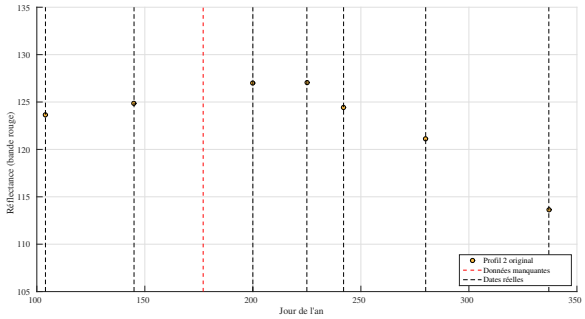
de plusieurs tuiles. Comme mentionné précédemment, ces différences sont problématiques puisque les algorithmes de classification ont besoin d'échantillons décrits par des vecteurs de variables de tailles identiques. Afin de s'affranchir de ce problème, il est nécessaire d'harmoniser temporellement les vecteurs de variables pour qu'ils représentent les mêmes dates. Pour ce faire, une interpolation temporelle linéaire est réalisée sur un même vecteur de dates. Comme montré dans Inglada et al. [2015], cette interpolation linéaire a peu de conséquences sur les performances des algorithmes de classification.

La Figure 3.6 montre deux profils de réflectance (bande rouge) acquis pour des dates différentes. La Figure 3.6a montre le premier profil pour neuf dates avec les ronds bleus. La ligne verticale en pointillé rouge correspond à une donnée manquante. De manière similaire, la Figure 3.6b montre un second profil représenté par les ronds jaunes pour des dates d'acquisition différentes que la Figure 3.6a. Une donnée est aussi manquante sur ce profil, représentée par la ligne verticale en pointillé rouge.

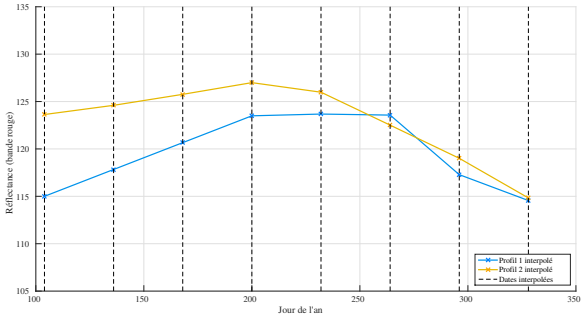
L'objectif est de classifier les deux échantillons décrits par les deux profils des Figures 3.6a et 3.6b avec le même algorithme de classification. Pour ce faire, les deux profils sont donc interpolés sur un nouveau vecteur de dates. Lors de cette interpolation, les données nuageuses et saturées ne sont pas prises en compte. La Figure 3.6c montre le résultat où le nouveau vecteur de dates est représenté par les lignes verticales en pointillé noir. La courbe jaune représente l'interpolation du profil de la Figure 3.6a, tandis que la courbe bleue représente l'interpolation de la Figure 3.6b.



(a) Profil 1 réel.



(b) Profil 2 réel.



(c) Profil 1 et 2 reconstruits sur un nouveau vecteur de dates.

FIGURE 3.6 – Reconstruction de données manquantes et utilisation de dates interpolées pour deux profils représentés par différentes dates.

3.2.3 Traitements appliqués

Dans l'ensemble de ces travaux de thèse, les images utilisées sont orthorectifiées, en réflectance TOC et corrigées des effets de pente.

Plus précisément, les images SPOT-4 et Sentinel-2 sont fournies par le pôle de données et de services surfaces continentales Theia, tandis que les images Landsat-8 par l'USGS. Les deux organismes ortho-rectifient les images, et effectuent les deux premières étapes des corrections radiométriques indiquées par la Figure 3.3. Ainsi, les images téléchargées sont corrigées géométriquement et en réflectance TOA.

Le passage en réflectance TOC est ensuite réalisé avec la chaîne de traitement *MACCS-ATCOR Joint Algorithm* (MAJA) [40] [Hagolle et al., 2015a]. Cette chaîne permet d'obtenir les images en réflectance TOC corrigées des effets de pente, les masques des nuages (avec leurs ombres) et les masques des données saturées.

Le traitement des données manquantes est ensuite spécifique à chaque zone d'étude. Il dépend principalement du nombre d'images et de tuiles utilisé. Par ailleurs, l'utilisation d'images à différentes résolutions spatiales nécessite un ré-échantillonnage afin d'avoir des images dans la série qui se superposent et de même résolution spatiale. Le reste des pré-traitements appliqués est détaillé par zone d'étude.

La première zone d'étude, représentée en rouge sur la Figure 3.1a, est composée de quinze images SPOT-4 et huit images Landsat-8 (Tableau A.1) qui ont des résolutions spatiales de 20 et 30 mètres respectivement. Les images Landsat-8 sont ré-échantillonnées à 20 mètres en utilisant une méthode d'interpolation spatiale bicubique. Ensuite, les données nuageuses et saturées sont interpolées par la méthode décrite à la Section 3.2.2. Pour cette zone d'étude, les dates réelles sont gardées.

La deuxième zone d'étude, en vert sur la Figure 3.1a, est composée uniquement d'images Landsat-8 pour lesquelles toutes les bandes multi-spectrales sont gardées sauf celle utilisée pour la détection de cirrus (Tableau 3.2). Comme le montre le Tableau A.2 (Annexe A.1), les dates d'acquisition ne sont pas identiques entre les huit tuiles. De plus, certaines tuiles contiennent moins de dates. Par exemple, la tuile D0004H0001 (dernière colonne) qui couvre une partie des Pyrénées est composée uniquement de neuf images à cause des conditions d'acquisition très défavorables (forte présence de neige et nuages). Comme expliqué à la Section 3.2.2, les vecteurs de variables sont harmonisées temporellement en appliquant une interpolation linéaire. Après cette opération, les échantillons présents sur les huit tuiles ont les mêmes vecteurs de dates. Comme la résolution temporelle théorique d'une série Landsat-8 est de seize jours, il a été décidé d'interpoler les données Landsat-8 sur un vecteur de dates allant du 19 avril 2013 au 29 novembre 2013 avec un pas de seize jours.

Concernant la troisième zone d'étude, en bleu sur la Figure 3.1a, seules les bandes spectrales à une résolution spatiale de 10 et 20 mètres de Sentinel-2 sont gardées. Les bandes à 20 mètres sont ré-échantillonnées à 10 mètres en utilisant une méthode d'interpolation spatiale bicubique. Les images acquises sur six tuiles à différentes dates sont interpolées sur un même vecteur de dates. Dans ce cas là, le vecteur de dates choisi va du 30 novembre 2015 au 15 octobre 2016 avec un pas de dix jours.

3.3 Données de référence utilisées

Outre les données satellitaires, les données de référence sont aussi nécessaires en entrée du système de classification. Comme le montre la Figure 2.1, elles sont utilisées pour l'apprentissage du classifieur et aussi pour l'évaluation des performances du système.

Dans le cadre de ces travaux de thèse, l'ensemble des données de référence est extrait de trois bases de données :

40. Anciennement *Multi-Sensor Atmospheric Correction and Cloud Screening* (MACCS).

1. Registre Parcellaire Graphique (RPG),
2. OCcupation des Sols à Grande Échelle (OCS-GE),
3. données terrains.

Les processus de collecte et de production de ces données sont décrits dans la suite pour chacune des données de référence.

3.3.1 Registre Parcellaire Graphique

Dans le cadre de la Politique Agricole Commune (PAC), les agriculteurs français doivent déclarer leurs parcelles agricoles afin de déterminer le montant des aides européennes. Ces informations sont consignées et mises à jour annuellement dans une base de données : le RPG. La base de données sous format vectoriel décrit 28 classes de végétation dont les cultures mais aussi les prairies ou encore les vignes (Annexe A, Tableau A.6). Actuellement, l'année N est mise à disposition des institutions publiques à la fin de l'année $N + 1$.

Afin de compléter le RPG, les agriculteurs doivent renseigner en ligne les cultures plantées sur des fonds d'images aériennes soit à l'îlot soit à la parcelle. L'îlot représente un ensemble contigu de parcelles culturales exploitées par un même agriculteur. Un îlot peut donc contenir différents types de cultures. L'exploitant agricole déclare le pourcentage de chacune des cultures par îlot. Depuis 2015/2016, les déclarations sont réalisées à la parcelle. De plus, les surfaces non agricoles comme les haies, les mares, les bosquets, ou les bâtiments sont mieux pris en compte que les années précédentes.

Dès qu'un agriculteur perçoit une aide pour une de ces parcelles, il doit déclarer toute l'exploitation, *i.e.* l'ensemble de ces parcelles. Cependant, comme tous les exploitants ne perçoivent pas l'aide européenne, le RPG ne référence pas de manière exhaustive toutes les cultures.

Dans le cadre des différentes études, seuls les îlots purs ont été gardés. Le Tableau 3.4 montre toutes les classes d'occupation des sols utilisées. Au cours des études, les années 2013 et 2014 du RPG ont été utilisées pour les trois zones d'étude.

Tableau 3.4 – Données du Registre Parcellaire Graphique (RPG) utilisées pour les trois zones d'études.

Blé
Colza
Maïs
Orge
Tournesol
Prairies (permanentes et temporaires)
Vignes
Vergers

3.3.2 OCcupation des Sols à Grande Échelle

En France, l'IGN produit depuis 2013 la base de données nationale OCS-GE. Cette donnée a la particularité d'être segmentée en quatre couches – occupation des sols, usages des sols, attributs morphologiques et éléments de caractérisation – afin notamment de bien différencier l'occupation des sols de l'usage. La production se fait en fusionnant toutes les

bases de données disponibles, puis en corrigeant les erreurs et en complétant l'information manquante par photo-interprétation. Pour l'ensemble de la France, le temps de production est estimé à trois ans.

Concernant la couche de l'occupation des sols, une nomenclature hiérarchique compatible avec CLC [41] est utilisée. La Figure 1.2e montre un extrait sur la zone urbaine de Toulouse, tandis que la Figure 3.7 montre un extrait sur une étendue plus grande aux alentours de Tarbes. Sur cette dernière Figure, les formations herbacées sont par exemple extraites du RPG.

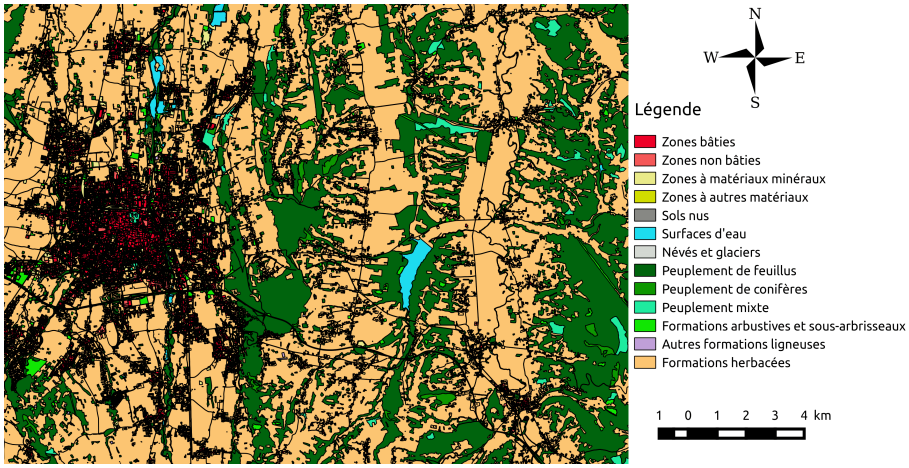


FIGURE 3.7 – OCcupation des Sols à Grande Échelle, Tarbes 2013.

Cette base de données s'appuie sur la mise à jour d'anciennes base de données. Les parties mise à jour et complétion sont pour le moment réalisées manuelles. Cependant, l'objectif de l'IGN est d'automatiser ces étapes afin de réduire les temps de production [Gressin, 2014].

3.3.3 Données terrain

La dernière base de données est composée de données collectées pendant les campagnes terrains réalisées au CESBIO. Dans cette approche, des experts du terrain vont sur place pour noter l'occupation des sols de différentes parcelles agricoles. L'emplacement et l'étendue des parcelles sont définies en amont à partir d'une image à très haute résolution spatiale.

Les données terrains utilisées représentent deux années : 2013 et 2016. Concernant l'année 2013, 1300 parcelles couvertes par une seule culture et représentant des conditions de développement similaires ont été suivies. Elles ont été visitées trois à quatre fois pendant des moments clés du cycle phénéologique. Concernant l'année 2016, plus de 3000 parcelles agricoles ont été suivies. Pour cette année, au minimum deux passages sont effectués.

41. CORINE *Land Cover* : base de données d'occupation des sols couvrant l'Europe et dont la dernière version date de 2012 (page 6)

Troisième partie

Stabilité et robustesse des algorithmes de classification

Chapitre 4

Études des algorithmes de classification sur de grandes étendues

Sommaire

4.1 Données en entrée du système de classification 68
4.1.1 Choix des variables en télédétection 68
4.1.2 Primitives spectrales et temporelles 69
4.1.3 Vecteurs de variables étudiés 73
4.2 Présentation des expérimentations 74
4.2.1 Images satellitaires et données de référence 75
4.2.2 Stratégie d'échantillonnage pour l'apprentissage supervisé . . . 76
4.2.3 Configuration des paramètres des classifieurs 78
4.3 Résultats des expérimentations 79
4.3.1 Comparaison des *Random Forests* et des *Support Vector Machines* 80
4.3.2 Sensibilité aux paramètres du *Random Forest* 85
4.3.3 Comparaison des différents vecteurs de variables 88
4.3.4 Stabilité du *Random Forest* sur de grandes étendues 89
4.4 Conclusion 91

Les séries temporelles d'images satellitaires à hautes résolutions sont une entrée essentielle dans les chaînes de traitement automatique dédiées à la cartographie de l'occupation des sols. Comme vu dans la section 1.4.1, le processus de classification dépend de l'algorithme de classification utilisé ainsi que des données fournies en entrée du système de classification.

Dans le contexte de la cartographie de l'occupation des sols, les algorithmes de classification doivent savoir exploiter les variabilités spectro-temporelles fournies par les données satellitaires. En effet, la variabilité des paysages induite par les différents climats, les activités humaines, ou encore les différences pédologiques complexifient les apparences des classes d'occupation des sols. En outre, cette variabilité est exacerbée lors de la classification sur de grandes étendues. Ainsi, ce chapitre s'intéresse à la classification de séries temporelles sur de grandes étendues.

La première section se focalise sur les données satellitaires à fournir en entrée du système de classification. La deuxième section présente les données utilisées ainsi que les configurations des études menées dont les résultats sont détaillés dans une quatrième partie. Enfin, une dernière partie conclut ce chapitre.

4.1 Données en entrée du système de classification

Dans le cadre de l'apprentissage supervisé, les données fournies en entrée du système de classification sont les échantillons d'apprentissage (Figure 2.1). Ces derniers sont décrits par un vecteur de variables extrait des données satellitaires, ainsi que d'une classe d'occupation des sols extraite des données de référence.

Bien qu'il soit possible d'utiliser seulement les bandes spectrales [Akbari et al., 2006], il est courant d'enrichir cette information de variables calculées à partir des images satellitaires : ce sont les primitives. Elles permettent d'améliorer la distinction des occupations des sols, et sont généralement divisées en trois catégories : 1) les primitives spectrales qui combinent l'information de différentes bandes spectrales, 2) les primitives spatiales qui utilisent le voisinage (*e.g.* géométrie ou texture) [Haralick, 1979; Lv et al., 2014; Trias-Sanz, 2006], et 3) les primitives temporelles extraites à partir de séries temporelles d'images satellitaires. L'apport des primitives spatiales sur les performances de classification a été démontré de nombreuses fois, grâce à leur prise en compte du voisinage et de leurs effets filtrages. L'information spatiale, et de manière plus générale contextuelle, est notamment très importante pour les images à très haute résolution spatiale [Blaschke, 2010]. Elle l'est toutefois moins pour des images à 20 ou 30 mètres de résolution. De plus, la grande nouveauté des séries temporelles d'images satellitaires repose sur la dimension temporelle. Le choix a donc été fait de se focaliser pour ces études sur l'apport de l'information spectrale et temporelle.

Ce chapitre discute en particulier du choix du vecteur de variables à fournir en entrée du système de classification. Une première partie est dédiée aux stratégies mises en place dans la littérature pour construire le vecteur de variables. Puis, une seconde partie décrit les primitives spectrales et temporelles.

4.1.1 Choix des variables en télédétection

Plusieurs stratégies sont possibles afin de choisir les variables à extraire des images satellitaires. Une première stratégie consiste à utiliser seulement certaines variables en lien avec le problème de classification, *e.g.* des indices pour caractériser la végétation [Xiao et al., 2005]. Cependant, cette stratégie peut conduire à l'utilisation d'un jeu de variables sous optimal si les primitives extraites ne sont pas suffisantes.

Une seconde stratégie consiste alors à calculer des centaines de primitives [Dalla Mura et al., 2010; Huang and Zhang, 2013]. Malheureusement une majorité des algorithmes de classification est sensible à la malédiction de la dimension, et donc incapable de gérer de grands volumes de données (Section 2.1). Par conséquent, des méthodes de réduction et de sélection de données ont été utilisées.

Les méthodes de réduction de dimension comme l'Analyse par Composantes Principales (ACP) transforment l'information physique des images satellitaires afin de réduire le nombre de variables et de supprimer l'information redondante [Potgieter et al., 2007]. Une autre approche consiste à sélectionner le meilleur sous-ensemble de variables parmi celles calculées [Camps-Valls et al., 2011], en se basant par exemple sur le score d'importance des variables calculé par le RF [Gressin et al., 2013; Paget et al., 2015]. Une autre stratégie de sélection consiste à s'appuyer sur des connaissances *a priori* du problème. Par exemple, il est courant lors de la classification de classes de cultures de sélectionner des images à deux saisons différentes afin de maximiser les différences spectrales entre les classes, *e.g.* cultures d'hiver et cultures d'été [Rodríguez-Galiano et al., 2012; Rogan et al., 2002]. Dans ce cas, une connaissance experte est nécessaire afin de déterminer le

meilleur sous-ensemble de variables à sélectionner.

En plus d'éviter la malédiction de la dimension, la réduction du nombre de variables permet aussi de réduire les temps d'apprentissage. Pourtant, ces méthodes sont coûteuses et complexes à mettre en place de manière opérationnelle. La sélection de primitives est souvent spécifique au jeu de données étudiées. D'une part, les variables sélectionnées pour une année ne formeront peut être pas le meilleur ensemble pour l'année suivante. D'autre part, les variables sélectionnées dans une zone éco-climatique et topographique spécifique ne conviendront peut être pas pour une zone d'étude différente [Arnau Rodes, 2016].

Le développement d'algorithmes de classification moins sensibles aux espaces de grande dimension permet d'envisager des stratégies plus faciles à automatiser et plus adaptées aux problèmes de classification sur de grandes étendues. Avec l'utilisation de ces classifieurs, une des stratégies la plus simple consiste à garder toute l'information temporelle et spectrale contenue dans les séries temporelles, *i.e.* chaque bande spectrale de toutes les images satellitaires. Les profils spectraux-temporels extraits pour différentes classes d'occupation des sols peuvent alors être suffisants (Figure 1.7). Cependant, la contribution de certaines primitives est encore incertaine dans le contexte de la classification de séries temporelles d'images satellitaires à hautes résolutions [Gómez et al., 2016]. De plus, l'ajout de primitives peut aider à gérer la forte variabilité des paysages lors d'études sur de grandes étendues. Par exemple, l'ajout d'information sur la phénologie des plantes peut permettre de mieux prendre l'ensemble des apparences d'une classe de culture, *e.g.* l'ajout de la longueur du plateau peut être une indication importante si les dates de levée sont par exemple décalées.

Pour des algorithmes moins sensibles aux problèmes de grande dimension, il pourrait être donc intéressant de calculer un grand nombre de primitives et laisser les algorithmes de classification choisir l'information la plus pertinente dans cet espace [Vieira et al., 2012]. Cette stratégie est adoptée dans ces travaux. Cependant, comme l'introduction d'un grand nombre de variables augmente les temps d'apprentissage, et nécessite aussi des fortes capacités de stockage, une étude préliminaire est effectuée pour déterminer le meilleur jeu de variables permettant d'atteindre la précision maximale tout en permettant une automatisation du processus.

4.1.2 Primitives spectrales et temporelles

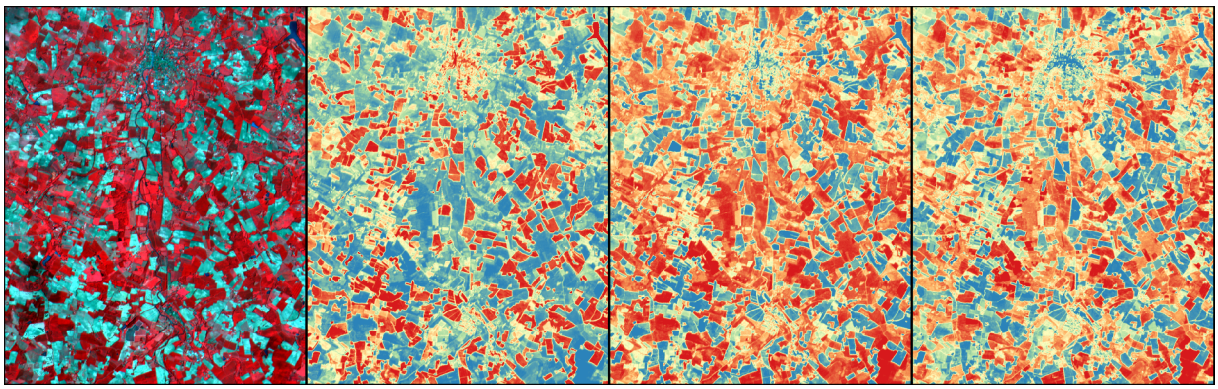
L'ajout d'informations supplémentaires aux bandes spectrales est donc courant dans la littérature sur la classification de données satellitaires : utilisation de données exogènes comme le MNT [Franklin, 1998; Rodríguez-Galiano et al., 2012], calcul de variables biophysiques [Waldner et al., 2015b], introduction de règles expertes [Osman et al., 2015; Waldner et al., 2015a], ou encore approches basées objets [Blaschke, 2010]. Bien que l'ensemble de ces approches ait montré un apport potentiel pour la cartographie de l'occupation des sols, de nombreuses questions méthodologiques persistent. À titre d'exemple, considérons l'ajout des trois composantes du MNT – altitude, pente et exposition – à une série temporelle d'images satellitaires composées de dix images représentées par dix bandes spectrales. La dimension du vecteur de variable est donc de 300 dimensions où l'information du MNT est sous-représentée, et donc a une influence limitée [Lucas et al., 2007]. Utiliser le MNT comme connaissance *a posteriori* en fixant des seuils sur certaines classes est une approche compliquée puisque les seuils doivent continuellement être mis à jour.

Cependant, ces informations ainsi que la prise en compte du contexte spatial ou de manière plus générale les approches basées objets [Blaschke, 2010] n'ont pas été étudiées

dans le cadre de ce manuscrit. La suite de cette section est donc dédiée aux primitives spectrales et temporelles.

Primitives spectrales

Le calcul de primitives spectrales combine différentes bandes spectrales en s'appuyant sur les propriétés physiques des bandes spectrales pour améliorer la discrimination entre les classes. Une des propriétés physiques bien connue, et exploitée, concerne la végétation qui a une réponse faible dans le visible due à la photo-synthèse, mais une réponse forte dans le proche infra-rouge à cause de la structure des plantes [Tucker, 1979]. Ainsi, l'indice de NDVI – différence normalisée entre les bandes spectrales rouge et infra-rouge – permet d'augmenter la discrimination des classes de végétation avec le sol nu et l'eau. Il est de plus corrélé à des propriétés de la végétation comme l'état de santé et la phénologie des plantes, la biomasse ou encore à la notion de rendement [DeFries and Townshend, 1994; Senf et al., 2015]. De manière similaire, l'indice de *Normalized Difference Water Index* (NDWI) permet d'augmenter la discrimination de l'eau du sol et de la végétation [McFeeters, 1996], et l'indice de *Normalized Difference Built-up Index* (NDBI) permet d'améliorer la détection de zones urbaines [Zha et al., 2003]. La Figure 4.1 montre une image satellitaire SPOT-4 en fausse couleur ainsi que les indices de NDVI, NDWI et NDBI associés.



(a) Image originale (b) NDVI (c) NDWI (d) NDBI

FIGURE 4.1 – Exemple d'indices spectraux. La couleur rouge représente des valeurs faibles de l'indice, et le bleu des valeurs élevées.

De plus, des indices spectraux spécifiques aux capteurs ont été proposés dans plusieurs travaux. Par exemple, les coefficients *tasseled cap* calculés pour les images Landsat permettent la caractérisation de zones de cultures agricoles à travers trois indices décorrélés (*brightness*, *greenness*, and *wetness*) [Baig et al., 2014; Crist and Cicone, 1984; Huang et al., 2002b; Kauth and Thomas, 1976].

Ainsi, l'utilisation de primitives peut aider les algorithmes de classification. Pour les algorithmes de classification linéaires comme le SVM linéaire, la frontière de décision recherchée est obtenue par des combinaisons linéaires des variables. L'utilisation de primitives calculées par des combinaisons non-linéaires des bandes spectrales peut donc faciliter la recherche de la frontière de décision. Pour les algorithmes de classification non-linéaires, la frontière de décision est simplifiée par l'ajout d'information non-linéaire, et donc la capacité de généralisation peut être augmentée, *e.g.* augmentation de la marge du SVM.

Dans la littérature, des centaines de primitives spectrales ont été proposées notamment pour l'étude de la végétation [Mróz and Sobieraj, 2004; Silleos et al., 2006; Yeom

Tableau 4.1 – Description des primitives spectrales utilisées.

Nom	Formule	Commentaires
Normalized Difference Vegetation Index	$\text{NDVI} = \frac{PIR-R}{PIR+R}$	Rouse et al. [1973]
Normalized Difference Water Index[*]	$\text{NDWI} = \frac{V-PIR}{V+PIR}$	McFeeters [1996]. Gao [1996] a proposé la même année la formulation suivante : NDWI = −NDBI. Les deux expressions sont utiles pour la détection du contenu liquide de la végétation. Wilson and Sader [2002] ont aussi nommé la formulation de Gao *Normalized Difference Moisture Index* (NDMI).
Modified Normalized Difference Water Index[*]	$\text{MNDWI} = \frac{V-MIR}{V+MIR}$	Xu [2006] corrige le défaut du NDWI de McFeeter en supprimant le bruit dû à la présence de structure urbaine. Cette nouvelle formulation est proposée avec la bande MIR du satellite Landsat-5 (1.55-1.75 μm.). Dans cette étude, l'indice MNDWI est calculé avec $MIR1$, et MNDWI2 avec $MIR2$.
Normalized Difference Built-up Index[*]	$\text{NDBI} = \frac{SWIR1-NIR}{SWIR1+NIR}$	Zha et al. [2003]
Modified Normalized DifferenceBuilt-up Index	$\text{MNDBI} = \frac{MIR2-PIR}{MIR2+PIR}$	Shingare et al. [2014]
Index-based Built-up Index[*]	$\text{IBI} = \frac{\text{NDBI}-\frac{\text{SAVI}+\text{MNDWI}}{2}}{\text{NDBI}+\frac{\text{SAVI}+\text{MNDWI}}{2}}$	Xu [2008]
Brightness	$Brightness = 0.3561 \times B + 0.3972 \times V + 0.3904 \times R + 0.6966 \times PIR + 0.2286 \times MIR1 + 0.1596 \times MIR2$	Les coefficients *tasseled cap* sont des combinaisons linéaires entre les bandes spectrales, proposés initialement pour les images Landsat [Crist and Cicone, 1984; Kauth and Thomas, 1976]. Ils décrivent les caractéristiques des parcelles agricoles, et de manière générale de la végétation. Bien que Baig et al. [2014] ait proposé les coefficients *tasseled cap* pour les images Landsat-8, ceux calculés pour les images Landsat-7 sont ici utilisés [Huang et al., 2002a].
Greenness	$Greenness = -0.3344 \times B - 0.3544 \times V - 0.4556 \times R + 0.6966 \times PIR - 0.0242 \times MIR1 - 0.2630 \times MIR2$	
Wetness	$Wetness = 0.2626 \times B + 0.2141 \times V + 0.0926 \times R + 0.0656 \times PIR - 0.7629 \times MIR1 - 0.5388 \times MIR2$	
Brillance[*]	norme des bandes spectrales	Brillance pour les bandes communes à SPOT-4 et Landsat-8, et Brillance2 pour l'ensemble des bandes Landsat-8.

Ubl pour la bande ultra-bleue (aérosol) ; B pour bleue ; V pour verte ; R pour rouge ; PIR pour proche infra-rouge ; MIR pour moyen infra-rouge
[*] primitives spectrales qui peuvent être calculées pour les images SPOT-4

et al., 2013]. Suite à plusieurs études préliminaires, une dizaine d'indices spectraux ont été sélectionnés pour ces travaux. Le détail de toutes ces primitives spectrales est donné par le Tableau 4.1.

Primitives temporelles

Contrairement aux primitives spectrales, les primitives temporelles ont reçu moins d'intérêt à cause du manque de séries temporelles d'images satellitaires à hautes résolutions. Cependant, certaines primitives temporelles ont prouvé leur capacité à améliorer les performances de classification, particulièrement pour la cartographie de la végétation [Jia et al., 2014a; Pittman et al., 2010; Valero et al., 2016]. Cette partie se propose de les décrire.

Calculables grâce au fort temps de revisite des nouvelles données satellitaires, les primitives temporelles extraites des séries temporelles peuvent aider le classifieur pour discriminer les occupations des sols qui évoluent au cours du temps comme les cultures [Waldner et al., 2015a]. L'extraction de primitives temporelles est donc réalisée généralement à partir des profils temporels d'indices de végétation comme le NDVI. Deux stratégies sont traditionnellement adoptées.

La première stratégie consiste à extraire des valeurs significatives des profils temporels comme le maximum, l'amplitude, ou encore la moyenne [Arvor et al., 2011; Pittman et al., 2010; Valero et al., 2016]. Calculées à partir de l'indice de végétation, ces valeurs sont généralement représentatives du cycle phénologique des plantes, *e.g.* un maximum de NDVI correspond la croissance maximale de la plante. Dans le cadre de ces travaux, un total de huit variables statistiques extraites à partir des profils de NDVI sont calculées : moyenne, minimum, maximum, amplitude, moyenne olympique (suppression des valeurs extrêmes), médiane et écart-type.

Une seconde stratégie consiste à extraire les paramètres phénologiques en modélisant les profils temporels d'indices de végétation. Basée sur ce principe, TIMESAT est la méthode la plus connue [Jönsson and Eklundh, 2002, 2004]. Le logiciel associé à TIMESAT permet le calcul de neuf paramètres phénologiques déduits du profil de NDVI comme le début de croissance, la valeur du pic ou la longueur du plateau. Au moins trois années de données acquises à dates régulières sont nécessaires pour 1) estimer le nombre de saisons, et 2) ajuster le profil de NDVI à un modèle. TIMESAT a principalement été utilisé pour des séries temporelles à basse résolution spatiale, notamment les données MODIS [Jia et al., 2014a]. D'autres méthodes d'ajustement ont aussi été proposées pour les données satellitaires [Beck et al., 2006; Eerens et al., 2014; Zhang et al., 2003].

Dans ces travaux, une modélisation à partir d'une double logistique développée au CESBIO est utilisée [Inglada, 2016]. Contrairement aux travaux de TIMESAT, cette méthode ne nécessite pas un échantillonnage régulier des données sur trois années. Le profil de NDVI à un temps t donné est donc modélisé par une double logistique utilisant six paramètres (équation (4.1)) :

$$\widetilde{NDVI}(t) = A\left(\frac{1}{1+e^{\frac{x_0-t}{x_1}}} - \frac{1}{1+e^{\frac{x_2-t}{x_3}}}\right) + B, \tag{4.1}$$

avec A l'amplitude, B le minimum, x_0 et x_2 les points d'inflexion, et x_1 et x_3 les taux d'accroissement et de décroissement de la courbe aux points d'inflexion x_0 et x_2 respectivement.

Plus précisément, l'algorithme effectue une modélisation du profil de NDVI pour l'ensemble de la saison. Il estime les paramètres du profil en deux étapes afin d'être robuste à la présence de plusieurs cycles de végétation dans la saison. A et B sont calculés à partir des valeurs de NDVI extraites de la série temporelle d'images satellitaires. Puis, les quatre paramètres restants x_i, $0 \leq i \leq 3$ sont estimés en utilisant l'algorithme itératif de Levenberg-Marquardt – algorithme non-linéaire de moindres carrés [Levenberg, 1944;

Marquardt, 1963].

La Figure 4.2 montre l'approximation d'un profil de NDVI (losange rouge) par une double logistique (courbe bleue). Les deux droites en cyan représentent les tangentes aux points d'inflexion x_0 et x_2.

De plus, quatre paramètres $t_i,\ 0 \leq i \leq 3$ sont aussi calculés pour décrire le cycle phénologique. Ces paramètres sont aussi indiqués sur la Figure 4.2. Les dates de début et de fin de croissance t_0 et t_3 sont les dates pour lesquelles les tangentes aux points d'inflexion x_0 et x_2 intersectent la valeur minimale B. De la même manière, t_1 et t_2 sont les dates pour lesquelles les tangentes aux points d'inflexion x_0 et x_2 intersectent la valeur maximale $A+B$. La différence $t_2 - t_1$ correspond à la longueur du plateau. Les équations de (4.2) à (4.5) indiquent comment l'ensemble de ces paramètres est calculé. $\widetilde{NDVI}'$ est la dérivée première par rapport au temps t.

$$t_0 = x_0 + \frac{B - \widetilde{NDVI}(x_0)}{\widetilde{NDVI}'(x_0)} \tag{4.2}$$

$$t_1 = \frac{A + B - (\widetilde{NDVI}(x_0) - x_0\widetilde{NDVI}'(x_0))}{\widetilde{NDVI}'(x_0)} \tag{4.3}$$

$$t_2 = \frac{A + B - (\widetilde{NDVI}(x_2) - x_2\widetilde{NDVI}'(x_2))}{\widetilde{NDVI}'(x_2)} \tag{4.4}$$

$$t_3 = x_2 + \frac{B - \widetilde{NDVI}(x_2)}{\widetilde{NDVI}'(x_2)} \tag{4.5}$$

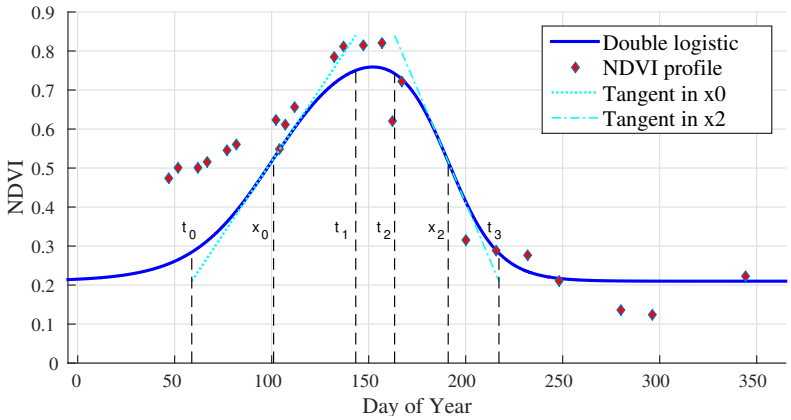


FIGURE 4.2 – Modélisation d'un profil de *Normalized Difference Vegetation Index* (NDVI) par une double logistique.

Finalement, un total de dix sept primitives est alors calculé à partir de la modélisation par la double logistique : A, B, $A+B$, $x_i,\ 0 \leq i \leq 3$, $\widetilde{NDVI}'(x_0)$, $\widetilde{NDVI}'(x_2)$, $t_i,\ 0 \leq i \leq 3$, $t_2 - t_1$, $t_3 - t_0$, une estimation de l'aire sous la courbe $A(t1 - t0)/2.0 + A(t3 - t2)/2.0 + A(t2 - t1)$, et l'erreur commise par l'algorithme des moindres-carrés.

4.1.3 Vecteurs de variables étudiés

L'objectif ici est de déterminer la contribution des primitives spectrales et temporelles lors de l'utilisation de séries temporelles comme celles fournies par Sentinel-2. La

stratégie la plus simple, qui consiste à utiliser l'ensemble des bandes spectrales seules, est étudiée. Toutefois, la contribution de l'information spectro-temporelle offerte par les nouvelles données satellitaires doit être étudiée. Pour ce faire, les primitives spectrales et temporelles détaillées précédemment (Section 4.1.2) sont ajoutées aux bandes spectrales [42] pour constituer de nouveaux vecteurs de variables. Les primitives spectrales sont calculées pour chaque image de la série temporelle, tandis que les primitives temporelles sont calculées sur le profil temporel du NDVI. Comme l'indice de NDVI est l'un des indices les plus utilisés, notamment lors d'études sur la cartographie de la végétation à partir de séries temporelles à basse résolution spatiale, il est aussi analysé séparément des autres primitives spectrales. Finalement, un total de cinq vecteurs de variables est proposé :

1. bandes spectrales seulement (BS)
2. bandes spectrales et primitives spectrales (BS-PS)
3. bandes spectrales et NDVI (BS-NDVI)
4. bandes spectrales et primitives temporelles (BS-PT)
5. bandes spectrales, NDVI et primitives temporelles (BS-NDVI-PS)

4.2 Présentation des expérimentations

Dans ces travaux, l'objectif est de tester différentes configurations de classification en utilisant toute l'information contenue dans la série temporelle. Ainsi, les jeux de variables décrits dans la Section 4.1.3 sont utilisés.

Le SVM et le RF ont été étudiés de nombreuses fois dans la littérature. Ainsi, le SVM est connu pour avoir des paramètres difficiles à régler (choix du noyau et valeur du paramètre de régularisation). De plus, il est sensible à la présence de données imparfaites, et son temps de calcul augmente avec le nombre de classes dû à la multiplication des apprentissages dans les approches « un-contre-tous » et « un-contre-un ». De même, la complexité du RF augmente avec le nombre d'arbres et le nombre d'échantillons. Par contre, l'algorithme est plus stable et plus simple à paramétrer. L'algorithme du RF semble donc plus adapté dans le contexte de la classification de séries temporelles d'images satellitaires sur de grandes étendues en quasi-temps réel.

L'objectif général est d'évaluer les performances du RF sur de grandes étendues en utilisant différents jeux de variables extraits de séries temporelles d'images satellitaires. Afin de réaliser cette évaluation, quatre études sont proposées. Dans un premier temps, le choix du RF est discuté en le comparant à celui du SVM. Dans un deuxième temps, une analyse de sensibilité sur les paramètres du RF est réalisée. Ensuite, l'utilisation des différents jeux de variables est comparée. Enfin, la stabilité de l'algorithme de classification en fonction des variables utilisées est testée sur une plus grande étendue.

Cette section décrit les configurations des différentes études. La première partie est dédiée à la description des images satellitaires et des données de référence utilisées. La deuxième partie aborde la question de la stratégie d'échantillonnage visant à partitionner la donnée de référence en deux sous-ensembles utilisés pour l'apprentissage et l'évaluation. Enfin, la dernière partie est consacrée à la configuration des paramètres des deux algorithmes de classification étudiés.

42. Dans le cas d'une série multi-capteurs, le nombre de bandes spectrales peut différer. Le choix a été fait dans les études d'ajouter toutes les bandes spectrales disponibles pour chaque capteur, pas seulement celles communes.

4.2.1 Images satellitaires et données de référence

Afin de réaliser les études sur la robustesse du RF, les deux zones d'études représentées en rouge et verte sur la Figure 3.1a sont utilisées. Pour rappel, la première zone d'étude est composée d'images SPOT-4 et Landsat-8, et la seconde zone uniquement d'images Landsat-8. Chaque image est corrigée au niveau de réflectance TOC et des effets de pente.

La Figure 4.3 détaille la distribution temporelle des images utilisées pour les deux capteurs SPOT-4 et Landsat-8. Les acquisitions SPOT-4 sont disponibles à partir de mi-février jusqu'à fin juin, tandis que les images Landsat-8 sont principalement acquises sur la fin de l'année où la présence de nuages et le temps de revisite de 16 jours du satellite diminuent la fréquence des images disponibles. Cette série temporelle permet de décrire le cycle de végétation sur quasiment une année, facilitant notamment la reconnaissance des classes de végétation comme les cultures.

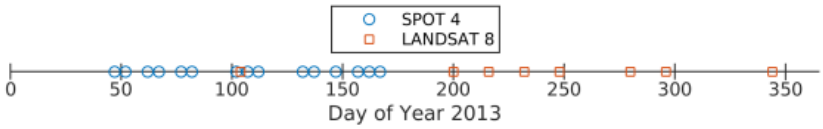


FIGURE 4.3 – Distribution temporelle des images SPOT-4 et Landsat-8 pour la première zone d'étude en fonction du jour de l'année (*Day of Year* en anglais).

Les données de référence utilisées pour cette zone sont principalement extraites des données terrain et de la base de données OCS-GE. Seules les vignes et les vergers sont extraits du RPG. Au total, dix-huit classes sont représentées (première colonne du Tableau 4.3). Cette zone d'étude couvre 16 902 km^2 (en bleu transparent sur la Figure 4.4). Les données de référence issues des données terrain sont spatialement localisées dans le Sud du Gers, et les données de l'OCS-GE sont seulement présentes sur le département des Hautes-Pyrénées. Seulement les classes vergers et vignes sont donc représentées au Nord de cette zone.

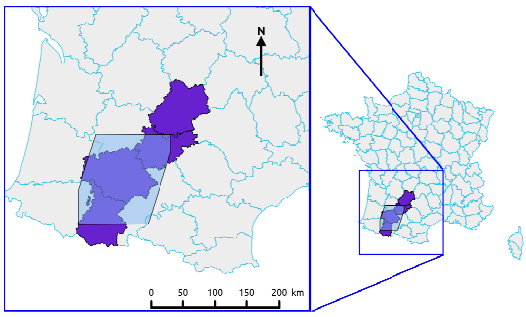


FIGURE 4.4 – Localisation des zones d'étude. En bleu transparent, la première zone d'étude. En violet foncé, la seconde zone d'étude.

Une seconde zone d'étude s'étendant sur 19 785 km^2 permet de réaliser une étude sur une plus grande étendue. Cette seconde zone est représentée en violet foncé sur la Figure 4.4. Contrairement à la Figure 3.1a où la taille de la zone est issue de la taille des images satellitaires, la superficie est ici donnée par la localisation des données de références utilisées. Sur cette seconde zone d'étude, seules des images Landsat-8 sont utilisées. Malheureusement le satellite ayant été lancé au cours de l'année 2013, la série temporelle commence le 19 avril et se termine le 29 novembre 2013.

Les données de référence utilisées représentent douze classes et sont extraites du RPG et de la base de données OCS-GE [43]. Le changement de nomenclature s'explique par l'agrandissement de la zone d'étude. D'une part, l'utilisation du RPG au lieu des données terrain diminue le nombre de classes de culture, et d'autre part les classes perméables, sol nu, forêts mixtes et vergers ne sont pas représentées sur l'ensemble de cette zone d'étude, et ont donc été supprimées.

La première zone d'étude est utilisée pour la comparaison du RF avec le SVM, l'analyse de sensibilité du paramétrage du RF ainsi que l'étude des différents vecteurs de variables. La seconde zone d'étude plus grande est alors utilisée pour évaluer la stabilité du classifieur. Le Tableau 4.2 met en évidence la taille des vecteurs de variables pour les deux zones d'études en fonction des primitives calculées.

Tableau 4.2 – Nombre total de variables pour chaque zone d'étude en fonction des jeux de variables utilisés. La première zone est composée de 15 images SPOT-4 et 8 images Landsat-8, tandis que la deuxième est composée d'images Landsat-8 interpolées sur 15 dates.

	BS	**BS-PS**	**BS-NDVI**	**BS-PT**	**BS-NDVI-PT**
1ère zone d'étude	116	302	139	141	164
2ème zone d'étude	105	285	120		

BS : bandes spectrales, PS : primitives spectrales, PT : primitives temporelles,
NDVI : *Normalized Difference Vegetation Index*

4.2.2 Stratégie d'échantillonnage pour l'apprentissage supervisé

L'échantillonnage des données de référence consiste à obtenir deux sous-ensembles d'échantillons indépendants utilisés d'une part pour l'apprentissage, et d'autre part pour l'évaluation des différents algorithmes de classification.

Un des choix les plus importants concerne le nombre d'échantillons à sélectionner par classe. Cette problématique est complexe. Le Chapitre 2 indique que le nombre d'échantillons d'apprentissage doit être suffisant afin d'éviter la malédiction de la dimension, *e.g.* sélectionner le nombre d'échantillons en suivant la règle des « $30p$ » avec p la dimension du vecteur de variables (Section 2.1).

Dans le contexte de la cartographie de l'occupation des sols, les différentes classes sont rarement représentées de manière équivalente. Ainsi, il est courant d'avoir un nombre déséquilibré d'échantillons entre classes.

Une stratégie évidente est de sélectionner le nombre d'échantillons d'apprentissage en suivant la distribution naturelle des classes dans la donnée de référence, *i.e.* les classes les plus représentées auront un plus grand nombre d'échantillons d'apprentissage. Cependant, certains classifieurs sont sensibles au problème fortement déséquilibré, *i.e.* le classifieur prédit seulement la classe majoritaire [Menardi and Torelli, 2014; Sun et al., 2009]. L'étude de Khoshgoftaar et al. [2007] montre que cette sensibilité diffère en fonction des classifieurs considérés. Par exemple, les arbres de décision binaires sont peu performants sur un problème déséquilibré. Les classes minoritaires seront moins présentes dans les nœuds terminaux. Si le vote majoritaire est utilisé pour définir la classe prédite par le nœud terminal, la classe minoritaire aura peu de chance de l'emporter.

43. Les données de référence OCS-GE étaient initialement disponibles que sur la partie Nord du département Hautes-Pyrénées. Après les études menées sur la première zone d'étude, les données sur les département du Gers, du Tarn-et-Garonne et du Lot ont été mises à disposition.

Afin de remédier à ces problèmes, des stratégies de sur- et de sous-échantillonnage ont été proposées dans la littérature [Chawla, 2005; Estabrooks et al., 2004]. Spécifiquement aux RF, deux solutions pour travailler sur les problèmes déséquilibrés ont été proposées : 1) sélectionner le même nombre d'échantillons par classe, et 2) donner plus d'importance aux classes minoritaires en pondérant le critère de Gini et le vote des nœuds terminaux [Chen et al., 2004; Thomas et al., 2006].

Mellor et al. [2015] montrent aussi que déséquilibrer le problème en faveur des classes les plus difficiles, peut améliorer les résultats de la classification. Malheureusement, les classes les plus difficiles sont généralement aussi les moins représentées. Au contraire, Colditz [2015] recommande d'utiliser un nombre d'échantillons par classe par rapport à la proportion des classes sur le terrain. Cette directive peut être difficile à appliquer si la donnée de référence est incomplète ou trop ancienne, car la distribution des échantillons par classe est alors biaisée, et non représentative de la réalité. Les recommandations de [Weiss and Provost, 2003], bien que plus nuancées, sont aussi de privilégier la distribution naturelle des données. Cependant, les auteurs montrent que choisir le même nombre d'échantillons par classe peut être favorable aux arbres de décision binaire.

Les résultats concernant le nombre d'échantillons d'apprentissage par classe à sélectionner sont donc divergents. Dans les travaux de cette thèse, cette problématique n'a pas été spécifiquement abordée. Il a été choisi de favoriser des problèmes équilibrés à l'apprentissage d'une part pour ne pas désavantager les algorithmes du RF et du SVM sensibles aux problèmes déséquilibrés [Akbani et al., 2004; Mellor et al., 2015], et d'autre part pour ne pas donner d'*a priori* à la distribution des classes.

La même problématique se pose pour les échantillons test car les mesures de performance, comme l'OA, sont influencées par un nombre d'échantillons déséquilibrés (Section 2.5). Cependant, les performances sont évaluées, dans ces travaux, sur l'ensemble des échantillons test disponibles afin de prendre en compte la distribution naturelle des classes de la zone d'étude [Congalton and Green, 2008].

La sélection des deux ensembles, apprentissage et test, se fait aléatoirement en veillant à respecter le principe d'indépendance : les échantillons test ne doivent pas être utilisés pour l'apprentissage. En télédétection, il est courant que les données de référence soient sous format vectoriel et composées de polygones représentant des objets homogènes ayant une même occupation des sols, *e.g.* une parcelle agricole. Pour s'assurer de l'indépendance des échantillons, la stratégie la plus simple consiste à ne pas utiliser des échantillons test appartenant aux mêmes polygones que les échantillons d'apprentissage.

En prenant en compte ces deux considérations – indépendance et nombre d'échantillons d'apprentissage et test par classe –, des stratégies d'échantillonnage spécifiques aux deux zones d'études ont été proposées. Les approches sont différentes car les deux zones d'étude ne sont pas identiques, et surtout les configurations des études sont différentes. Les stratégies adoptées sont décrites dans la suite, le Tableau 4.3 indique les nombres d'échantillons d'apprentissage et test pour les deux zones.

Première zone d'étude

Pour la première zone d'étude, une première étape consiste à diviser en deux ensembles disjoints de même taille les polygones de la donnée de référence. Dans le sous-ensemble apprentissage, une deuxième étape consiste à sélectionner aléatoirement le même nombre d'échantillons d'apprentissage par classe. Cette sélection aléatoire est réalisée au niveau pixel, et est donc indépendante des polygones. Pour les classes sous-représentées, le nombre maximal d'échantillons d'apprentissage disponible est utilisé. Aucune stratégie de sur- ou sous-échantillonnage n'est utilisée. Concernant les échantillons test, tous les

pixels disponibles sont sélectionnés. Cette stratégie est répétée cinq fois afin de s'assurer que les résultats ne soient pas trop optimistes ou pessimistes en fonction des échantillons d'apprentissage et test sélectionnés. Ces tirages aléatoires permettent aussi d'évaluer statistiquement les résultats en calculant les intervalles de confiance (Section 2.5.2) [44].

Pour l'apprentissage, le nombre d'échantillons est tout d'abord fixé à 5000 par classe. Cependant, l'influence sur les performances de classification du nombre d'échantillons d'apprentissage par classe a été analysée avec 750, 2000 et 10 000 échantillons. Le Tableau 4.3 montre le nombre d'échantillons total utilisé par classe pour chacune de ces configurations.

Seconde zone d'étude

La seconde zone d'étude est utilisée seulement pour évaluer la stabilité de l'algorithme de classification pour des zones spatialement éloignées de la zone d'apprentissage. Pour ce faire, la zone d'apprentissage est spatialement localisée dans un cercle de 15 kilomètres de rayon. Elle est représentée par le cercle rouge sur la Figure 4.5. Les différents zones d'évaluation utilisées sont quant à elle représentées par les rectangles bleus, de 15 $\times$ 20 km.^{2} distant de 15 kilomètres les uns des autres. Les matrices de confusion seront calculées dans chacun des dix-neuf rectangles afin d'évaluer l'influence sur les performances de classification de la distance de ces zones à la zone d'apprentissage.

Un total de 15 000 échantillons d'apprentissage par classe, quand c'est possible, est aléatoirement sélectionné dans le cercle rouge [45]. Afin de limiter les temps d'analyse, 35 000 échantillons par classe, quand c'est possible, sont sélectionnés à l'extérieur de la zone d'apprentissage mais dans la zone d'étude (en gris foncé sur la Figure 4.5). Puis, les échantillons tests sont extraits pour les dix-neuf zones d'évaluation (rectangles bleus sur la Figure 4.5). Le Tableau 4.3 montre les nombres d'échantillons d'apprentissage et test utilisés. Pour cette seconde zone d'étude, l'échantillonnage n'est pas réitéré plusieurs fois puisque l'ensemble des échantillons disponibles dans la zone d'apprentissage est quasiment utilisé, et que les échantillons d'apprentissage sont spatialement indépendants des échantillons tests.

4.2.3 Configuration des paramètres des classifieurs

Afin de comparer les meilleurs résultats obtenus par les deux algorithmes de classification sélectionnés, SVM et RF, les valeurs de leurs hyper-paramètres sont optimisées à l'aide d'une grille de recherche.

Pour le SVM, le noyau gaussien est choisi [46]. Il requiert l'optimisation de deux hyper-paramètres : le paramètre de régularisation C ainsi que l'écart-type de la gaussienne γ (Section 2.4.1). L'optimisation est réalisée par valida-

44. Les polygones ayant différentes tailles, la séparation en deux sous-ensembles de polygones implique des variations dans le nombre d'échantillons, *i.e.* pixels, disponibles d'apprentissage et test pour les différents tirages aléatoires. Afin de garder un nombre d'échantillons identiques, le nombre d'échantillons minimal disponible pour chaque classe est d'abord identifié. Ensuite, le tirage aléatoire est appliqué aux échantillons test pour se placer dans ce cas le plus défavorable. Cette information n'est pas explicitée puisque ce tirage aléatoire n'écarte que quelques échantillons test, et permet seulement d'assurer un nombre constant d'échantillons test pour les différents tirages.

45. Comme la zone d'apprentissage est spatialement localisée, il a été choisi d'augmenter le nombre d'échantillons d'apprentissage par rapport à la première zone d'étude afin de prendre en compte les différentes apparences des classes présentes.

46. Le choix du noyau découle d'études préliminaires (*benchmarking*) à ces travaux, et n'est donc pas discuté.

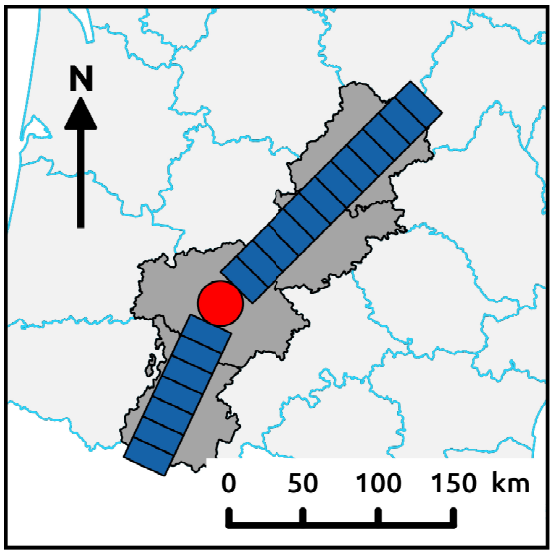


FIGURE 4.5 – Seconde zone d'étude en gris foncé avec les échantillons d'apprentissage spatialement localisés dans le cercle rouge, et les échantillons test localisés dans les zones rectangulaires bleues.

tion croisée sur dix partitions : $C = \{0,1; 0,5; 2,5; 12,5; 62,5; 312,5\}$, et $\gamma = \{10^{-5}; 1,50.10^{-4}; 2,25.10^{-3}; 3,38.10^{-2}; 5,51.10^{-1}\}$. Les valeurs de C sont donc obtenues sur une grille logarithmique pour cinq itérations avec une valeur minimale de 0.1 et un pas logarithmique de 5. De même, les valeurs de γ sont obtenues sur une grille logarithmique pour quatre itérations avec une valeur minimale de 10^{-5} et un pas logarithmique de 15. Dans le cas du SVM, les vecteurs de variables sont normalisés en soustrayant la moyenne et en divisant par l'écart-type pour chaque variable. Cette standardisation assure que la distance à l'hyperplan ne soit pas dominée par une seule variable ayant une forte dynamique [Han et al., 2011].

Pour le RF, les quatre hyper-paramètres – nombre d'arbres K, nombre de variables sélectionnées aléatoirement à chaque nœud m, profondeur maximale max_depth et nombre minimal d'échantillons par nœud $min\_samples$ – sont aussi optimisés en sélectionnant la combinaison qui atteint le meilleur OA moyen calculé sur les cinq jeux de tests pour chaque vecteur de variables[47]. Les valeurs testées sont les suivantes : $K = \{50; 100; 150; 200; 400\}$, $m = \left\{2; \sqrt{p}; p/3; p/2; p\right\}$ avec p la dimension du vecteur de variables, $max\_depth = \{10; 25; 50\}$, et $min\_samples = \{1; 10; 25; 50; 70\}$. Une étude spécifique sera aussi dédiée à l'influence du paramétrage du RF sur les performances de la classification.

4.3 Résultats des expérimentations

Cette partie présente l'évaluation des performances du RF lors de la classification sur de grandes étendues en utilisant différents jeux de variables. Dans un premier temps, le choix du RF est discuté en le comparant avec l'algorithme du SVM. Dans un deuxième temps, une analyse de sensibilité sur les paramètres du RF est réalisée. Ensuite, l'utilisation des différents jeux de variables est comparée. Enfin, la stabilité de l'algorithme de classification en fonction des variables utilisées est testée sur une plus grande étendue en

47. L'optimisation des paramètres du RF n'est donc pas indépendante des échantillons test. Cependant, une étude sur l'erreur OOB (Section 2.4.2) permet d'obtenir les mêmes configurations optimales.

Tableau 4.3 – Nombre d'échantillons d'apprentissage et validation pour différentes configurations.

	1ère zone d'étude					2ème zone d'étude	
Nombre d'échantillons	A 750	A 2000	A 5000	A 10000	T	A	T
Imperméables	750	2000	5000	10 000	10 321	2968	13 280
Perméables	750	1045	1045	1045	1045	-	-
Sol nu	750	2000	2493	2493	2493	-	-
Eau	750	2000	5000	7630	7630	899	16 383
Feuillus	750	2000	5000	10 000	202 829	15 000	14 851
Conifères	750	2,000	5,000	10 000	26 113	2179	22 801
Mixtes	750	2000	5000	10 000	24 454	-	-
Arbustes	750	2000	5000	10 000	26 519	906	24 709
Blé	750	2000	5000	10 000	20 923	15 000	13 330
Colza	750	2000	4350	4350	4350	8 412	10 645
Maïs	750	2000	5000	10 000	30 037	8065	12 762
Orge	750	1556	1556	1556	1556	10 720	12 577
Soja	750	2000	2268	2268	2268	-	-
Sorgho	257	257	257	257	257	-	-
Tournesol	750	2000	5000	10 000	10 730	15 000	12 840
Prairies	750	2000	5 000	10 000	11 756	15 000	13 203
Vergers	750	2000	5000	10 000	18 120	-	-
Vignes	750	2000	5000	10 000	52 188	639	2480
Total	**13 007**	**32 858**	**71 969**	**129 599**	**453 589**	**94 788**	**157 831**

A : apprentissage
T : test

contraignant spatialement la localisation des échantillons d'apprentissage (Section 2.3).

4.3.1 Comparaison des *Random Forests* et des *Support Vector Machines*

Afin de comparer les deux algorithmes de classification, trois études sont menées : 1) comparaison des performances de classification pour deux jeux de variables, 2) sensibilité des algorithmes au nombre d'échantillons d'apprentissage, et 3) comparaison des temps de calcul.

Précision et influence des données en entrée

Cette étude compare les performances du SVM et du RF pour deux jeux de variables différents : BS (le cas le plus simple) et BS-PS (le cas avec le plus grand nombre de primitives) (Section 4.1.3). Le Tableau 4.4 montre les précisions (UA), rappels (PA) et F-Scores (F) par classe ainsi que l'OA avec l'intervalle de confiance à 95 % obtenus pour les jeux de variables. Les valeurs en gras représentent les meilleurs résultats pour les F-Scores et l'OA.

Une première analyse de l'OA indique que le RF obtient de meilleurs résultats que le SVM pour les deux jeux de variables, et plus précisément le RF combiné avec les variables BS-PS obtient le meilleur OA. Cependant, les différences d'OA entre les deux classifieurs ne sont pas toujours observées en analysant les F-Scores. Comme indiqué dans la Section 2.5, l'OA est une mesure globale fortement affectée en cas de problème déséquilibré. Par exemple, les échantillons de feuillus comptent pour la moitié des échantillons test (Tableau 4.3). La différence de F-Score de 8 % pour cette classe entre le RF et le SVM avec le jeu de données BS-PS explique donc en grande partie la différence observée dans l'OA. Par ailleurs, la différence d'OA observée en fonction du jeu de variables utilisé est plus élevée pour le SVM, mais reste tout de même assez faible (inférieur à 2 %).

Tableau 4.4 – Précisions, rappels et F-Scores par classe et *Overall Accuracy* (OA) moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le RF et le SVM. Les données en entrée sont BS et BS-PS.

		RF BS			**SVM BS**			**RF BS-PS**			**SVM BS-PS**		
K, m, max_depth, $min\_samples$	C, γ	400, 10 50, 10			$2,5$, $2,25e^{-3}$			400, 17, 50, 10			$2,5$, $1,5e^{-4}$		
		UA	PA	F	UA	PA	F	UA	PA	F	UA	PA	F
Imperméables		87,4	76,7	90,4	89,8	88,9	89,2	86,6	93,7	89,9	92,0	93,5	**92,6**
Perméables		76,7	34,8	47,8	69,75	32,0	43,8	73,1	34,1	46,4	69,7	54,7	**61,2**
Sol nu		49,5	52,6	49,9	44,0	28,8	32,6	55,2	56,9	**55,3**	47,4	42,8	43,6
Eau		98,5	97,1	97,7	99,8	77,5	87,2	98,8	99,6	**99,2**	99,8	86,3	92,4
Feuillus		93,3	82,1	87,2	92,8	75,8	83,3	93,9	83,2	**88,1**	90,7	72,2	80,2
Conifères		59,8	67,4	63,2	51,5	55,2	52,8	62,5	67,2	**64,6**	53,5	56,8	54,4
Mixtes		21,2	31,6	24,9	19,2	31,1	23,6	24,3	35,8	**28,1**	13,1	26,8	17,5
Arbustes		73,1	88,4	80,0	39,0	92,8	54,7	75,6	90,0	**82,0**	65,7	90,2	75,8
Blé		91,0	92,8	91,9	90,1	83,6	86,7	91,6	92,9	**92,2**	90,3	92,5	91,4
Colza		90,5	94,6	92,5	94,3	64,1	76,2	91,1	95,3	93,2	92,9	94,5	**93,7**
Maïs		94,9	91,9	93,4	93,1	80,2	86,1	94,7	92,8	**93,8**	93,2	92,0	92,6
Orge		69,7	47,9	56,7	68,3	19,9	30,6	71,0	51,6	**59,7**	61,2	43,9	51,1
Soja		82,2	69,9	75,1	90,1	31,1	45,9	86,6	75,1	**80,1**	85,8	67,8	74,9
Sorgho		15,8	1,5	2,7	10,0	0,5	0,9	35,2	1,4	2,6	24,7	3,1	**5,3**
Tournesol		79,9	89,5	84,4	79,3	70,0	74,2	83,7	90,8	**87,1**	82,2	89,0	85,3
Prairies		68,8	83,7	75,4	74,2	73,7	73,9	69,2	84,6	76,0	75,9	83,8	**79,6**
Vergers		92,3	88,0	90,1	96,3	87,7	91,8	94,1	89,6	91,8	95,5	94,1	**94,8**
Vignes		96,3	94,8	95,5	98,2	93,3	95,7	96,8	95,5	96,2	98,2	96,6	**97,4**
OA		82,1 ± 3,6			75,52 ± 2,44			**83,3 ± 3,9**			77,1 ± 5,2		

UA : *user's accuracy* (précision) ; PA : *producer's accuracy* (rappel)
BS : bandes spectrales ; PS : primitives spectrales
Random Forest (RF) : K, le nombre d'arbres ; m, le nombre de primitives sélectionnées aléatoirement à chaque noeud ; max_depth, la profondeur maximale des arbres ; $min\_samples$, le nombre minimal d'échantillons par nœud pour continuer les divisions
Support Vector Machine (SVM) : C, le paramètre de régularisation ; γ, l'écart-type du noyau gaussien

Une seconde analyse par classe montre que les variations de F-Score pour des combinaisons « classifieur - données en entrée » peuvent dépasser 20 % (*e.g.* l'orge et le maïs), dues entre autre à une configuration déséquilibrée dans les échantillons test qui affectent les valeurs de F-Score. Par exemple, le nombre très déséquilibré entre les échantillons des surfaces perméables et imperméables engendre de grandes variations de F-Score entre le RF et le SVM pour les variables BS-PS.

L'analyse des classes avec moins de 5000 échantillons d'apprentissage (Tableau 4.3) montre qu'il y a un entre-deux avec la littérature sur les problèmes déséquilibrés : parfois le SVM est meilleur que le RF [Lin and Chen, 2013], d'autres fois c'est le contraire [Khalilia et al., 2011]. Si le nombre d'échantillons est vraiment trop bas, comme pour le sorgho, les précisions sont très faibles pour les deux classifieurs (sous 10 %).

Outre le problème du nombre déséquilibré d'échantillons, les différences entre classifieurs sont aussi d'ordre thématique : le SVM est meilleur pour les surfaces urbaines, les prairies, les vignes et les vergers, tandis que le RF est plus précis sur les forêts et majoritairement sur les cultures. Cependant, lorsque le RF est meilleur, la différence de F-Score est généralement plus élevée que lorsque le SVM est meilleur.

Dans la littérature, de nombreuses études ont comparé le RF avec le SVM soulignant des performances comparables entre les deux classifieurs lors d'utilisation de différentes données de télédétection [Dalponte et al., 2013; Duro et al., 2012; Ghosh et al., 2014; Hasan et al., 2012; Meyer et al., 2016; Nery et al., 2016; Waske and van der Linden, 2008]. Le Tableau 4.4 montre une légère amélioration avec l'utilisation du RF probablement due à sa capacité à prendre en compte des données représentées dans des espaces de grandes dimensions (Tableau 4.2). En effet, l'algorithme du RF sélectionne à chaque nœud la meilleure variable pour faire sa division. Cette opération peut être vue comme une sélection interne de variables facilitant la construction de la règle de décision en minimisant l'effet des variables inutiles et bruitées. Bien qu'une partie de la littérature mette en avant l'insensibilité du SVM au phénomène de Hugues, les résultats de cette étude et de d'autres travaux récents diffèrent [Foody et al., 2016; Gressin et al., 2013; Löw et al., 2013].

Influence du nombre d'échantillons d'apprentissage

Le Tableau 4.4 montre les résultats du RF et du SVM obtenus pour le même nombre d'échantillons d'apprentissage. Cependant, ce dernier peut modifier les performances des deux classifieurs. Ainsi, les Tableaux 4.5 et 4.6 montrent les précisions, rappels, F-Scores et OA moyens obtenus pour le SVM et le RF respectivement en utilisant 750 - 2000 - 5000 et 2000 - 5000 - 10000 échantillons par classe respectivement (Tableau 4.3). Les variables utilisées sont BS-PS puisqu'elles donnaient les meilleurs résultats dans l'étude précédente.

Le SVM n'est pas entraîné avec 10000 échantillons par classe à cause des temps d'apprentissage trop long pour cette configuration. Par ailleurs, le RF n'est pas appris avec 750 échantillons d'apprentissage par classe puisque ces performances augmentent avec le nombre d'échantillons. Une étude avec 750 échantillons par classe pour le RF n'est donc pas nécessaire.

Le Tableau 4.5 montre que le SVM obtient le meilleur OA avec 2000 échantillons d'apprentissage par classe. Cependant, l'analyse des F-Score par classe révèle que les meilleurs résultats sont généralement obtenus avec 5000 échantillons d'apprentissage par classe. La différence d'environ 4 % dans le F-Score des feuillus entre les deux configurations, entraîne la différence observée dans l'OA. Bien que les résultats soient moins bons avec 750 échantillons d'apprentissage par classe, l'amélioration obtenue en augmentant le nombre d'échantillons d'apprentissage n'est pas si importante.

Selon le Tableau 4.6, le RF obtient son meilleur OA et ses meilleurs F-Scores avec 10000 échantillons d'apprentissage par classe. Cependant, comme pour le SVM, les différences de précision observées en augmentant le nombre d'échantillons d'apprentissage sont faibles. De plus, n'utiliser que 2000 échantillons d'apprentissage par classe est bénéfique pour certaines classes, *e.g.* les surfaces perméables. Cette configuration diminue l'écart entre les échantillons de la classe surfaces perméables (classe minoritaire avec environ 1000 échantillons d'apprentissage) et ceux des autres classes (Tableau 4.3), conduisant à une augmentation du F-Score de la classe surfaces perméables. La même analyse peut être faite pour la classe orge qui sera plus clairement distinguée de la classe blé en limitant le nombre d'échantillons d'apprentissage par classe à 2000 ou 5000.

Temps de calcul

Le SVM et le RF ont des résultats similaires avec les données BS-PS, mais un autre point important de comparaison est le temps de calcul. Or le SVM est connu pour avoir un temps d'apprentissage plus long que celui du RF [Pal, 2005]. Afin de comparer les temps

Tableau 4.5 – Précisions, rappels, et F-Scores par classe et *Overall Accuracy* (OA) moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le SVM pour différents nombres d'échantillons d'apprentissage. Les données en entrée sont BS-PS.

	SVM BS-PS 750			**SVM BS-PS 2000**			**SVM BS-PS 5000**		
C, γ	$12, 5,\ 1, 5e^{-4}$			$12, 5,\ 1, 5e^{-4}$			$2, 5,\ 1, 5e^{-4}$		
	UA	PA	F	UA	PA	F	UA	PA	F
Imperméables	92,8	87,2	89,8	92,6	91,2	91,8	92,0	93,5	**92,6**
Perméables	46,0	71,7	56,0	58,4	61,8	60,0	69,7	54,7	**61,2**
Sol nu	28,9	46,8	34,4	36,6	48,8	40,5	47,4	42,8	**43,6**
Eau	99,8	82,3	90,0	99,8	84,5	91,3	99,8	86,3	**92,4**
Feuillus	93,2	74,1	82,4	93,3	77,0	**84,3**	90,7	72,2	80,2
Conifères	51,4	59,5	**54,8**	53,9	55,9	54,5	53,5	56,8	54,4
Mixtes	18,6	35,1	24,2	19,8	36,6	**25,6**	13,1	26,8	17,5
Arbustes	56,6	86,3	67,9	60,6	87,8	71,3	65,7	90,2	**75,8**
Blé	90,6	89,8	90,2	90,1	91,4	90,7	90,3	92,5	**91,4**
Colza	90,5	94,7	92,5	91,7	94,7	93,1	92,9	94,5	**93,7**
Maïs	93,7	90,4	92,0	93,3	91,6	92,4	93,2	92,0	**92,6**
Orge	50,2	54,7	**52,2**	53,6	48,2	50,6	61,2	43,9	51,1
Soja	79,0	72,5	**75,1**	81,1	70,7	74,8	85,8	67,8	74,9
Sorgho	5,8	3,8	4,4	15,1	3,6	**5,6**	24,7	3,1	5,3
Tournesol	78,9	88,5	83,3	81,5	88,5	84,7	82,2	89,0	**85,3**
Prairies	72,4	79,5	75,7	75,0	81,5	78,0	75,9	83,8	**79,6**
Vergers	93,8	92,3	93,1	94,8	93,5	94,1	95,5	94,1	**94,8**
Vignes	97,7	95,0	96,3	98,0	96,0	97,0	98,2	96,6	**97,4**
OA	77,7 ± 3,8			**79,4 ± 2,7**			77,1 ± 5,17		

UA : *user's accuracy* (précision) ; PA : *producer's accuracy* (rappel)
BS : bandes spectrales ; PS : primitives spectrales
Séparateurs à Vastes Marges (SVM) : C, le paramètre de régularisation ; γ, la largeur du noyau gaussien

Tableau 4.6 – Précisions, rappels, et F-Scores par classe et *Overall Accuracy* (OA) moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le RF pour différents nombres d'échantillons d'apprentissage. Les données en entrée sont BS-PS.

	RF BS-PS 2000			**RF BS-PS 5000**			**RF BS-PS 10000**		
K, m, *max_depth*, *min_samples*	400, 2, 50, 10			400, 17 50, 10			400, 17, 50, 25		
	UA	PA	F	UA	PA	F	UA	PA	F
Imperméables	87,7	90,8	89,1	86,6	93,7	89,9	86,0	94,4	**89,9**
Perméables	55,3	46,6	**50,5**	73,1	34,1	46,4	81,5	23,6	36,3
Sol nu	40,2	63,2	48,3	55,2	56,9	**55,3**	64,0	49,0	54,6
Eau	99,1	99,5	**99,3**	98,8	99,6	99,2	99,0	99,6	99,3
Feuillus	93,7	82,7	87,7	93,9	83,2	88,1	93,8	84,5	**88,8**
Conifères	63,1	71,2	**66,7**	62,5	67,2	64,6	62,0	66,6	64,1
Mixtes	25,2	35,0	**28,6**	24,3	35,8	28,1	24,9	34,2	28,0
Arbustes	71,9	86,3	78,3	75,6	90,0	82,0	75,7	91,2	**82,6**
Blé	92,3	91,8	92,1	91,6	92,9	**92,2**	90,5	93,4	91,9
Colza	90,2	94,4	92,2	91,1	95,3	93,2	93,0	94,0	**93,5**
Maïs	95,1	91,9	93,5	94,7	92,8	**93,8**	94,3	93,1	93,7
Orge	58,5	58,6	58,5	71,0	51,6	**59,7**	77,1	43,8	55,6
Soja	82,0	73,9	77,3	86,6	75,1	**80,1**	88,9	69,7	77,7
Sorgho	15,0	1,4	2,6	35,2	1,4	**2,6**	12,5	0,4	0,8
Tournesol	80,6	89,2	84,7	83,7	90,8	**87,1**	83,5	90,7	86,9
Prairies	68,3	83,2	74,9	69,2	84,6	**76,0**	69,2	84,6	**76,0**
Vergers	91,8	87,7	89,7	94,1	89,6	91,8	94,4	89,7	**92,0**
Vignes	96,1	94,7	95,4	96,8	95,5	96,2	96,9	95,8	**96,3**
OA	82,7 ± 3,8			83,3 ± 3,9			**83,8 ± 3,5**		

UA : *user's accuracy* (précision) ; PA : *producer's accuracy* (rappel)
BS : bandes spectrales ; PS : primitives spectrales
Random Forest (RF) : K, le nombre d'arbres ; m, le nombre de primitives sélectionnées aléatoirement à chaque nœud ; *max_depth*, la profondeur maximale des arbres ; *min_samples*, le nombre minimal d'échantillons par nœud pour continuer les divisions

de calcul de cette étude, le jeux de variables BS-PS est utilisé avec 5000 échantillons d'apprentissage par classe. Dans ce contexte, le Tableau 4.7 montre les temps d'apprentissage pour les deux classifieurs [48]. Un premier cas (**A**) donne les temps d'apprentissage quand le paramétrage optimal est connu (les configurations utilisées sont celles du Tableau 4.4). Un second cas (**B**) donne le temps nécessaire pour l'apprentissage lorsque l'étape d'optimisation des hyper-paramètres est réalisée. Pour le SVM, la procédure décrite dans la Section 4.2.3 est appliquée. Pour le RF, seul le paramètre m est optimisé. Les autres paramètres sont fixés : $K = 100$, $max\_depth = 50$ et $min\_samples = 25$. Les réslutats obtenus dans la Section 4.3.2 montreront que $K = 100$ est un réglage classique pour le RF et que l'influence des paramètres max_depth et $min\_samples$ est négligeable.

Tableau 4.7 – Temps de calcul pour l'apprentissage (avec les écarts-types) en secondes pour le SVM et le RF. La ligne (A) montre les temps d'apprentissage pour un jeu de paramètre, tandis que la ligne (B) montre les temps d'apprentissage avec le temps d'optimisation des paramètres (réalisée sur une grille de recherche).

	SVM BS-PS	**RF BS-PS**
A	265 ± 9 s.	3246 ± 375 s.
B	209856 ± 51497 s.	4207 ± 80 s.
	(≃ 2 jours et 9 h.)	(≃ 1 h.)

BS : bandes spectrales ; PS : primitives spectrales
SVM : *Support Vector Machine* ; RF : *Random Forest*

Connaissant le paramétrage optimal, le SVM est douze fois plus rapide que l'algorithme du RF. Cependant, paramétrer le SVM est rarement évident et nécessite souvent la phase d'optimisation réalisée par validation croisée, qui fait exploser les temps de calcul de quelques minutes à plusieurs jours. Au contraire, le RF est moins sensible à ces paramètres, ce qui résulte d'un temps d'apprentissage moins élevé lorsque l'optimisation des paramètres est faite.

Par ailleurs, aucune des implémentations des deux algorithmes ne tirent bénéfice de la parallélisation. Pour l'algorithme du SVM, il est possible d'apprendre les modèles « un-contre-un » en parallèle [49]. Pour l'algorithme du RF, il est possible d'apprendre en parallèle les différents arbres du modèle puisque la construction de chaque arbre est indépendante. Pour la construction de 100 arbres, le gain de la parallélisation est équivalent pour le RF et le SVM pour un problème à quinze classes.

Pal [2005] et Meyer et al. [2016] tirent les mêmes conclusions au sujet des temps d'apprentissage dans le contexte de la classification de zones agricoles et d'analyse pluviométrique respectivement. Comme décrit par le Tableau 4.7, ces études montrent que le temps et les expériences nécessaires pour configurer le RF sont moins importants que pour le SVM.

L'ensemble des analyses réalisées montre que le RF obtient des résultats meilleurs pour des temps de calcul plus faibles.

4.3.2 Sensibilité aux paramètres du *Random Forest*

L'objectif de cette partie est d'évaluer l'influence des paramètres du RF sur les performances de la classification en utilisant les données satellitaires de la première zone d'étude

48. Les temps de calcul sont obtenus pour un processeur Intel Xeon CPU ES – 2620 / 2.1 GHz, 32 Go. Le facteur de parallélisation de l'ordinateur utilisé est possible grâce aux 24 CPU disponibles.

49. Pour c classes, une approche « un-contre-un » nécessite l'apprentissage de $\frac{c(c-1)}{2}$ modèles.

Tableau 4.8 – *Overall Accuracy* (OA) moyennés sur cinq tirages aléatoires obtenus pour le RF en utilisant différentes valeurs de paramètres. Les valeurs de m sont telles que $m = 2$, $m = \sqrt{p}$, $m = p/3$, $m = p/2$, et $m = p$ avec p la dimension du vecteur de variables.

m \ K	**50**	**100**	**150**	**200**	**400**
BS $p = 116$					
2	80,9	81,3	81,5	81,5	81,6
10	81,3	81,7	81,7	81,9	**82,0**
38	80,9	81,2	81,4	81,5	81,5
58	80,6	80,9	81,1	81,1	81,3
116	79,2	79,5	79,6	79,6	79,8
BS-PS $p = 302$					
2	82,2	82,6	82,8	82,7	82,9
17	82,7	82,3	83,0	83,1	**83,2**
100	82,0	82,9	82,6	82,5	82,7
151	81,8	82,2	82,2	82,3	82,4
302	80,4	80,8	80,9	80,9	81,0
BS-NDVI $p = 139$					
2	81,0	81,4	81,5	81,6	81,7
11	81,4	81,8	81,9	81,9	**82,0**
46	81,1	81,5	81,5	81,6	81,6
69	80,8	81,3	81,2	81,3	81,5
139	79,5	79,8	79,8	80,0	80,0
BS-PT $p = 141$					
2	80,6	81,1	81,3	81,4	81,4
11	81,3	81,6	81,7	81,8	**81,8**
47	80,8	81,4	81,4	81,5	81,6
70	80,6	81,2	81,1	81,3	81,3
141	79,6	79,9	80,0	80,1	80,1
BS-NDVI-PT $p = 164$					
2	80,9	81,2	81,3	81,3	81,4
12	81,3	81,7	81,8	81,9	**82,0**
54	81,1	81,3	81,5	81,5	81,6
82	80,8	81,1	81,3	81,3	81,3
164	79,4	79,7	79,7	79,8	79,9

BS : bandes spectrales ; PS : primitives spectrales ; NDVI : *Normalised Difference Vegetation Index* ; PT : primitives temporelles
K, le nombre d'arbres dans la forêt ; m, le nombre de primitives sélectionnées aléatoirement à chaque nœud

présentée dans la Section 4.2.1. Les cinq vecteurs de variables – BS, BS-NDVI, BS-PS, BS-PT et BS-NDVI-PT – décrits à la Section 4.1.3 sont étudiés. Pour rappel, le RF nécessite la configuration de quatre paramètres : 1) K, le nombre d'arbres dans la forêt ; 2) m, le nombre de primitives sélectionnées aléatoirement à chaque nœud ; 3) *max_depth*, la profondeur maximale de chaque arbre et 4) *min_samples* le nombre minimum d'échantillons autorisé par nœud.

Une première étude vise à montrer l'influence des paramètres K et m sur les performances de classification. Pour ce faire, les valeurs des paramètres *max_depth* et *min_samples* sont fixées dans un premier temps à 25. Le Tableau 4.8 affiche les valeurs d'OA moyennées sur cinq tirages aléatoires obtenus pour les cinq vecteurs de variables avec différentes valeurs de K et de m. Les valeurs en gras correspondent aux valeurs d'OA les plus élevées.

Ces résultats permettent de tirer une conclusion générale à propos du choix des données en entrée : pour toutes les valeurs de paramètres, le vecteur de variables BS-PS obtient le

meilleur résultat. Une discussion spécifique au choix des données en entrée est faite dans la Section suivante.

Concernant le paramètre K, l'OA augmente doucement lorsque le nombre d'arbres augmente. Comme les temps de calcul augmentent linéairement avec le nombre d'arbres, K peut être fixé à 100 sans perte majeure de précision. Ces résultats sur la valeur de K sont en accord avec d'autres études [Rodríguez-Galiano et al., 2012].

Concernant le paramètre m, deux stratégies existent dans la littérature. La première consiste à utiliser une petite valeur de m afin de réduire la corrélation entre les différents arbres, et donc obtenir une meilleure généralisation [Rodríguez-Galiano et al., 2012]. La seconde est d'utiliser la » valeur par défaut « , $m = \sqrt{p}$ [Liaw and Wiener, 2002]. Dans cette étude, le Tableau 4.8 montre que les meilleures performances sont obtenues pour la valeur par défaut (deuxième colonne). Cependant, la différence d'OA est faible pour les valeurs de $m = 2$, $m = \sqrt{p}$ et $m = p/3$ (les trois premières colonnes) à valeur de K fixée. Ces résultats sont aussi cohérents avec ceux de Cutler et al. [2007] et Boulesteix et al. [2012] qui affirment qu'il est préférable de paramétrer m pour chaque nouvelle donnée, et ce, même si le gain de performances est faible.

Par ailleurs, Tatsumi et al. [2015] reportent que lorsque le nombre d'échantillons d'apprentissage augmente, la valeur de m est moins sensible. Pour eux, le gain de précision sera imputé principalement à l'augmentation du nombre de » bons « échantillons d'apprentissage. Une autre explication est aussi possible : l'augmentation du nombre d'échantillons d'apprentissage conduit à augmenter la diversité parmi les arbres car les différences entre échantillons *bootstrap* sont augmentées. Pour cette étude, le faible gain en paramétrant m peut donc être dû au grand nombre d'échantillons d'apprentissage utilisés (5000 par classe).

De plus, le Tabkeau 4.8 montre que les différences d'OA entre $m = \sqrt{p}$ (seconde colonne) et $m = p$ (dernière colonne) peuvent atteindre plus de 2 % pour tous les vecteurs de variables. Contrairement à Oliveira et al. [2012], une trop grande valeur de m entraîne un déclin de l'OA. Le cas $m = p$ implique que toutes les variables soient utilisées à la construction de chaque nœud pour déterminer la meilleure division, ce qui entraine du sur-apprentissage.

Les hyper-paramètres *max_depth* et *min_samples* sont généralement moins étudiés dans la littérature. Leur influence sur les performances de la classification est ici analysée. Le Tableau 4.9 affiche l'OA moyenné sur cinq tirages aléatoires pour les vecteurs de variables BS (le cas le plus simple) et BS-PS (le cas donnant les meilleurs résultats) avec différentes valeurs de paramètres : $min\_samples = \{1, 10, 25, 50, 70\}$ et $max\_depth = \{10, 25, 50\}$. Les valeurs des paramètres K et m sont fixées à l'aide des résultats précédents : $K = 100$ et $m = \sqrt{p}$.

Le Tableau 4.9 montre que l'arrêt précoce de la construction des arbres ($max\_depth = 10$) donne une règle de décision imprécise, ce qui conduit à des faibles précisions. Par ailleurs, la différence d'OA entre $max\_depth = 25$ et $max\_depth = 50$ sont faibles. Ainsi, la valeur de 25 est choisie pour diminuer légèrement les temps de calcul.

Concernant la valeur de *min_samples*, les meilleurs résultats sont obtenus pour $min\_samples = 10$ ou $min\_samples = 25$. Cependant, une diminution de la valeur de *min_samples* a une très faible influence sur les performances de la classification.

Les Tableaux 4.8 et 4.9 montrent aussi que les OA varient de manière cohérente entre les différents jeux de primitives. Les variations d'OA entre la meilleure configuration et les autres configurations sont plus élevées pour les bandes spectrales seules (BS) que pour BS-NDVI et BS-PS. Ainsi, augmenter le nombre de primitives ou ajouter des primitives complexes rend le RF moins sensible à la configuration de ces paramètres.

Tableau 4.9 – *Overall Accuracy* (OA) moyennés sur cinq tirages aléatoires obtenus pour le RF en utilisant différentes valeurs de paramètres.

max_depth \ min_samples	**1**	**10**	**25**	**50**	**70**
BS $p = 116$					
10	78,1	78,3	78,1	78,1	78,0
25	81,3	**81,8**	81,7	81,3	80,8
50	81,2	**81,8**	81,6	81,2	80,9
BS-PS $p = 302$					
10	80,0	79,8	80,0	80,1	79,9
25	82,1	82,3	82,3	82,2	81,8
50	82,1	82,4	**82,5**	82,2	81,9

BS : bandes spectrales ; PS : primitives spectrales
max_depth, la profondeur maximale de chaque arbre ; *min_samples* le nombre minimum d'échantillons autorisé par nœud

4.3.3 Comparaison des différents vecteurs de variables

Le but ici est d'évaluer l'impact des différents vecteurs de variables, présentés dans la Section 4.1.3, sur les performances du RF. Cette étude est encore réalisée sur la première zone d'étude (Section 4.2.1). Afin de comparer seulement les meilleures performances, l'ensemble des hyper-paramètres du RF sont optimisés pour chaque vecteur de variables en suivant la procédure de la Section 4.2.3 . Le Tableau 4.10 affiche les précisions, rappels et F-Scores par classes et l'OA moyennés sur cinq tirages aléatoires. Les valeurs en gras montrent les meilleures valeurs de F-Score et d'OA.

En comparant seulement les OA, le vecteur de variables BS-PS obtient les meilleurs scores. Cependant, l'OA résultant est seulement supérieur de 1 %. De plus, les intervalles de confiance obtenus par les quatre autres vecteurs de variables chevauchent celui des BS-PS. L'analyse par classe des F-Scores donne une conclusion similaire : ajouter de l'information spectrale améliore le F-Score de quasiment toutes les classes, mais les différences de F-Score sont faibles. Ainsi, les résultats du cas le plus simple BS sont très proches de celui de BS-PS. L'absence de gain significatif en ajoutant des primitives spectrales est probablement due à la redondance d'information dans les primitives spectrales avec les bandes spectrales.

Par ailleurs, les résultats obtenus par les jeux de primitives BS-NDVI, BS-PT et BS-NDVI-PT sont très similaires pour toutes les classes. Les primitives temporelles proposées, calculées à partir du profil temporel de NDVI, échouent à apporter de l'information pertinente. En effet, cette information temporelle est déjà contenue dans le profil temporel de NDVI. Comme les primitives temporelles augmentent les temps de calcul (temps d'apprentissage et pour les calculer), sans améliorer les performances de la classification, elles ne seront pas étudiées plus en détail dans la suite.

L'analyse des F-Score montre aussi que les classes surfaces perméables, sols nus, forêts mixtes, orge et sorgho obtiennent des F-Scores en-dessous de 60 %. Comme indiqué par le Tableau 4.3, le nombre d'échantillons d'apprentissage pour ces classes, excepté la classe forêts mixtes, sont sous la barre des 5000 (classes minoritaires). Comme les échantillons *bootsrap* sont sélectionnés aléatoirement parmi l'ensemble des échantillons d'apprentissage pour la construction de chaque arbre, les classes avec le moins d'échantillons d'apprentissage sont désavantagées. Les forêts mixtes, qui correspondent à un mélange de conifères et de feuillus, obtiennent des F-Scores sous les 30 % pour tous les vecteurs de variables. À 20 mètres de résolution spatiale, les forêts mixtes représentent un mélange principalement au niveau macroscopique, *i.e.* une majorité des pixels feuillus ou conifères sont mixtes.

Tableau 4.10 – Précisions, rappels, F-Scores par classe et *Overall Accuracy* (OA) moyennés avec l'intervalle de confiance à 95 % obtenus pour cinq tirages aléatoires en utilisant le RF pour différents vecteurs de variables. Les paramètres du RF sont optimisés pour chaque vecteur de variables.

	BS			**BS-PS**			**BS-NDVI**			**BS-PT**			**BS-NDVI-PT**		
K, m, *max_depth*, *min_samples*	400, 10, 50, 10			400, 17, 50, 10			400, 11, 50, 10			400, 11, 25, 10			400, 12, 50, 10		
	UA	PA	F	UA	PA	F	UA	PA	F	UA	PA	F	UA	PA	F
Imperméables	87,4	93,7	90,4	86,6	93,7	89,9	87,0	93,5	90,1	88,0	93,6	**90,7**	87,5	93,4	90,3
Perméables	76,7	34,8	**47,8**	73,1	34,1	46,4	73,6	32,9	45,3	76,4	33,8	46,8	72,7	32,3	44,6
Sol nu	49,5	52,6	49,9	55,2	56,9	**55,3**	50,4	55,3	51,6	51,3	55,7	52,5	51,5	57,4	53,3
Eau	98,5	97,1	97,7	98,8	99,6	99,2	99,2	99,4	99,3	99,3	99,3	99,3	99,4	99,4	**99,4**
Feuillus	93,3	82,1	87,2	93,9	83,2	**88,1**	93,6	81,9	87,2	93,3	81,9	87,1	93,5	81,7	87,1
Conifères	59,8	67,4	63,2	62,5	67,2	**64,6**	60,3	67,7	63,7	60,5	68,0	63,9	60,5	68,3	64,0
Mixtes	21,2	31,6	24,9	24,3	35,8	**28,1**	22,4	33,5	26,1	21,6	32,0	25,1	22,1	33,0	25,7
Arbustes	73,1	88,4	79,9	75,6	90,0	**82,0**	73,6	88,5	80,2	73,4	88,5	80,1	73,8	88,3	80,2
Blé	91,0	92,8	91,9	91,6	92,9	**92,2**	91,1	92,9	92,0	90,9	92,5	91,7	91,1	92,6	91,9
Colza	90,5	94,6	92,5	91,1	95,3	**93,1**	90,9	94,4	92,6	90,0	94,0	92,0	90,7	94,2	92,4
Maïs	94,9	91,9	93,4	94,7	92,8	**93,8**	94,7	92,1	93,4	94,6	92,1	93,3	94,5	92,1	93,3
Orge	69,7	47,9	56,7	71,0	51,6	**59,7**	68,5	50,3	57,8	69,3	48,5	56,9	66,8	49,5	56,7
Soja	82,2	69,9	75,1	86,6	75,1	**80,1**	82,7	71,0	76,0	83,2	70,3	75,7	82,6	70,6	75,6
Sorgho	15,8	1,5	2,7	35,2	1,4	2,6	16,2	1,6	**3,0**	15,0	1,4	2,6	15,6	1,6	**3,0**
Tournesol	79,9	89,5	84,4	83,7	90,8	**87,1**	80,3	89,9	84,8	79,3	88,7	83,7	80,1	89,4	84,4
Prairies	68,8	83,7	75,4	69,2	84,6	**76,0**	68,5	83,4	75,1	67,8	83,7	74,8	67,8	83,4	74,7
Vergers	92,3	88,0	90,1	94,1	89,6	**91,8**	92,2	88,2	90,1	91,6	87,6	89,6	91,7	88,1	89,9
Vignes	96,3	94,8	95,5	96,8	95,5	**96,2**	96,4	94,7	95,6	96,2	94,5	95,4	96,4	94,6	95,5
OA	82,1 ± 3,6			**83,3 ± 3,9**			82,2 ± 4,0			82,0 ± 4,0			82,1 ± 4,2		

UA : *user's accuracy* (précision) ; PA : *producer's accuracy* (rappel)
BS : bandes spectrales ; PS : primitives spectrales ; NDVI : *Normalised Difference Vegetation Inde* ; PT : primitives temporelles
Random Forest (RF) : K, le nombre d'arbres ; m, le nombre de primitives sélectionnées aléatoirement à chaque noeud ; *max_depth*, la profondeur maximale des arbres ; *min_samples*, le nombre minimal d'échantillons par nœud pour continuer les divisions

Cette catégorie est donc plus difficile à classifier sans prise en compte du contexte spatial.

Les résultats suggèrent que le vecteur de variables BS-PS soit celui donnant le plus d'information, mais la précision globale de la classification est légèrement améliorée par rapport à l'utilisation des autres vecteurs de variables. Cependant, la contribution des différentes primitives peut changer lors de la classification sur de grandes étendues où les paysages présentent de fortes variabilités. Dans ce contexte particulier, la variabilité des données pourrait être mieux prise en compte en ajoutant de l'information spectrale. Cette problématique est l'objet de la section suivante.

4.3.4 Stabilité du *Random Forest* sur de grandes étendues

L'objectif de cette Section est d'évaluer la précision du RF et la contribution des primitives spectrales lors de la classification sur de grandes étendues. Pour ce faire, le second jeu de données composé exclusivement des données Landsat-8 est utilisé (Section 4.2.1). En outre, la zone d'apprentissage circulaire en rouge et les dix neuf zones de validation en bleu de la Figure 4.5 sont considérées. L'évaluation est réalisée en calculant une matrice de confusion pour chaque zone de validation. L'objectif est d'évaluer l'influence de la distance à la zone d'apprentissage sur les performances de la classification.

Afin de confirmer les résultats précédents, les deux jeux de variables ayant obtenus les meilleurs résultats ainsi que le cas le plus simple sont utilisés en entrée du système de classification. Cette étude vise donc à déterminer si les vecteurs de variables BS-PS et BS-NDVI

sont plus robustes et plus stables que les bandes spectrales seules lors d'une classification sur une grande étendue. L'hypothèse est que l'introduction d'indices spectraux normalisés comme le NDVI, moins sensible aux conditions climatiques et topographiques, doivent aider l'algorithme de classification sur de grandes étendues. Les paramètres du RF sont fixés de la manière suivante : $K = 100$, $m = \sqrt{p}$, $max\_depth = 25$ et $min\_samples = 25$.

La Figure 4.6 représente la distribution des échantillons test en fonction de leur classe d'occupation des sols dans toutes les zones de validation. L'axe des abscisses représente la distance (en kilomètres) à la zone d'apprentissage. Les distances positives et négatives représentent des zones de validation localisées au nord-est et au sud-ouest respectivement de la zone d'apprentissage. La droite verticale en pointillé rouge représente la position de la zone d'apprentissage.

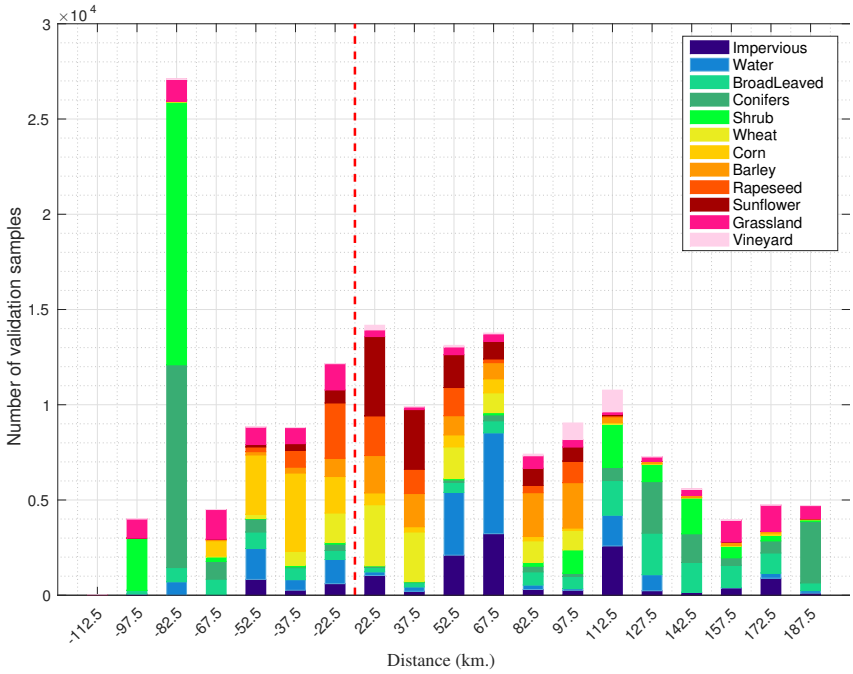


FIGURE 4.6 – Distribution des échantillons test pour chaque classe d'occupation des sols en fonction de la distance à la zone d'apprentissage. La droite verticale en pointillés rouges indique la position de la zone d'apprentissage.

La Figure 4.6 donne donc une idée des variations du paysage sur le gradient nord-sud de la zone d'étude. De -67,5 à 97,5 km. les classes de cultures sont principalement représentées, tandis que de 112,5 à 187,5 km. les forêts détiennent la majorité. Dans le sud-est (-112,5 à -82,5 km.) et le nord-est (à 187,5 km.), la répartition des échantillons test change fortement, ceci est dû à la présence des Pyrénées et du Massif Central. Le nombre d'échantillons test très faible à -112,5 km. est dû à une forte présence de neige et la difficulté pour collecter des informations dans cette zone montagneuse.

La Figure 4.7 montre les valeurs d'OA calculées pour chacune des zones de validation pour les trois vecteurs de variables BS, BS-PS et BS-NDVI.

La Figure 4.7 montre que les comportements des trois vecteurs de variables sont assez similaires. Cependant, ajouter des primitives spectrales améliore légèrement les résultats pour les zones les plus proches de la zone d'apprentissage. Pour les zones éloignées de la zone d'apprentissage, les vecteurs de variables BS et BD-NDVI obtiennent globalement de meilleurs résultats que le jeu de variables BS-PS. Cependant, les valeurs d'OA sont souvent faibles, en-dessous de 50 %, pour ces zones.

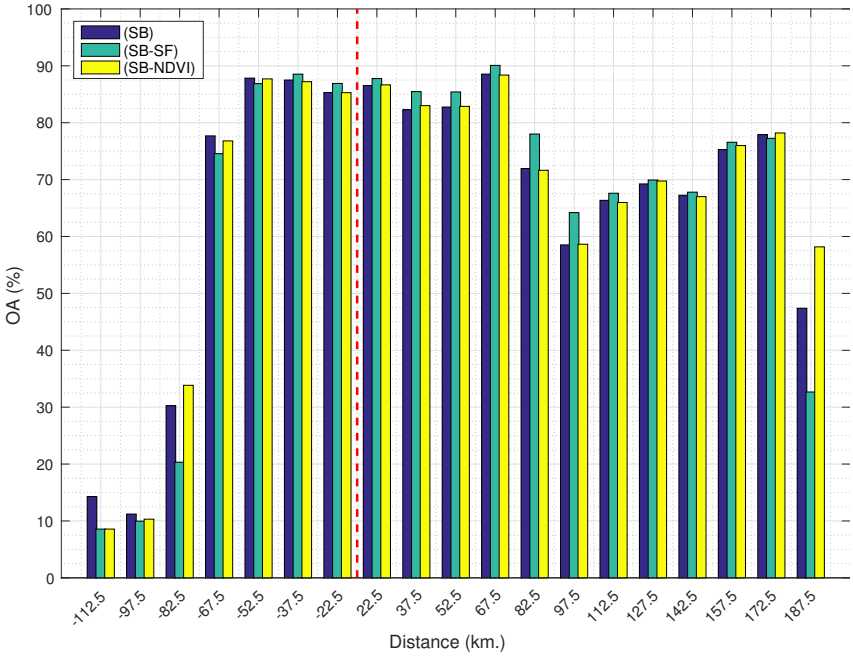


FIGURE 4.7 – OA obtenues pour les trois jeux de variables en fonction de la distance à la zone d'apprentissage. La droite verticale en pointillés rouges indique la position de la zone d'apprentissage. (Légende en anglais. SB : *spectral bands*, SF : *spectral features*)

Plus précisément, les valeurs d'OA restent constantes de -52,5 à 67,5 km. où les paysages sont principalement composés de cultures et de surfaces urbaines, comme ceux de la zone d'apprentissage. Ainsi, le RF peut obtenir de bonnes performances tant que les paysages restent similaires à ceux de la zone d'apprentissage. Alors que les performances chutent, OA en-dessous de 40 %, pour des zones de validation éloignées de plus de -82.5 km. (présence des montagnes). Dans ces zones, les paysages sont significativement différents de ceux de la zone d'apprentissage. Entre 67.5 et 97.5 km, les valeurs d'OA chutent d'environ 25 % : les paysages sont modifiés avec la présence d'arbustes et de vignes (dont les pratiques de cultures changent de ceux de la zone d'apprentissage). Puis, l'OA augmente légèrement de 112,5 à 172,5 km. pour les zones de validation où les cultures sont peu présentes.

Pour conclure, les résultats obtenus montrent que l'addition d'information spectrale supplémentaire ne résulte pas nécessairement en une amélioration de la précision, alors qu'elle accroît les temps de calcul. Par ailleurs, ces résultats peuvent servir pour mettre en place une stratégie d'échantillonnage. En effet, les performances de classification restent inchangées lorsque les paysages sont similaires. Ainsi, la stratification des données d'entrée, *i.e.* la séparation de la zone d'étude en plus petites zones en fonction des types de paysages, peut aider à améliorer la qualité de la carte produite. Par exemple, une stratification par zone éco-climatique est utilisée pour la production de la carte du CES OSO [Inglada et al., 2017].

4.4 Conclusion

Plusieurs enjeux de la cartographie de l'occupation des sols sur de grandes étendues ont été discutés dans ce chapitre. Le potentiel du RF pour la classification de séries temporelles d'images satellitaires à hautes résolutions a été montré.

L'algorithme du RF a obtenu des résultats comparables aux SVM avec un meilleur

compromis entre la précision et les temps de calcul, notamment grâce à un paramétrage plus évident. Le RF a aussi montré être moins influencé par l'utilisation de différents jeux de variables. L'algorithme du RF est donc un outil approprié pour gérer la quantité de données fournie par les nouvelles données satellitaires. Les résultats du SVM et du RF ont aussi montré des complémentarités, notamment pour les classes d'occupation des sols avec de faibles précisions. La fusion des résultats des deux classifieurs peut donc mener à des résultats plus précis que les résultats d'un seul classifieur (principe des méthodes d'ensemble).

L'étude de primitives spectrales et temporelles a aussi été proposée, afin notamment de caractériser les classes dynamiques comme celles de végétation. Bien qu'elles améliorent les performances de la classification, le gain en précision est très faible comparé au surcoût induit par le calcul de ces primitives et l'augmentation de la dimension du problème de classification. Plus précisément, il était attendu que les primitives temporelles ajoutent de l'information pertinente qui aide l'algorithme de classification à discriminer les différentes occupations des sols. Cependant, le RF gère déjà l'information temporelle en exploitant les signatures spectrales des séries temporelles. Ce qui permet de diminuer les confusions entre les classes d'occupation des sols qui varient au cours du temps. De plus, le paramétrage du RF a peu d'influence sur les performances de la classification. Pour conclure, l'utilisation seule des bandes spectrales est donc un bon compromis entre la précision et les temps de calcul.

La procédure de classification a aussi été testée dans le contexte particulier où les échantillons d'apprentissage sont spatialement localisés. Plus précisément, les performances de classification ont été évaluées avec différents jeux de variables. L'étude révèle que les primitives spectrales n'aident pas l'algorithme de classification même lorsque la zone d'apprentissage est loin. De manière générale, les performances de classification diminuent lorsque les paysages sont différents de ceux de la zone d'apprentissage. Par conséquent, il peut être intéressant de stratifier les échantillons d'apprentissage en fonction des paysages étudiés. Des travaux précédents avaient notamment été menés pour la sélection des échantillons par région éco-climatiques, topographiques ou encore pédologiques [Arnau Rodes, 2016].

Au moment de l'étude, les données Sentinel-2 n'étaient pas encore disponibles. Cependant, la résolution spatiale de 10 mètres, ainsi que la présence de plusieurs canaux dans le rouge et le proche infra-rouge pourrait permettre d'améliorer les résultats notamment lors de la classification sur de grandes étendues.

Chapitre 5

Influence des données mal étiquetées sur la qualité des cartes d'occupation des sols

Sommaire

5.1 Problématique des données mal étiquetées 94
5.1.1 Données mal étiquetées en télédétection 94
5.1.2 Apprentissage supervisé en présence de données mal étiquetées 96
5.2 Présentation des expérimentations 97
5.2.1 Données utilisées 98
5.2.2 Génération du bruit d'étiquetage 102
5.2.3 Stratégie d'échantillonnage 104
5.2.4 Configuration des algorithmes de classification 105
5.3 Résultats des expérimentations 105
5.3.1 Influence du nombre de classes 105
5.3.2 Influence du vecteur de variables 107
5.3.3 Influence du nombre d'échantillons 107
5.3.4 Complexité des algorithmes de classification 109
5.3.5 Étude d'un bruit aléatoire systématique 111
5.3.6 Comparaison du *Random Forest* et du *Support Vector Machine* 113
5.4 Conclusion 115

L'utilisation de méthodes de classification supervisée pour l'obtention de cartes d'occupation des sols nécessite des échantillons d'apprentissage pour entraîner l'algorithme de classification. Ces échantillons d'apprentissage sont décrits par un vecteur de variables extrait des données satellitaires et d'une étiquette fournie par la donnée de référence. L'étiquette est une indication sur la classe d'occupation des sols des échantillons d'apprentissage, qui est utilisée dans la prise de décision pour attribuer la classe de nouvelles observations. La qualité des étiquettes est donc directement liée à la précision de la règle de décision apprise par l'algorithme de classification.

Bien que cette étiquette représente généralement la vraie classe de l'échantillon, il arrive fréquemment que des échantillons d'apprentissage soient mal étiquetés. Dans ce cas là, l'étiquette fournie par la donnée de référence ne correspond pas à la classe de la

vérité terrain. Cette présence de données mal étiquetées s'explique en grande partie par la difficulté d'obtenir des données de référence fiables et de bonne qualité. En effet, la collecte des données de référence est en télédétection une tâche fastidieuse et coûteuse.

Ce chapitre s'intéresse spécifiquement aux conséquences de l'utilisation d'échantillons d'apprentissage mal étiquetés sur le processus de classification. Dans un premier temps, la problématique liée à la présence de données mal étiquetées parmi les échantillons d'apprentissage est décrite. Dans un deuxième temps, les études menées afin d'évaluer l'influence sur le processus de classification des échantillons d'apprentissage mal étiquetés sont présentées. Ensuite, les résultats de ces études sont détaillés. Enfin, les conclusions sont tirées.

5.1 Problématique des données mal étiquetées

La présence de données mal étiquetées est assimilée à un bruit sur les étiquettes, parfois appelée erreur d'étiquetage. Cette section s'intéresse aux problématiques liées à la présence de données mal étiquetées dans les échantillons d'apprentissage. Dans un premier temps, des généralités sur le bruit présent dans les échantillons d'apprentissage sont introduites, et spécifiquement les sources de bruit en télédétection sont décrites. Dans un second temps, les conséquences de ce bruit sur le processus de classification, notamment lors de la phase d'apprentissage, sont abordées.

5.1.1 Données mal étiquetées en télédétection

Les données utilisées en entrée de tout système de classification peuvent contenir du bruit qui impacte la précision du vecteur de variables et l'exactitude des étiquettes des échantillons. Ce bruit présent dans les données fournies en entrée du système de classification est traditionnellement divisé en deux catégories [Nettleton et al., 2010; Sáez et al., 2014; Zhu and Wu, 2004]. La première catégorie correspond au bruit sur le vecteur de variables (*feature noise* en anglais) qui est principalement ajouté lors de l'acquisition des données et de leurs pré-traitements. La seconde catégorie représente les données mal étiquetées (*class label noise* en anglais) qui correspondent à un désaccord entre l'étiquette de la donnée de référence et la vérité sur le terrain. Les données mal étiquetées sont généralement dues au processus de collecte des données de référence.

Dans le domaine de la télédétection, le bruit sur le vecteur de variables est dû aux pré-traitements appliqués sur les images satellitaires. Par exemple, les étapes d'orthorectification, de corrections géométriques, radiométriques, et aussi le calcul de primitives peuvent apporter des incertitudes sur le vecteur de variables. Par ailleurs, la présence d'obstacles sur les images (*e.g.* les nuages, les ombres, des objets en mouvement) cache la surface au sol d'intérêt. Les valeurs de réflectance utilisées dans le vecteur de variables sont alors erronées. Une autre source d'erreur est due à la numérisation de données d'archives utilisées pour des études sur le long terme. La présence d'artefacts et de déformations conduit à des imprécisions géométriques et radiométriques.

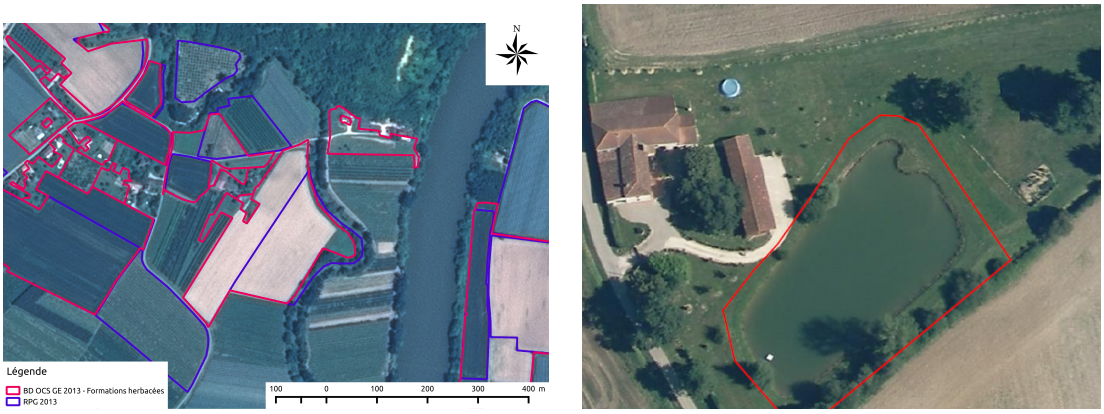
Concernant la seconde source de bruit, la présence des données mal étiquetées est de plus en plus courante dans les systèmes de classification actuels due à la difficulté de collecter de manière fiable les étiquettes des échantillons. Deux types de données mal étiquetées sont identifiés dans la littérature : 1) les échantillons contradictoires, *i.e.* le même échantillon qui est utilisé plusieurs fois avec différentes étiquettes ; et 2) les erreurs d'étiquetage, *i.e.* lorsque l'étiquette associée à un échantillon est différente de la vérité

terrain [Zhu and Wu, 2004]. Le terme données mal étiquetées fait référence dans ce manuscrit seulement aux erreurs d'étiquetage, le cas des échantillons contradictoires n'étant pas abordés.

Dans le contexte de la cartographie de l'occupation des sols, des erreurs d'étiquetage peuvent être ajoutées lors de la collecte de données. Elles sont dues à un manque d'expérience de l'opérateur ou à un manque d'information sur le terrain (mauvaise saison). Par exemple, un opérateur ne pourra pas distinguer au mois de septembre les cultures d'hiver qui sont déjà récoltées. De plus, les classes d'occupation des sols à discriminer peuvent être très complexes et les définitions des occupations des sols peuvent être ambiguës. Par exemple, le blé et l'orge sont deux cultures très similaires, difficiles à discriminer sur le terrain et encore plus à partir d'images drones ou satellitaires. Des erreurs informatiques et humaines peuvent aussi causer la modification de l'étiquette par mégarde.

Par ailleurs, la combinaison de plusieurs données de référence, parfois nécessaire pour couvrir de grandes étendues, peut mener à la présence de données mal étiquetées. En effet, les désaccords entre bases de données et les hétérogénéités spatiales affectent la précision des étiquettes référencées.

La Figure 5.1 montre deux types d'imperfections présents dans les bases de données. La Figure 5.1a montre par exemple les désaccords sur la définition des polygones de culture pour l'année 2013 entre deux bases de données. Les polygones bleus représentent l'information du RPG , tandis que les polygones rouges représentent l'information de la donnée OCS-GE. Certains polygones sont présents seulement pour l'une des deux bases de données, tandis que d'autres polygones sont scindés en plusieurs polygones pour l'autre base de données. Des désaccords sur la géométrie des polygones sont aussi visibles.



(a) Désaccords entre deux bases de données sur des polygones de culture

(b) Imprécision géométrique sur la définition d'un polygone en eau

FIGURE 5.1 – Illustration de différentes imperfections dans les bases de données.

Une autre source d'erreur d'étiquetage est due aux imprécisions géométriques, qui peuvent être présentes si les données satellitaires ne sont pas parfaitement superposables aux données de référence [Foody, 2002]. Dans l'exemple de la Figure 5.1b, le polygone d'eau qui apparaît en rouge est trop large par rapport à la taille réelle de l'étendue d'eau. Il englobe alors des arbres et des surfaces en herbe. Pourtant ces pixels seront étiquetés comme de l'eau alors qu'ils représentent d'autres classes d'occupation des sols. De plus, l'année d'acquisition des données satellitaires doit être cohérente avec celle de la donnée de référence au risque d'introduire des erreurs sémantiques si les occupations des sols ont changé. Par exemple, l'utilisation du RPG 2014 pour la classification d'une série temporelle acquise en 2016 va avoir pour conséquence l'introduction de données

mal étiquetées. En effet, la rotation des cultures conduit à de nombreux changements d'occupations des sols pour les parcelles agricoles.

Le bruit dépend donc de la qualité des images satellitaires et de la donnée de référence. Il influence la précision du vecteur de variables ainsi que l'étiquette des échantillons d'apprentissage. Bien que les deux types de bruit ont un impact sur le système de classification, plusieurs études montrent que la présence de bruit sur le vecteur de variables est moins nuisible que la présence de données mal étiquetées [Nettleton et al., 2010; Zhu and Wu, 2004]. Selon Frénay and Verleysen [2014], deux raisons expliquent ce résultat : 1) chaque échantillon a plusieurs primitives mais une seule étiquette, et 2) les classifieurs les plus robustes utilisent des stratégies pour prendre en compte l'importance des variables, et donc donner moins de poids aux variables les plus bruitées. Ainsi, ces travaux s'intéressent uniquement à la présence de données mal étiquetées.

Par ailleurs, la stratégie d'échantillonnage utilisée conduit à la présence des données mal étiquetées à la fois dans les échantillons d'apprentissage et dans les échantillons test (Chapitre 2). D'un côté, les échantillons test servent à évaluer les performances de l'algorithme de classification. D'un autre côté, les échantillons d'apprentissage, utilisés dans les processus de classification supervisée, guident le classifieur pour déterminer sa règle de décision. Dans la suite, seule la problématique de la présence des données mal étiquetées dans les données d'apprentissage est abordée.

5.1.2 Apprentissage supervisé en présence de données mal étiquetées

L'utilisation d'échantillons d'apprentissage mal étiquetés est connue pour avoir des conséquences sur le processus de classification, en particulier sur les performances des algorithmes de classification supervisée. Afin de réduire la sensibilité des algorithmes de classification aux données mal étiquetées, une solution proposée est de rendre plus robustes les classifieurs.

Par exemple, l'algorithme d'*AdaBoost* a tendance à sur-pondérer les données mal étiquetées au fil des itérations [Dietterich, 2000b]. Ainsi, l'algorithme *ORBoost* (*Outlier Removal Boosting*) annule au cours des itérations le poids des échantillons les plus surpondérés par *AdaBoost* [Karmaker and Kwek, 2006]. Le nombre d'échantillons dont le poids doit être annulé est un paramètre difficile à définir. D'autres variantes proposent alors de d'abord identifier les échantillons mal étiquetés, puis de de pondérer positivement ou négativement leur influence [Cao et al., 2012; Sun et al., 2016].

Un autre exemple d'algorithme robuste est le CN2 qui simplifie la règle de décision de l'arbre de décision C4.5 [Clark and Niblett, 1989]. En effet, les arbres de décision binaire sont connus pour avoir des règles de décision très complexes qui s'adaptent parfaitement aux données d'apprentissage [John, 1995].

Cependant, l'utilisation de ces algorithmes plus robustes est inefficace lorsque le nombre de données mal étiquetées est élevé [Frénay and Verleysen, 2014].

Des études plus générales ont analysé la robustesse de plusieurs classifieurs. Par exemple, Folleco et al. [2009] évaluent les performances de onze classifieurs pour plusieurs problèmes de classification, souvent déséquilibrés. Ils concluent que la présence de données mal étiquetées impacte différemment les classifieurs. Dans ces travaux, les méthodes d'ensemble, notamment le RF, sont plus robustes que les autres classifieurs. De même, la combinaison des prédictions de plusieurs algorithmes de classification améliore les résultats en présence de données mal étiquetées par rapport à l'utilisation d'un seul classifieur [Sáez et al., 2013]. Dans le contexte de la cartographie des sols, Brodley and

Friedl [1999] montrent que les performances de trois classifieurs – arbre de décision binaire C4.5, 1-Plus Proches Voisins (PPV) et analyse discriminante linéaire – décroissent linéairement lorsque le nombre d'échantillons d'apprentissage mal étiquetés augmente.

Outre les performances de classification, la complexité des classifieurs est aussi impactée en présence de données mal étiquetées. Par exemple, le chemin moyen parcouru par les échantillons d'apprentissage dans des arbres de décision peut augmenter, conduisant à une augmentation des temps de calcul. L'influence des données mal étiquetées est aussi mesurée en s'intéressant directement à la complexité des échantillons d'apprentissage [Garcia et al., 2015]. Par exemple, la séparabilité entre les classes diminue lorsque la présence d'échantillons mal étiquetés augmente. L'impact des données mal étiquetées a aussi été évalué sous certaines contraintes par exemple en limitant le nombre d'échantillons d'apprentissage [Mellor et al., 2015]. Finalement, les tâches connexes à l'entraînement de l'algorithme de classification, comme la sélection de primitives, sont aussi affectées [Pechenizkiy et al., 2006].

Bien que certaines études quantifient l'influence des données mal étiquetées sur les performances de la classification, à notre connaissance, aucune n'a été menée dans le contexte de la classification de séries temporelles d'images satellitaires sur de grandes étendues. L'absence de ces études s'explique par la difficulté d'avoir un jeu de données réel soit « propre », soit avec une connaissance parfaite des données mal étiquetées [Carlotto, 2009].

Ces travaux cherchent à caractériser l'influence de la présence de données mal étiquetées dans l'ensemble d'apprentissage sur le système de classification, notamment sur ses performances. Afin d'étudier cette problématique sur les données mal étiquetées, la terminologie suivante est adoptée :

— classe vérité terrain c_{vt} – la classe réelle, celle sur le terrain au moment de l'acquisition des images [50],

— classe donnée de référence c_r – l'étiquette fournie par la donnée de référence qui est notamment utilisée pour l'apprentissage,

— classe prédite c_p – la classe prédite par l'algorithme de classification.

5.2 Présentation des expérimentations

Dans ces travaux, l'objectif est d'évaluer l'influence des échantillons d'apprentissage mal étiquetés sur les performances de la classification en fonction 1) de différents niveaux de bruit d'étiquetage, et 2) du choix de l'algorithme de classification. Plus précisément, l'impact des données mal étiquetées sur les performances de classification du RF, du SVM-RBF et du SVM-linéaire est étudié pour différentes configurations de classification, *e.g.* nombre de classes, vecteurs de variables ou encore nombre d'échantillons.

Ce type d'étude est difficile à conduire à cause en partie du manque de jeux de données répertoriant les erreurs d'étiquetage. Pour surmonter ces limitations, le problème des données mal étiquetées est souvent analysé sur un jeu de données simulées et un jeu de données réelles pour lesquels les données mal étiquetées sont ajoutées de manière artificielle [Garcia et al., 2015; Natarajan et al., 2013; Xiao et al., 2015a]. L'utilisation de données simulées est complémentaire à l'utilisation de données réelles. En effet, les données simulées permettent de contrôler le niveau de bruit, tandis que les données réelles

50. Dans le cas d'étude sur des séries temporelles, il est possible que l'occupation du sol change au cours de la série temporelle, *e.g.* construction d'un bâtiment ou double cultures. Dans le contexte de ces études, seule la classe présente majoritairement dans le temps est considérée.

représentent mieux la complexité du problème de classification. Dans ces travaux, une stratégie similaire est adoptée.

Cette partie présente les expérimentations. Dans un premier temps, les données utilisées pour les différentes études sont décrites dans la Section 5.2.1. En particulier, le principe de génération des données simulées est développé. Dans un deuxième temps, la procédure pour générer les données mal étiquetées est détaillée dans la Section 5.2.2. Puis, la stratégie d'échantillonnage permettant d'obtenir les ensembles d'échantillons d'apprentissage et test est présentée dans la Section 5.2.3. Enfin, la configuration des paramètres des algorithmes de classification étudiés (RF, SVM-RBF et SVM-linéaire) est donnée à la Section 5.2.4.

5.2.1 Données utilisées

Dans ces travaux de thèse, les études sur l'influence des données mal étiquetées sont menées sur deux jeux de données. Le premier est constitué de données simulées pour lesquelles un bruit arbitraire sur les étiquettes et les vecteurs de variables peut être ajouté. Ce bruit est donc connu et contrôlé. Le second correspond à un jeu de données réel où les vecteurs de variables sont extraits de données satellitaires et l'étiquette d'une donnée de référence. Ainsi, la diversité des paysages est mieux représentée, mais il est possible que du bruit réel sur les vecteurs de variables et sur les étiquettes soit présent. Les deux jeux de données décrivent une année complète d'un cycle de végétation, qui facilite la reconnaissance des occupations des sols comme les cultures.

Données simulées

La création d'un jeu de données simulées décrivant une série temporelle d'images satellitaires optiques proches de la réalité n'est pas une tâche simple. Bien qu'il existe des travaux pour prédire les réflectances TOC [Gao et al., 2006], il est presque impossible de créer des modèles réalistes garantissant que les propriétés statistiques des données simulées soient proches de celles des images satellitaires.

En revanche, certaines études montrent que le cycle phénologique de la végétation peut être modélisé en utilisant des indices de végétation comme le NDVI. La Section 4.1.2 a notamment montré que la communauté de télédétection a proposé plusieurs méthodes pour modéliser les profils de NDVI : les fonctions gaussiennes asymétriques [Jönsson and Eklundh, 2004], les fonctions logistiques par morceaux [Zhang et al., 2003] ou encore la double logistique [Beck et al., 2006; Fisher et al., 2006].

Ainsi, les données simulées proposées dans ces travaux sont basées sur la génération de profils de NDVI décrivant des classes de végétation. Le modèle utilisé pour l'extraction de paramètres phénologiques dans le Chapitre 4 est utilisé ici pour simuler des profils de NDVI :

$$NDVI(t) = A\left(\frac{1}{1+e^{\frac{x_0-t}{x_1}}} - \frac{1}{1+e^{\frac{x_2-t}{x_3}}}\right) + B, \tag{5.1}$$

avec A l'amplitude, B le minimum, x_0 et x_2 les points d'inflexion, et x_1 et x_3 les taux d'accroissement et de décroissement de la courbe aux points d'inflexion x_0 et x_2 respectivement. Le paramètre t représente un vecteur de dates qui peut par exemple représenter une année civile pour la simulation de classes de végétation.

L'équation (5.1) modélise le cycle phénologique d'une seule culture. Cependant, la somme de plusieurs double logistique peut permettre de simuler la croissance de plusieurs

cultures et aussi la simulation de classes plus complexes. Dans le cadre de ces travaux, les profils NDVI simulés représentent seulement une culture par année. Les profils représentant le colza seront néanmoins simulés par la somme de deux double logistique afin de simuler la floraison du colza en avril.

De plus, une repousse de végétation est simulée par une gaussienne afin d'obtenir des profils plus réalistes. Elle est ajoutée au cycle principal de végétation décrit par l'équation (5.1). Un bruit blanc sur chaque date est également ajouté à l'équation (5.1) pour obtenir une variabilité réaliste sur les profils. La Figure 5.2 montre un exemple d'un cycle principal de végétation suivi par une repousse de végétation.

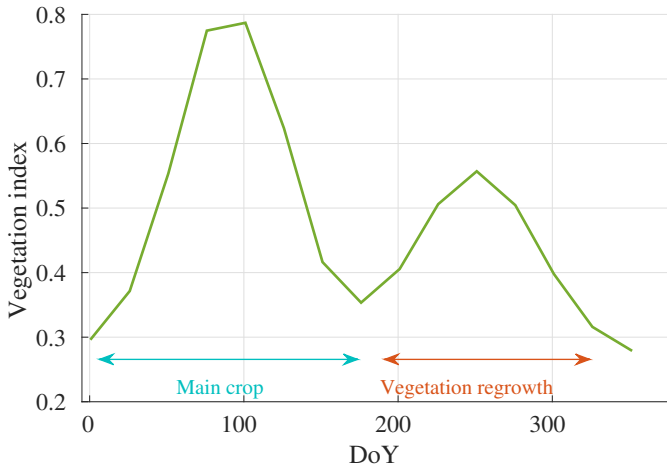


FIGURE 5.2 – Exemple d'un profil de végétation au cours de l'année (DoY : *Day of Year* en anglais).

Le modèle décrit par l'équation (5.1) est utilisé pour simuler les dix classes de végétation qui constituent les jeux de données proposés. Au total, il y a cinq classes de cultures d'été (maïs, maïs ensilage [51], sorgho, tournesol et soja), trois classes de cultures d'hiver (blé, colza et orge) et deux classes de forêt (persistant et caduc). Les profils sont simulés en s'appuyant sur des connaissances expertes afin qu'ils soient les plus réalistes possible.

Pour réduire les différences entre les données simulées et réelles, le concept de polygones a été introduit dans la procédure de simulation. En effet, l'occupation des sols est principalement référencée sous forme de polygones dans les bases de données utilisées. Les polygones permettent de décrire des structures particulières, *e.g.* une parcelle agricole, un bâtiment, un ensemble forestier. Ainsi, les échantillons extraits d'un même polygone sont généralement très similaires.

Suivant cette particularité, la génération de profils n'est pas faite de manière indépendante, mais par polygone. Ainsi, chaque échantillon est défini par un vecteur de variables, une occupation des sols, mais aussi un identifiant polygone. Les échantillons provenant d'un même polygone ont généralement des profils similaires. En conséquence, les échantillons simulés partageant le même identifiant polygone seront plus similaires que les échantillons appartenant à la même classe mais à un autre polygone.

Le processus de simulation repose sur le choix de deux types de paramètres : 1) les paramètres globaux dont les valeurs sont identiques pour toutes les classes, et 2) les paramètres de classe dont les valeurs dépendent de la classe simulée.

Trois paramètres globaux sont ici utilisés : le vecteur de dates t (équation (5.1)), le

51. Contrairement au maïs, le maïs ensilage – utilisé pour nourrir le bétail – est récolté lorsqu'il est encore un peu vert, causant une baisse rapide de la valeur de NDVI pendant la sénescence.

nombre d'échantillons par polygone n, et le nombre de polygones par classe nbp[52]. Dans ce cas là, multiplier le nombre de classes par le produit $n \times nbp$ donne le nombre total d'échantillons simulés. Pour ces études, $n = 10$, $nbp = 100$ et t varie de 1 à 351 par pas de 25 jours, soit 15 dates simulées.

Pour les paramètres de classes, un intervalle allant d'un minimum (min) à un maximum (max) est défini pour chacun des paramètres A, B, et x_i, $0 \leq i < 4$ de l'équation (5.1). Le Tableau 5.1 montre les différentes valeurs de min et max pour les dix classes de végétation simulées. Les paramètre du colza sont sur deux lignes car deux double logistique sont utilisées.

Tableau 5.1 – Valeurs minimales et maximales des paramètres de la double logistique pour dix classes d'occupation des sols. Le colza est simulée par la somme de deux double logistique.

		A		B		x_0		x_1		x_2		x_3	
Cultures d'été	**Maïs**	0,57	0,72	0,15	0,30	100	200	05	25	250	310	10	30
	Maïs ensilage	0,57	0,72	0,15	0,30	100	200	05	25	250	310	05	10
	Sorgho	0,62	0,77	0,15	0,30	120	190	20	40	290	295	25	30
	Tournesol	0,67	0,82	0,15	0,30	102	192	15	40	180	240	05	20
	Soja	0,67	0,82	0,15	0,30	140	220	15	45	270	320	20	45
Cultures d'hiver	**Blé**	0,52	0,67	0,20	0,35	30	90	05	25	125	175	05	25
	Colza	0,70	0,80	0,05	0,20	30	45	15	25	80	90	03	12
		0,60	0,70	0,05	0,15	85	95	03	12	135	145	05	15
	Orge	0,52	0,67	0,20	0,35	30	90	05	25	120	170	05	25
Forêts	**Persistant**	0,01	0,02	0,55	0,70	0	365	100	150	0	365	100	150
	Caduc	0,20	0,35	0,40	0,50	23	27	15	20	315	320	15	20

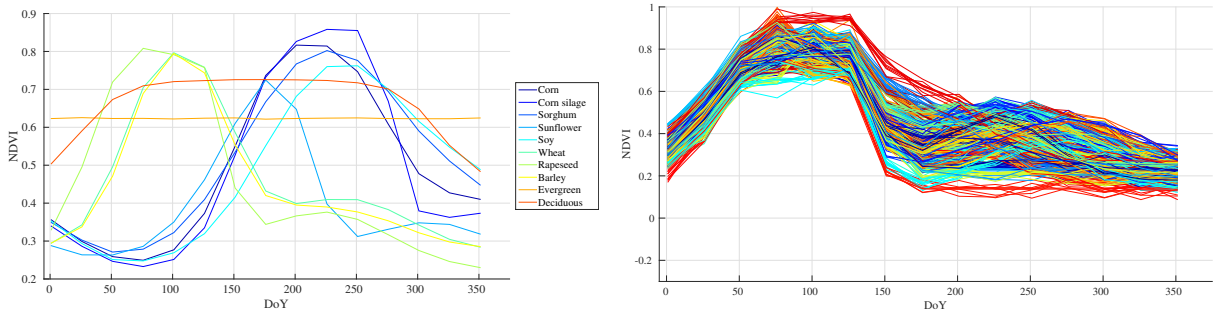
Comme mentionné auparavant, chaque échantillon a un identifiant polygone. Ainsi, chaque polygone a ses propres paramètres. Ces derniers sont choisis aléatoirement en utilisant une distribution normale $\mathcal{N}(\mu, \sigma)$, avec $\mu = \frac{max-min}{2}$ et $\sigma = \frac{max-\mu}{3.0}$ (Tableau 5.1). Afin d'éviter que les n échantillons du polygone soient identiques, les valeurs des six paramètres sont légèrement modifiées pour chaque échantillon en ajoutant un bruit gaussien. La même repousse de végétation, modélisée par une distribution gaussienne, est assignée pour chaque échantillon partageant le même identifiant polygone. Chaque échantillon est finalement contaminé par un bruit blanc indépendant.

La Figure 5.3 montre des exemples de profils NDVI en fonction du jour de l'année (DoY) obtenus par la procédure de simulation. Plus précisément, la Figure 5.3a montre les profils NDVI moyennés pour l'ensemble des échantillons appartenant à la même classe (500 profils par classe).

La Figure 5.3b montre les profils NDVI de colza pour 500 échantillons (10 échantillons répartis dans 50 polygones). Les profils de même couleur représentent des échantillons appartenant au même polygone. Sur ces profils, la repousse de végétation est visible entre les jours de l'année 175 et 325. Elle est plus ou moins importante en fonction des polygones. Par ailleurs, la floraison du colza est visible. Elle se traduit par une légère diminution puis ré-augmentation dans la valeur de NDVI en avril vers le jour de l'année 100.

À l'aide de ce processus de simulation, trois jeux de données sont générés. Le Tableau 5.2 montre que les jeux de données sont composés d'un nombre différent de classes afin de pouvoir étudier l'influence du nombre de classes sur les performances de classification en présence de données mal étiquetées.

52. Les paramètres n et nbp peuvent être des paramètres de classe si l'on souhaite simuler des problèmes déséquilibrés. Pour ces études, seuls les problèmes équilibrés sont utilisés afin de ne mesurer que l'effet des données mal étiquetées sur les performances de la classification.



(a) Profils NDVI moyens pour chaque occupation des sols (légende en anglais).

(b) Profils NDVI simulés pour la classe de colza. Les échantillons appartenant au même polygone ont leur profil NDVI de la même couleur.

FIGURE 5.3 – Exemple de profils simulés de *Normalized Difference Vegetation Index* (NDVI) pour l'ensemble du cycle phénologique (légende en anglais). DoY : *Day of Year*.

TABLEAU 5.2 – Occupation des sols utilisés pour chaque jeu de données simulées.

	2-classes	**5-classes**	**10-classes**
Occupation des sols	Maïs Maïs ensilage	Maïs Maïs ensilage Sorgho Tournesol Soja	Maïs Maïs ensilage Sorgho Tournesol Soja Blé Colza Orge Persistant Caduc

Données réelles

Pour les données réelles, les vecteurs de variables sont extraits des images satellitaires et les étiquettes associées sont fournies par une donnée de référence. Contrairement aux données simulées, les données réelles ne sont pas garanties sans données mal étiquetées, mais elles représentent mieux la complexité du problème de classification. Afin de pouvoir comparer les résultats avec les données simulées, seules des classes de végétation sont étudiées pour les données réelles.

Les vecteurs de variables sont extraits de la série temporelle composée des images SPOT-4 et Landsat-8. Les données du RPG 2013 sont utilisés dans cette étude comme données de référence. Les occupations des sols sélectionnées sont deux cultures d'été (tournesol et maïs), et trois cultures d'hiver (orge, blé et colza).

Le RPG est connu pour pouvoir contenir des erreurs d'étiquetage [53]. Afin de limiter cette présence de données mal étiquetées, deux pré-traitements sont appliqués. Le premier consiste à éroder les polygones de 80 mètres (équivalent à la taille de quatre pixels), et le second consiste à éliminer les polygones de petites tailles. Le Tableau 5.3 montre le nombre total de polygones disponibles une fois les pré-traitements réalisés.

Afin d'assurer une configuration expérimentale similaire à celle mise en place pour les données simulées, *nbp* polygones par classe sont sélectionnées aléatoirement dans la donnée de référence. Puis, n échantillons sont extraits aléatoirement de chaque polygone.

53. Les données terrain utilisées dans la Section 4.2.1 sont probablement moins bruitées que le RPG. Cependant, le nombre de polygones et de pixels par polygone était trop limité pour les études proposées.

Tableau 5.3 – Nombre de polygones disponibles par classe dans le Registre Parcellaire Graphique (RPG).

Classe	Nb. de polygones disponibles
Blé	1197
Maïs	883
Orge	125
Colza	164
Tournesol	851

De manière similaire aux données simulées, les échantillons sont dans un premier temps décrits par leur profil temporel de NDVI calculés pour chaque image de la série SPOT-4 Landsat-8. Cependant, l'information des données satellitaires ne se résume pas au NDVI. Ainsi, les deux vecteurs de variables donnant les meilleurs OA au Chapitre 4 sont aussi utilisés dans ces études : bandes spectrales et NDVI (BS-NDVI) et bandes spectrales et primitives spectrales (BS-PS). Augmenter le nombre de variables permet alors d'étudier l'influence des données mal étiquetées en fonction du vecteur de variables utilisé.

Dans la suite du manuscrit, ces données réelles sont nommées données SPOT-Landsat. Au total, douze configurations montrées par le Tableau 5.4 sont utilisées. En particulier, le Tableau 5.4 affiche le nombre d'échantillons par polygone n, le nombre de polygones par classe nbp et la taille du vecteur de variables. Comme pour les données simulées, le nombre total d'échantillons peut être calculé en multipliant le nombre de classes par $n \times nbp$.

Les jeux de données 1 à 6 sont composés de 1000 échantillons par classe, *i.e.* le même nombre d'échantillons que pour les données simulées. Dans ces premiers jeux de données, le nombre d'échantillons est proche de la taille des vecteurs de variables BS-NDVI et BS-PS. Dans ces cas là, les algorithmes de classification peuvent être affectés par la malédiction de la dimension. Afin d'éviter le phénomène de Hughes (Section 2.1), les jeux de données 7 à 12 sont aussi étudiés. Pour ces données, 40 échantillons sont sélectionnés aléatoirement dans 120 polygones pour chaque classe d'occupation des sols. Ainsi, le nombre d'échantillons par classe est de 4800.

5.2.2 Génération du bruit d'étiquetage

Afin d'évaluer l'influence des données mal étiquetées sur les différents jeux de données décrits précédemment, une procédure pour générer artificiellement des données mal étiquetées est nécessaire [Brodley and Friedl, 1999; Mellor et al., 2015; Teng, 1999]. Dans la littérature, deux types de bruit sur les étiquettes des échantillons sont distingués : le bruit aléatoire et le bruit déterministe [Xiao et al., 2015b].

Le bruit aléatoire se produit de manière indépendante de la donnée en entrée. Tous les échantillons ont la même probabilité d'être corrompus, et les étiquettes erronées associées aux échantillons corrompus sont aléatoires [Teng, 1999]. Dans le contexte de la cartographie de l'occupation des sols, le bruit aléatoire peut survenir lorsqu'il y a une erreur humaine ou informatique.

Contrairement au bruit aléatoire, le bruit déterministe dépend de la donnée en entrée qui aura une influence sur l'étiquette erronée attribuée. Il intervient généralement lorsque l'information utilisée pour discriminer la classe n'est pas suffisante [Kolcz and Cormack, 2009; Takenouchi et al., 2008]. Par exemple, la sortie terrain peut être programmée à la mauvaise période, la résolution spatiale de l'image satellitaire utilisée pour la photo-

Tableau 5.4 – Description des jeux de données SPOT-Landsat.

Numéro	**Nom**	**Classes utilisées**	***n***	***nbp***	**Taille du vecteur de variables**
1	NDVI 2-classes 2000-échantillons	M/T	10	100	23
2	BS-NDVI 2-classes 2000-échantillons	M/T	10	100	139
3	BS-PS 2-classes 2000-échantillons	M/T	10	100	302
4	NDVI 5-classes 5000-échantillons	B/M/O/ C/T	10	100	23
5	BS-NDVI 5-classes 5000-échantillons	B/M/O/ C/T	10	100	139
6	BS-PS 5-classes 5000-échantillons	B/M/O/ C/T	10	100	302
7	NDVI 2-classes 9600-échantillons	M/T	40	120	23
8	BS-NDVI 2-classes 9600-instance	M/T	40	120	139
9	BS-PS 2-classes 9600-échantillons	M/T	40	120	302
10	NDVI 5-classes 24000-instance	B/M/O/ C/T	40	120	23
11	BS-NDVI 5-classes 24000-échantillons	B/M/O/ C/T	40	120	139
12	BS-PS 5-classes 24000-échantillons	B/M/O/ C/T	40	120	302

n : nombre d'échantillons par polygone. *nbp* : nombre de polygones par classe.
M : Maïs. T : Tournesol. B : Blé. O : Orge. C : Colza.
BS : bandes spectrales (aérosol, bleu, vert, rouge, proche infra-rouge, moyen infra-rouge).
PS : primitives spectrales (NDVI, NDWI, MNDWI, NDBI, MNDBI, IBI, *tasseled cap*, brillance)

interprétation peut être insuffisante ou encore une forte ressemblance entre plusieurs occupations des sols peut exister. Afin de générer du bruit déterministe, des permutations entre classes très similaires en termes de variables ou des suggestions d'experts sont utilisées [Brodley and Friedl, 1999; Brodley et al., 1996; Mellor et al., 2015]. Dans le domaine de la sécurité comme le filtrage des spams ou la détection d'intrusion, le bruit déterministe correspond au bruit qui va maximiser l'erreur de classification [Biggio et al., 2011; Xiao et al., 2012]. Dans ce cas, les étiquettes associées aux échantillons corrompus amplifient les confusions déjà présentes. Plusieurs travaux montrent que le bruit déterministe est plus nuisible que le bruit aléatoire [Rebbapragada and Brodley, 2007].

En outre, le bruit peut être ajouté pour l'ensemble des classes avec le même niveau, mais seulement entre les classes majoritaires, *i.e.* celles avec les plus grands nombres d'échantillons d'apprentissage [Zhu and Wu, 2004; Zhu et al., 2003].

Dans le cadre de ces travaux, un bruit aléatoire est ajouté entre toutes les classes. Ainsi, le même nombre de données mal étiquetées est ajouté pour chaque classe. Un niveau de bruit de x % décrit le pourcentage d'échantillons mal étiquetés par classe. Par exemple, un niveau de bruit de 5 % implique que toutes les classes ont 5 % de données mal étiquetées.

Le bruit aléatoire appliqué considère que le choix de l'étiquette corrompue est équiprobable entre toutes les classes excepté celle de la vérité terrain. Ainsi, un échantillon ne peut pas être ré-étiqueté avec son étiquette initiale comme dans les travaux de Teng [1999] et Zhu and Wu [2004].

Dans ces travaux, vingt niveaux de bruit allant de 5 à 100 % par pas de 5 % sont étudiés. Le bruit est généré pour l'ensemble des jeux de variables présentés à la Section 5.2.1. Chaque niveau de bruit est indépendant, *i.e.* si un échantillon est corrompu avec un niveau de bruit à 5 %, il n'est pas nécessairement corrompu à un niveau de bruit à 10 %.

Dans la littérature, la corrélation entre les échantillons n'est pas prise en compte durant la génération du bruit : les procédures classiques ajoutent le bruit au niveau des échantillons. Dans notre contexte, l'ajout du bruit au niveau pixel n'est pas réaliste puisque les erreurs d'étiquetage impacte le plus souvent l'ensemble des échantillons appartenant à un polygone.

Afin de mieux respecter la nature spécifique des données de télédétection, la procédure proposée est donc adaptée : l'ajout du bruit est réalisé au niveau des polygones, *i.e.* tous les échantillons appartenant à un même polygone seront mal étiquetés avec la même étiquette erronée. Cependant, si l'ajout de bruit résulte en un nombre d'échantillons corrompus supérieur au niveau de bruit désiré, les échantillons à l'intérieur du polygone sont aléatoirement sélectionnés pour respecter le niveau de bruit donné.

5.2.3 Stratégie d'échantillonnage

La procédure d'échantillonnage vise à diviser les échantillons extraits de la donnée de référence en deux ensembles indépendants : un pour l'apprentissage, et l'autre pour l'évaluation. Pour chaque jeu de données, le nombre d'échantillons par classe est présenté dans la Section 5.2.1.

Pour tous les jeux de données (simulées et SPOT-Landsat), la séparation des échantillons en deux ensembles est réalisée au niveau des polygones : 50 % sont utilisés pour l'apprentissage, et 50 % pour l'évaluation. La séparation au niveau des polygones permet de s'assurer que des échantillons d'un même polygone ne soient pas utilisés à la fois pour l'apprentissage et l'évaluation. Pour chaque jeu de données, la procédure de séparation est répétée dix fois. Ainsi, les résultats ne sont pas influencés par une séparation spécifique

des données de référence.

Pour chaque jeu de données, les dix ensembles d'apprentissage indépendants sont ensuite corrompus par la procédure de génération de bruit décrite à la Section 5.2.2. Enfin, l'évaluation est réalisée au niveau du pixel avec les échantillons test qui ne contiennent pas de données mal étiquetées.

5.2.4 Configuration des algorithmes de classification

Les études proposées sont réalisées avec trois algorithmes de classification : le SVM-RBF, le SVM-linéaire et le RF. Les performances de ces algorithmes dépendent en partie de la configuration de leurs paramètres.

Pour le SVM, les paramètres C et γ ont besoin d'être optimisés lors de l'utilisation d'un noyau gaussien. Sinon, seul le paramètre C a besoin d'être optimisé pour le noyau linéaire. L'optimisation des paramètres est effectuée pour chaque jeu d'apprentissage, leurs valeurs peut donc différer en fonction du niveau de bruit.

Afin de réaliser cette optimisation, une grille de recherche logarithmique est utilisée afin de déterminer la meilleure configuration pour chaque jeu de données. Plus précisément, l'optimisation est réalisée en deux étapes en utilisant une validation croisée sur cinq partitions : une première recherche a une résolution grossière $\{2^{-5}, 2^{-4}, ..., 2^{4}\}$, et une seconde a une résolution plus fine $\left\{val \times 2^{-1}, val \times 2^{-\frac{4}{5}}, val \times 2^{-\frac{3}{5}}, ..., val \times 2^{\frac{4}{5}}\right\}$, avec val la valeur sélectionnée à la résolution grossière.

De plus, les vecteurs de variables sont standardisés en soustrayant la moyenne et en divisant par l'écart-type pour chaque variable. Cette standardisation assure que la distance à l'hyperplan ne soit pas dominée par une seule variable ayant une forte dynamique [Han et al., 2011].

Pour le RF, le Chapitre 4 a montré que les paramètres influençaient peu les performances de la classification. Ainsi, les valeurs des paramètres sont fixées de la manière suivante : le nombre d'arbres $K = 200$, le nombre de variables aléatoires sélectionnées à chaque nœud $m = \sqrt{p}$ avec p la dimension du vecteur de variables, la profondeur maximale $max\_depth = 25$ et le nombre minimal d'échantillons par nœud $min\_samples = 10$.

5.3 Résultats des expérimentations

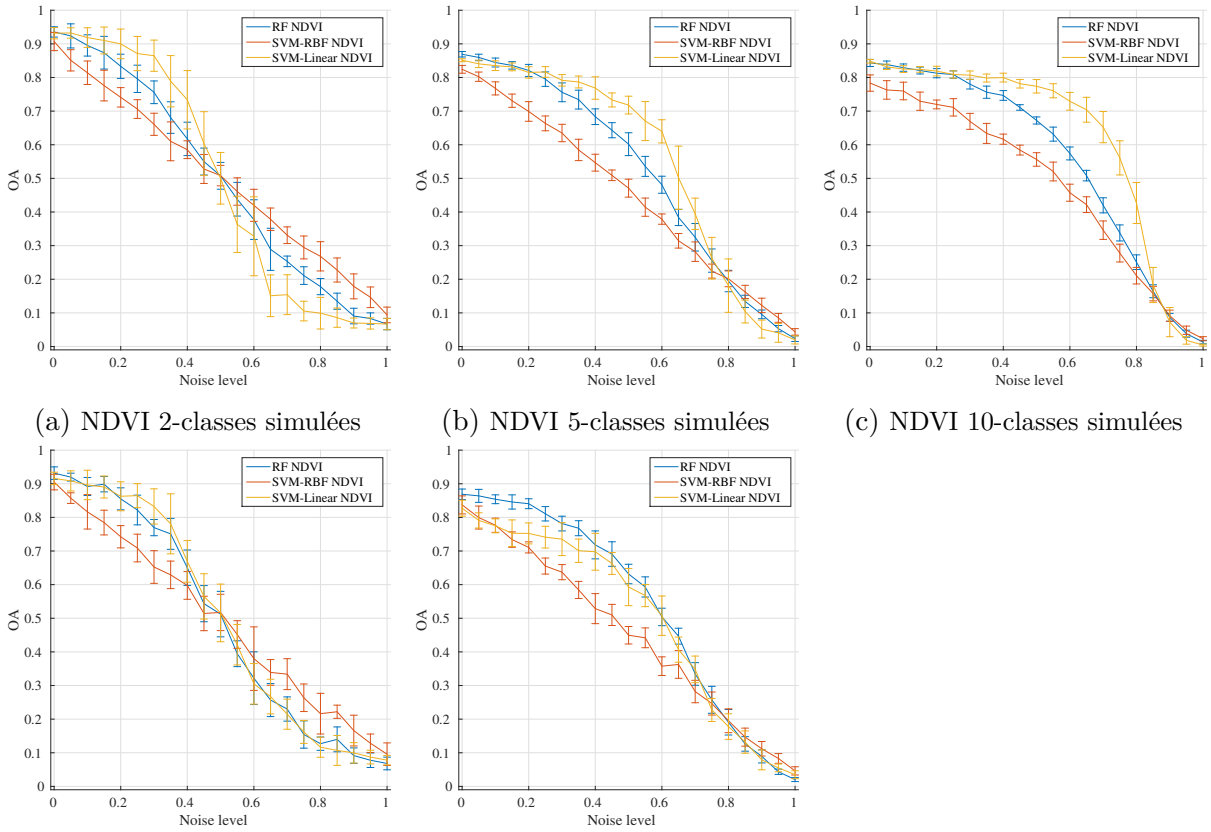
L'objectif de cette section est d'évaluer l'influence des échantillons d'apprentissage mal étiquetés sur les performances de la classification en fonction 1) du niveau de bruit, et 2) du choix de l'algorithme de classification. Pour ce faire, le RF, le SVM-RBF et le SVM-linéaire sont testés sur les données simulées et les données SPOT-Landsat.

Plusieurs configurations sont étudiées. Premièrement, l'influence du nombre de classes, du vecteur de variables et du nombre d'échantillons d'apprentissage est évaluée lors de l'ajout d'un bruit aléatoire. Deuxièmement, la complexité des algorithmes de classification est étudiée. Puis, l'influence d'un autre type de bruit est analysée. Finalement, les différences entre les trois algorithmes de classification sont discutées.

5.3.1 Influence du nombre de classes

Premièrement, l'influence du bruit aléatoire sur les valeurs d'OA obtenues par les algorithmes de classification est évaluée pour un nombre de classes différent. Pour les problèmes à deux classes, le bruit aléatoire peut être considéré comme un bruit déterministe puisqu'il est ajouté entre deux occupations des sols similaires.

La Figure 5.4 montre l'OA moyenné sur dix tirages aléatoires en fonction du niveau de bruit. La première ligne montre les résultats pour les trois jeux de données simulées (Tableau 5.2), tandis que la seconde ligne montre les résultats pour deux jeux de données SPOT-Landsat (numéro 1 et 4 du Tableau 5.4). Chaque courbe représente un algorithme de classification différent : le RF en bleu, le SVM-RBF en rouge et le SVM-linéaire en jaune. Les barres d'erreurs représentent les écarts-types calculées pour l'OA sur les dix tirages aléatoires à chaque niveau de bruit.



(a) NDVI 2-classes simulées (b) NDVI 5-classes simulées (c) NDVI 10-classes simulées

(d) NDVI 2-classes SPOT-Landsat (e) NDVI 5-classes SPOT-Landsat

FIGURE 5.4 – *Overall Accuracy* (OA) moyenné sur dix tirages aléatoires en fonction du niveau de bruit. Les barres d'erreurs représentent les écarts-types. Les résultats sont obtenus pour le RF, le SVM-RBF, et le SVM-linéaire sur les données décrites par les profils de NDVI tels que (**a**) les données simulées à deux classes ; (**b**) les données simulées à cinq classes ; (**c**) les données simulées à dix classes ; (**d**) les données SPOT-Landsat à deux classes ; (**e**) les données SPOT-Landsat à cinq classes.

La comparaison entre les trois classifieurs montre que le SVM-RBF obtient les résultats avec les plus faibles précisions. Excepté pour les données simulées à dix classes, les valeurs de l'OA pour le SVM-RBF décroît quasiment de manière linéaire avec le niveau de bruit. D'autre part, les Figures 5.4b-5.4d montrent que le RF et le SVM-linéaire obtiennent des résultats similaires pour des niveaux de bruit jusqu'à 25 %. Pour ces faibles niveaux de bruit, l'apparition de zones plates confirme la faible influence des données mal étiquetées sur les deux algorithmes de classification. Pour les niveaux de bruit supérieur à 25 %, le SVM-linéaire est plus robuste excepté pour les données SPOT-Landsat à cinq classes (Figure 5.4e).

La Figure 5.4 met aussi en évidence la cohérence des résultats obtenus pour les données simulées et les données SPOT-Landsat. En effet, les résultats sont très similaires.

Cependant, le RF est plus robuste que le SVM-linéaire pour les données à cinq classes SPOT-Landsat, ce qui n'est pas le cas sur les données simulées. Dans ce cas réel, le RF prend mieux en compte la complexité des données que le SVM-linéaire.

Pour conclure, cette première étude montre que le RF et le SVM-linéaire sont plus robustes que le SVM-RBF quelque soit le nombre de classes en présence d'un bruit aléatoire dans un espace de petite dimension.

5.3.2 Influence du vecteur de variables

Deuxièmement, l'influence du bruit aléatoire sur les performances de classification est évaluée pour différents jeux de variables. Plus précisément, les jeux de données SPOT-Landsat numérotés de 1 à 6 dans le Tableau 5.4 sont utilisés. En utilisant ces jeux de données, des problèmes de classification à deux et cinq classes sont analysés.

La Figure 5.5 montre les valeurs d'OA moyennés sur dix tirages aléatoires en fonction du niveau de bruit pour les données SPOT-Landsat (numéro 1 à 6 du Tableau 5.4). La première ligne montre les résultats pour un problème de classification à deux classes, tandis que la seconde ligne montre les résultats pour un problème de classification à cinq classes. La première colonne montre les résultats obtenus pour des échantillons décrits par les profils de NDVI, la deuxième par les bandes spectrales avec le NDVI (BS-NDVI) et la troisième par les bandes spectrales et un ensemble de primitives spectrales (BS-PS). Chaque courbe représente un algorithme de classification différent : le RF en bleu, le SVM-RBF en rouge et le SVM-linéaire en jaune.

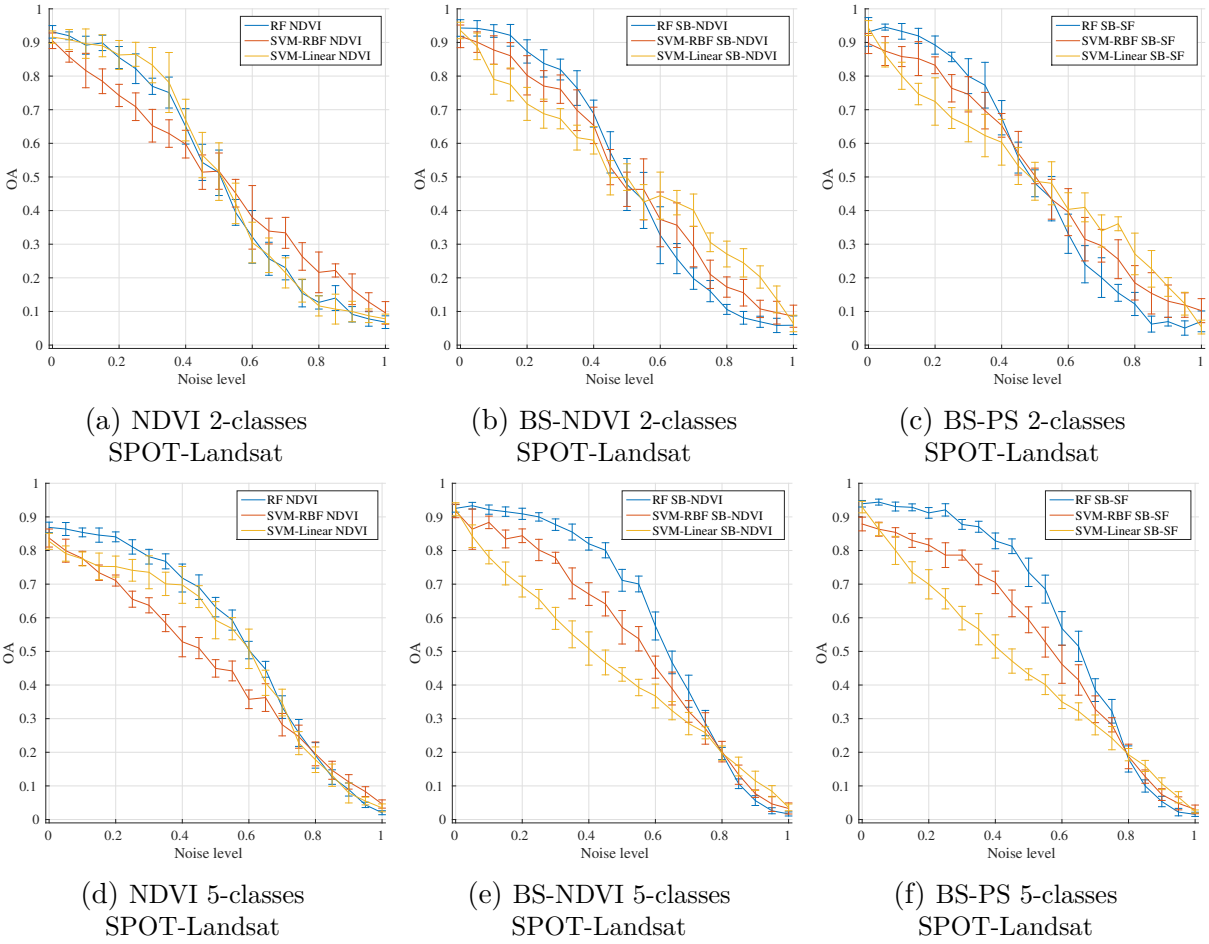
Pour les vecteurs de variables BS-NDVI et BS-PS, les performances du SVM-linéaire décroissent linéairement avec le niveau de bruit (Figures 5.5b, 5.5c, 5.5e et 5.5f). Ces faibles performances du SVM-linéaire ne sont pas observées avec l'utilisation seule du NDVI (Figures 5.4d, 5.4e). Au contraire, les performance du SVM-RBF décroissent quasi-linéairement lors de l'utilisation seule du NDVI (Figures 5.4d, 5.4e). Une analyse plus approfondie sur ces différences entre le SVM-linéaire et le SVM-RBF est fournie dans la Section 5.3.6.

Concernant l'algorithme du RF, ses performances dépassent celles des deux SVM pour les vecteurs de variables BS-NDVI et BS-PS pour des niveaux de bruit inférieurs à 50 %. De plus, l'ajout de primitives spectrales augmente les performances du RF comparé à l'utilisation du profil de NDVI seul. Cette augmentation est fortement visible sur la seconde ligne de la Figure 5.5 pour le RF. L'utilisation de l'ensemble des bandes spectrales permet d'une part de mieux décrire les échantillons, et d'autre part d'augmenter la diversité dans les arbres construits par le RF. En revanche, aucune différence significative ne peut être observée entre les vecteurs de variables BS-NDVI et BS-PS, ce qui est en accord avec les résultats du Chapitre 4.

Pour conclure, le RF est plus robuste que le SVM en présence de bruit aléatoire lorsque les bandes spectrales et les primitives spectrales sont utilisées.

5.3.3 Influence du nombre d'échantillons

Les résultats précédents montrent que l'ajout de primitives spectrales peut aider à améliorer les performances de la classification. Cependant, la grande dimension des problèmes de classification précédents peut affecter les performances des algorithmes de classification, notamment du SVM. Pour évaluer ce phénomène, les mêmes données que précédemment sont utilisées. Cependant, le nombre d'échantillons d'apprentissage et de test sont différents. Pour les données à deux classes, le nombre d'échantillons d'apprentissage et test



(a) NDVI 2-classes SPOT-Landsat

(b) BS-NDVI 2-classes SPOT-Landsat

(c) BS-PS 2-classes SPOT-Landsat

(d) NDVI 5-classes SPOT-Landsat

(e) BS-NDVI 5-classes SPOT-Landsat

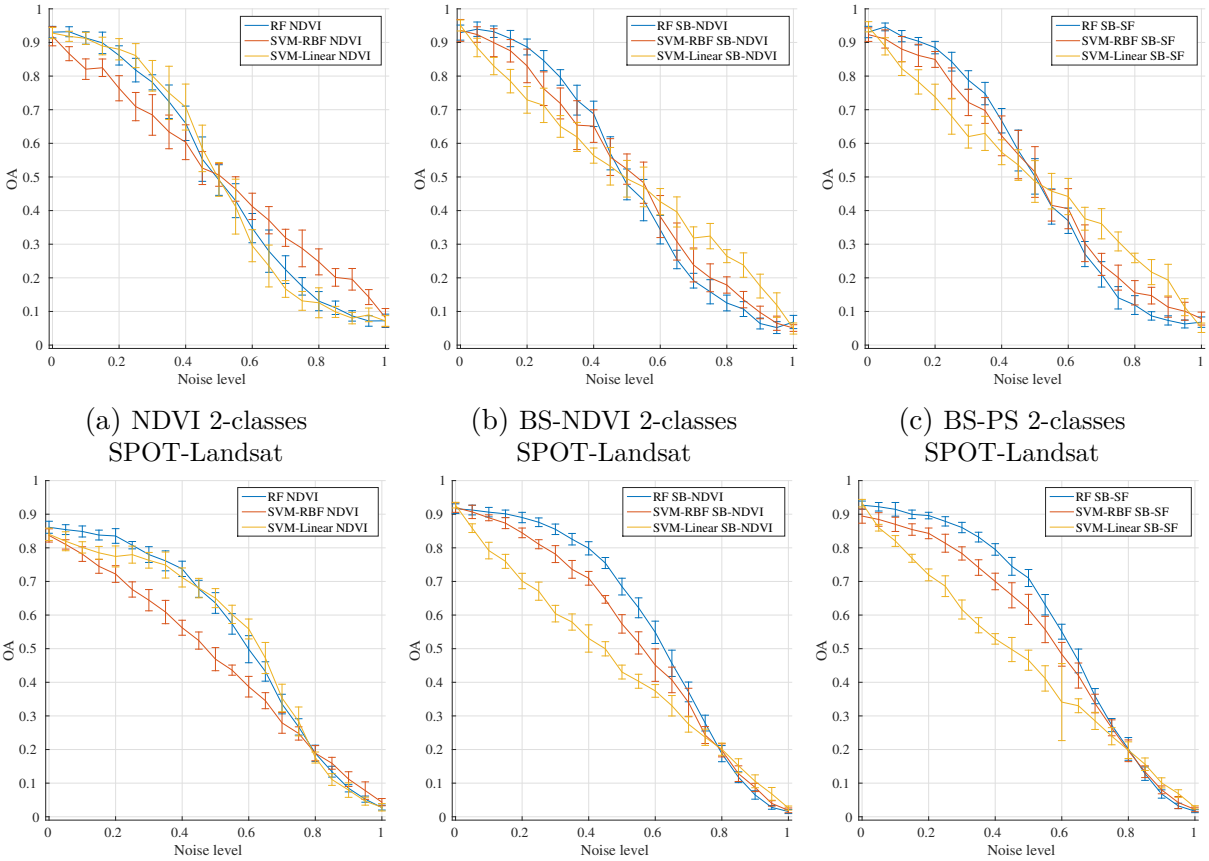
(f) BS-PS 5-classes SPOT-Landsat

FIGURE 5.5 – *Overall Accuracy* (OA) moyenné sur dix tirages aléatoires en fonction du niveau de bruit. Les barres d'erreurs représentent les écarts-types. Les résultats sont obtenus pour le RF, le SVM-RBF, et le SVM-linéaire pour les données SPOT-Landsat tel que (**a**) NDVI, deux classes ; (**b**) BS-NDVI, deux classes ; (**c**) BS-PS, deux classes ; (**d**) NDVI, cinq classes ; (**e**) BS-NDVI, cinq classes ; (**f**) BS-PS, cinq classes.

par classe est désormais de 2400 (500 dans les études précédentes). Pour les données à cinq classes, les jeux de données numérotés de 7 à 12 dans le Tableau 5.4 sont utilisés. Le nombre d'échantillons d'apprentissage et test est aussi de 2400 par classe.

Suivant le même style que précédemment, la Figure 5.6 montre les performances de classification pour différents niveaux de bruit. La première ligne montre les résultats pour des problèmes de classification à deux classes avec un total de 4800 échantillons d'apprentissage, tandis que la seconde ligne montre les résultats pour des problèmes de classification à cinq classes avec un total de 12000 échantillons d'apprentissage. De la gauche à droite, les vecteurs de variables NDVI, BS-NDVI et BS-PS sont utilisés. Chaque courbe représente un algorithme de classification différent : le RF en bleu, le SVM-RBF en rouge et le SVM-linéaire en jaune.

La Figure 5.6 met en évidence que l'augmentation du nombre d'échantillons d'apprentissage par classe n'affecte par significativement les résultats. Pour toutes les configurations – classifieur, variables, nombre de classes –, les tendances sont identiques à celles avec moins d'échantillons. Néanmoins, les performances du RF décroissent légèrement par rapport au cas de la Figure 5.5, tandis que celles des SVM augmentent faiblement.



(a) NDVI 2-classes SPOT-Landsat

(b) BS-NDVI 2-classes SPOT-Landsat

(c) BS-PS 2-classes SPOT-Landsat

(d) NDVI 5-classes SPOT-Landsat

(e) BS-NDVI 5-classes SPOT-Landsat

(f) BS-PS 5-classes SPOT-Landsat

FIGURE 5.6 – *Overall Accuracy* (OA) moyenné sur dix tirages aléatoires en fonction du niveau de bruit. Les barres d'erreurs représentent les écarts-types. Les résultats sont obtenus pour le RF, le SVM-RBF, et le SVM-linéaire pour les données SPOT-Landsat avec 4800 échantillons d'apprentissage par classe tel que (**a**) NDVI, deux classes ; (**b**) BS-NDVI, deux classes ; (**c**) BS-PS, deux classes ; (**d**) NDVI, cinq classes ; (**e**) BS-NDVI, cinq classes ; (**f**) BS-PS, cinq classes.

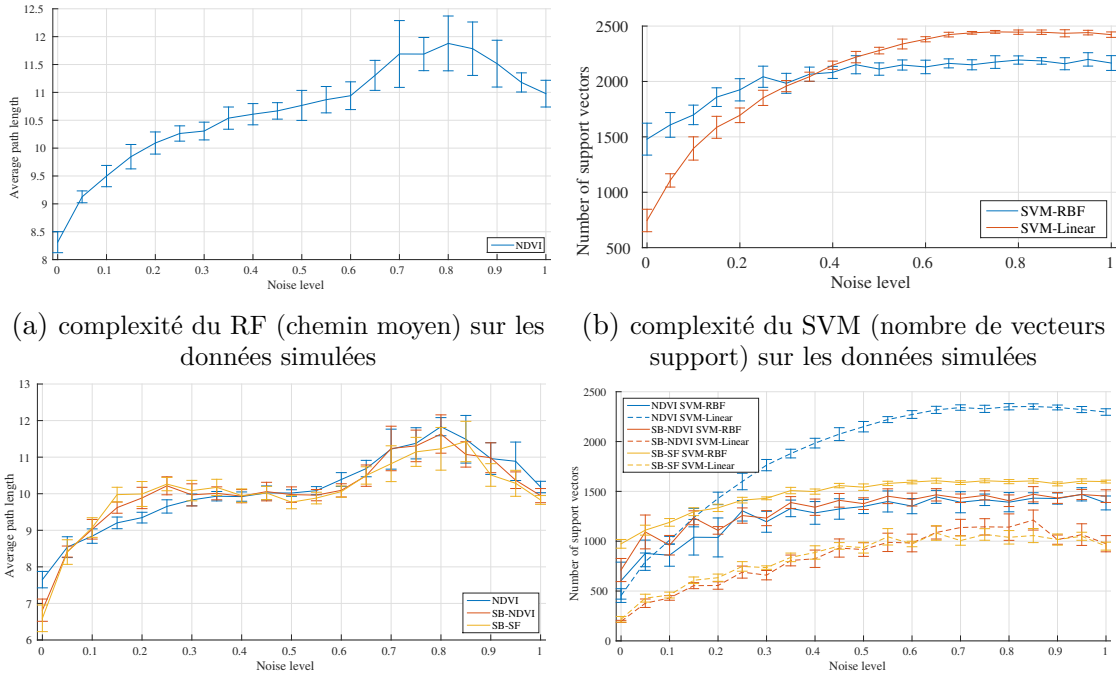
5.3.4 Complexité des algorithmes de classification

La complexité des algorithmes de classification a aussi été étudiée. L'objectif est d'analyser l'influence des données d'apprentissage mal étiquetées sur la complexité des modèles appris par les algorithmes. Dans le cas du RF, la complexité est évaluée en calculant le chemin moyen parcouru pour chaque échantillon d'apprentissage. Un petit chemin moyen signifie un modèle plus simple à apprendre, et une phase de décision plus rapide. Concernant les algorithmes du SVM, le nombre de vecteurs support (Section 2.4.1) est sélectionné pour représenter la complexité de l'algorithme. Théoriquement, le nombre de vecteurs support dépend à la fois de la distribution des données et de la valeur du paramètre de régularisation C. Un petit nombre de vecteurs support correspond à un modèle plus simple à apprendre.

Cette étude est menée sur les données simulées et SPOT-Landsat à cinq classes avec un total de 2500 échantillons d'apprentissage (jeu de données numéro 4 du Tableau 5.4) pour un bruit aléatoire. Les vecteurs de variables NDVI, BS-NDVI et BS-PS sont utilisés.

La Figure 5.7 montre la complexité des algorithmes du RF et du SVM en fonction du niveau de bruit. La première ligne correspond aux données simulées à cinq classes, tandis que la seconde ligne correspond au niveau de bruit pour les données SPOT-Landsat.

Dans le cas des données SPOT-Landsat, chaque courbe représente un vecteur de variables : NDVI en bleu, BS-NDVI en rouge, et BS-PS en jaune. Les Figures 5.7a et 5.7c montrent le chemin moyen parcouru par les échantillons d'apprentissage sur l'ensemble des arbres construits par les modèles RF. Les Figures 5.7b et 5.7d montrent le nombre de vecteurs support pour les algorithmes de SVM. Sur la Figure 5.7d, les lignes en traits pleins représentent les résultats obtenus pour le SVM-RBF, tandis que les lignes en pointillé correspondent aux résultats du SVM-linéaire.



(a) complexité du RF (chemin moyen) sur les données simulées

(b) complexité du SVM (nombre de vecteurs support) sur les données simulées

(c) complexité du RF (chemin moyen) sur les données SPOT-Landsat

(d) complexité du SVM (nombre de vecteurs support) sur les données SPOT-Landsat

FIGURE 5.7 – Complexité des algorithmes de classification en fonction du niveau de bruit pour les données simulées et SPOT-Landsat à cinq classes. Les barres d'erreurs représentent l'écart-type calculé à chaque niveau de bruit : (**a**) RF pour les données simulées ; (**b**) pour les données simulées ; (**c**) RF pour les données SPOT-Landsat ; (**d**) SVM pour les données SPOT-Landsat.

Comme attendu, la complexité des algorithmes de classification augmente avec le niveau de bruit. Par conséquent, la présence de données mal étiquetées impacte aussi négativement les besoins pour l'apprentissage, *e.g.* le temps de calcul et le stockage en mémoire.

Pour le RF, la complexité du modèle est similaire pour les différents vecteurs de variables. Lorsqu'aucune donnée mal étiquetée n'est présente (à 0 %), les vecteurs de variables BS-NDVI et BS-PS ont des complexités plus faibles que lors de l'utilisation seule du NDVI. En effet, l'utilisation d'un plus grand nombre de variables permet d'améliorer la qualité de la règle de partitionnement définit à chaque nœud. Ainsi, moins de séparations sont nécessaires pour obtenir une bonne partition des échantillons d'apprentissage. Par ailleurs, le chemin moyen parcouru pour les données simulées est supérieur à celui des données SPOT-Landsat, car la taille du vecteur de variables des données simulées est inférieure (15 contre 23 pour les profils de NDVI). À noter qu'un chemin moyen élevé n'implique pas nécessairement un temps de calcul plus élevé. En effet, le temps calculatoire pour les arbres de décision est donné principalement par le temps nécessaire pour l'étape d'apprentissage. Ce dernier sera d'autant plus long que le nombre de variables sera élevé. Dans le cas de données simulées, le chemin moyen est plus long. La phase de

décision sera donc plus longue que pour les données SPOT-Landsat. Cependant, les temps d'apprentissage sont plus courts car le nombre de variables est moins élevé.

Plus spécifiquement, les Figures 5.7a et 5.7c montrent que les courbes de complexité du RF suivent trois étapes. Ces étapes sont analysées en suivant l'évolution de l'OA donnée par la Figure 5.4. Tout d'aborde, les Figures 5.4b et 5.4e montrent une faible diminution de l'OA. Pour ces faibles niveaux de bruit, le RF compense la présence des données mal étiquetées en augmentant la complexité des modèles construits sans que cela affecte les performances de la classification. Puis, la complexité du RF reste stable alors que l'OA diminue de façon constante. Dans ce cas, le RF assimile les données mal étiquetées à des données normales. Pour les niveaux de bruit supérieurs à 50 %, la complexité du RF augmente à nouveau à cause de la présence excessive des données mal étiquetées.

Concernant la complexité du SVM-RBF, l'ajout de variables entraîne l'utilisation de plus de vecteurs support pour construire la règle de décision du modèle. Dans cette configuration, le noyau gaussien projette les données dans un espace de plus grande dimension, conduisant à la possibilité d'utiliser plus de vecteurs support. Cette différence peut donc expliquer le nombre de vecteurs support supérieur pour le SVM-RBF par rapport au SVM-linéaire. Cependant, le SVM-linéaire construit sur les données décrites par le profil de NDVI est une exception, puisque le nombre de vecteurs support est plus grand que pour le modèle du SVM-RBF construit sur les données décrites aussi par le NDVI pour les niveaux de bruit élevés.

5.3.5 Étude d'un bruit aléatoire systématique

Les précédentes études ont été menées pour un bruit aléatoire dit non-systématique. La nouvelle étiquette des échantillons corrompus est à chaque fois tirée aléatoirement. Afin d'analyser un bruit plus réaliste, un bruit aléatoire systématique est étudié [54].

Dans ce cas, les étiquettes corrompues pour une classe sont systématiquement changées vers une même autre classe. Dans cette étude, le bruit aléatoire systématique est ajouté sur les jeux de données simulées et SPOT-Landsat à cinq classes décrits par le profil NDVI [55]. Le nombre d'échantillons d'apprentissage et test par classe est de 500. Comme pour le bruit aléatoire (non-systématique), vingt niveaux de bruit allant de 5 à 100 % par pas de 5 % sont étudiés.

Le Tableau 5.5 montre la nouvelle étiquette attribuée aux échantillons corrompus en fonction de leur étiquette initiale. Pour le jeux de données simulées, le choix des étiquettes corrompues est fait de manière aléatoire. Pour les données réelles, le bruit est ajouté d'une part entre les cultures d'hiver (blé, orge, colza) et d'autre part entre les cultures d'été (maïs et tournesol) [56].

La Figure 5.8 montre les valeurs d'OA moyennées sur dix tirages aléatoires en fonction du niveau de bruit. La Figure 5.8a montre les résultats pour les données simulées, tandis que la Figure 5.8b pour les données SPOT-Landsat (jeu de données numéro 4 du Tableau 5.4). Chaque courbe représente un algorithme de classification différent : le RF en bleu, le SVM-RBF en rouge et le SVM-linéaire en jaune. Les échantillons sont décrits par leur profil de NDVI.

54. Le bruit systématique étant utilisé exclusivement dans cette partie, il n'a pas été présenté dans la Section 5.2.2.

55. Dans le cas de la classification binaire, les bruits aléatoires systématique et non-systématique sont identiques.

56. La terminologie bruit aléatoire systématique est gardée car les classes de cultures d'hiver sont bien permutées aléatoirement

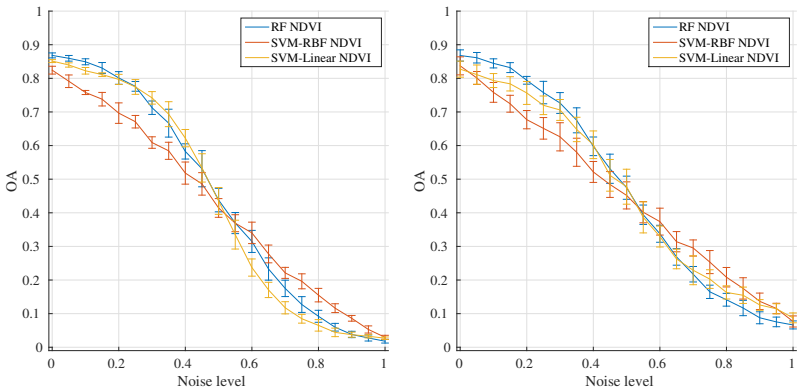
Tableau 5.5 – Choix des étiquettes corrompues dans le cas de l'ajout d'un bruit aléatoire systématique.

Données simulées		**Données réelles**	
Étiquette initiale	**Étiquette corrompue**	**Étiquette initiale**	**Étiquette corrompue**
maïs	maïs ensilage	blé	colza
maïs ensilage	sorgho	maïs	tournesol
sorgho	tournesol	orge	blé
tournesol	soja	colza	orge
soja	maïs	tournesol	maïs

L'objectif est de comparer ces résultats avec ceux obtenus dans la même configuration mais pour un bruit aléatoire non-systématique. Ainsi, les résultats des Figures 5.8a et 5.8b sont comparés avec ceux des Figures 5.4b et 5.4e respectivement.

La comparaison de ces Figures montre que le bruit aléatoire systématique a un impact similaire au bruit aléatoire non-systématique. Cependant, le bruit aléatoire systématique est plus nuisible. Dans ce cas, les valeurs d'OA restent stables seulement jusqu'à 10 % de bruit pour le RF et le SVM-linéaire. Pour les niveaux de bruit supérieur à 10 %, les valeurs d'OA décroissent plus rapidement. Par exemple, les valeurs d'OA du RF à un niveau de bruit de 40 % pour les données SPOT-Landsat sont de 72 % et 60 % pour le bruit non-systématique et systématique respectivement. Comme pour le bruit aléatoire non-systématique, les valeurs d'OA décroissent linéairement avec l'augmentation du niveau de bruit pour le SVM-RBF. Ce dernier est l'algorithme de classification le plus impacté par la présence du bruit aléatoire systématique.

Le bruit aléatoire systématique a donc une plus grande influence sur les performances de la classification qu'un bruit aléatoire non-systématique. De plus, le RF est plus robuste à ce type d'erreurs pour les données simulées et SPOT-Landsat.



(a) NDVI 5-classes simulées

(b) NDVI 5-classes SPOT-Landsat

Figure 5.8 – *Overall Accuracy* (OA) moyenné sur dix tirages aléatoires en fonction du niveau de bruit pour les algorithmes de classification RF, SVM-RBF et SVM-linéaire. Les barres d'erreur représentent les écarts-types. Le vecteur de variables NDVi est utilisé pour : (**a**) les données simulées à cinq classes ; (**b**) les données SPOT-Landsat à cinq classes.

5.3.6 Comparaison du *Random Forest* et du *Support Vector Machine*

Les résultats obtenus dans les parties précédentes montrent comment les performances du système de classification décroissent avec l'augmentation du nombre d'échantillons d'apprentissage mal étiquetés. En comparant l'ensemble des résultats, il semble que le RF est moins sensible à la présence de données mal étiquetées que le SVM. De plus, le RF a des comportements similaires en présence de données mal étiquetées quelque soit la configuration étudiée : changement du nombre de classes, augmentation du nombre d'échantillons ou modification du vecteur de variables.

Ces résultats sont conformes avec la littérature sur le sujet. Lors de la comparaison de onze algorithmes de classification incluant le SVM et le RF, Folleco et al. [2009] trouvent que le RF est l'algorithme le plus robuste à la présence de données mal étiquetées. De manière similaire, le RF obtient aussi de meilleurs résultats que le SVM-RBF en présence de bruit dans le contexte de la détection de fraude [Bhattacharyya et al., 2011]. Cependant, les valeurs des paramètres des algorithmes ne sont pas optimisées dans ces derniers travaux. Ce qui peut être plus nuisible pour le SVM-RBF. De manière similaire, il a été montré que l'algorithme du RF est plus robuste qu'un seul arbre de décision binaire en présence d'échantillons d'apprentissage mal étiquetés [Rodríguez-Galiano et al., 2012].

La robustesse du RF à la présence de données mal étiquetées peut être reliée à sa bonne capacité de généralisation. Cette dernière repose sur la construction d'un ensemble d'arbres de décision décorrélés grâce à l'utilisation d'échantillons *bootstrap*, d'un nombre suffisant d'arbres et de la procédure de séparation utilisée à chaque nœud. Ainsi, il est admis que le RF, avec un paramétrage précautionneux [Segal, 2004], est moins susceptible d'être en position de sur-apprentissage que d'autres algorithmes de classification.

Les classifieurs SVM appris avec deux noyaux différents semblent avoir des comportements complémentaires. Le SVM-linéaire est plus robuste à la présence de données mal étiquetées avec un vecteur de variables de petite taille, tandis que le SVM-RBF est plus robuste dans un espace de plus grande dimension.

Dans le cas d'un vecteur de variables de petite taille, *e.g.* l'utilisation seule du NDVI, les faibles performances du SVM-RBF peuvent être dues au sur-apprentissage. La projection des données dans un espace de plus grande dimension autorise le SVM-RBF a trouver facilement un hyperplan optimal qui suit parfaitement les données d'apprentissage. Comme vu au Chapitre 2, la valeur du paramètre γ permet de contrôler l'écart-type de la gaussienne définie dans l'équation (2.10). Lorsque la valeur de γ est petite, le modèle est trop contraint et ne peut pas capturer la structure des données. Dans ce cas, le modèle résultant se comporte similairement à un classifieur linéaire. Au contraire, un modèle appris avec une valeur de γ plus élevée est très sensible aux données d'apprentissage, *i.e.* il essaye coûte que coûte de ne pas avoir d'échantillons d'apprentissage mal-classifiés. Ainsi, le modèle perd sa capacité de généralisation, et a des difficultés à classer de nouveaux échantillons.

Les valeurs du paramètre γ obtenues pour l'algorithme SVM-RBF sont analysées pour mieux comprendre le comportement du classifieur. Le Tableau 5.6 montre les valeurs de γ sélectionnées par le processus de validation croisée sur cinq partitions pour les données SPOT-Landsat à cinq classes composées d'un total de 2500 échantillons d'apprentissage pour les trois vecteurs de variables étudiés. Les faibles performances du SVM-RBF avec le vecteur de variables NDVI (montrées à la Figure 5.5d) correspondent à de fortes valeurs de γ (> 0.1) (première ligne du Tableau 5.6). Dans ce cas, les faibles valeurs d'OA indiquent que l'algorithme du SVM-RBF fait du sur-apprentissage. Au contraire, les Fi-

gures 5.5e et 5.5f montrent des performances meilleures pour des niveaux de bruit faibles qui correspondent à des petites valeurs de γ (deuxième et troisième ligne du Tableau 5.6). Dans le cas du vecteur de variables NDVI, la dimension de l'espace est petite, et le modèle ne peut pas correctement capturer la variabilité des données. Ainsi, la procédure de validation croisée utilisée pour optimiser la valeur des paramètres C et γ semble conduire au sur-apprentissage dans le cas d'espace de petite dimension [Cawley and Talbot, 2010].

Tableau 5.6 – Valeurs du paramètre γ optimisées par la validation croisée sur cinq partitions pour l'algorithme du SVM-RBF pour les données SPOT-Landsat à cinq classes composées d'un total de 2500 échantillons d'apprentissage.

Niveau de bruit	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
NDVI	0,0825	0,5000	0,1895	0,3789	0,3299	0,4353
BS-NDVI	0,0156	0,0179	0,0179	0,0179	0,0312	0,0359
BS-PS	0,0156	0,0179	0,0156	0,0156	0,0156	0,0179

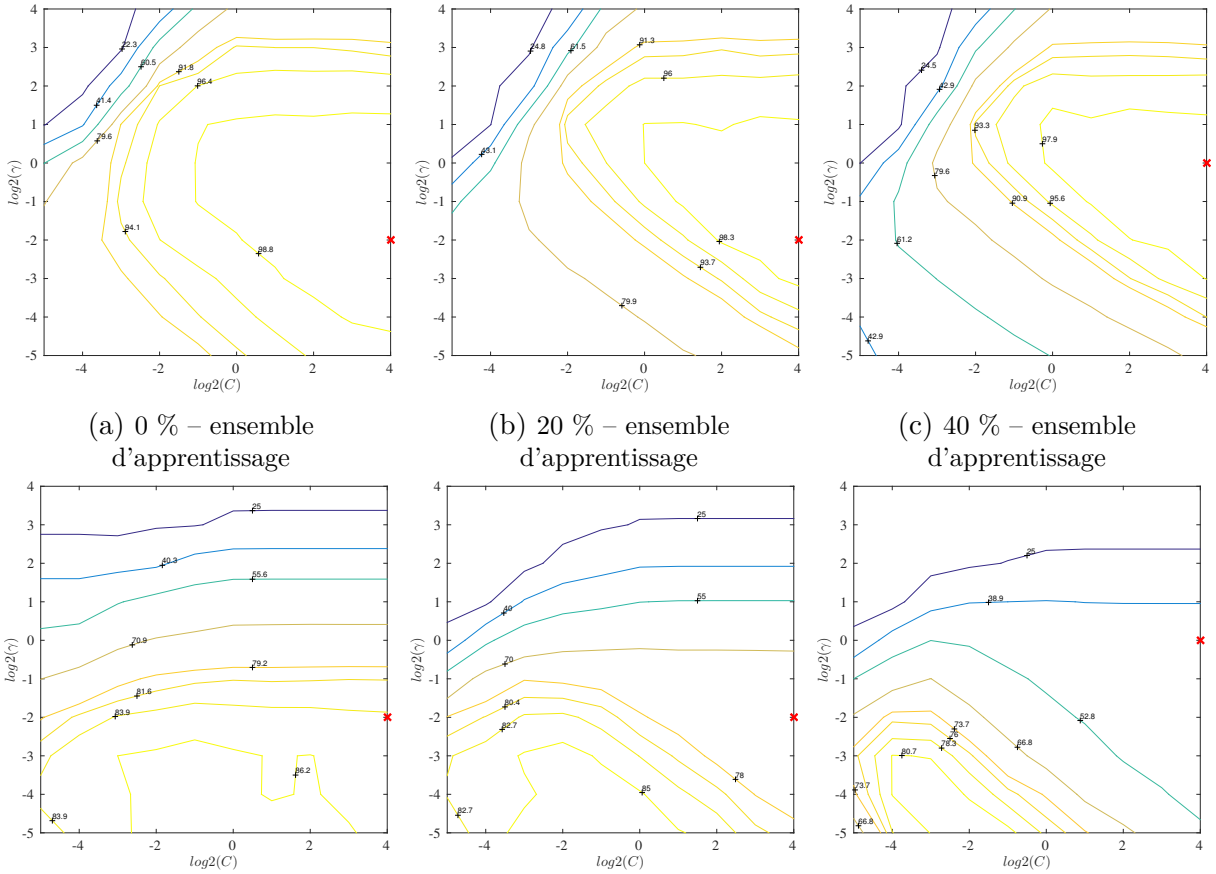
Afin de valider cette dernière hypothèse, les performances de la validation croisée sont évaluées en présence d'échantillons d'apprentissage mal étiquetés. Pour l'algorithme SVM-RBF, toutes les configurations $C - \gamma$ de la grille de recherche grossière détaillées dans la Section 5.2.4 sont étudiées. Les études sont menées sur les données SPOT-Landsat à cinq classes décrites par le profil de NDVI. Les données sont composées de 500 échantillons d'apprentissage et test par classe. L'objectif est ici de comparer les valeurs de paramètres optimales obtenues par validation croisée à partir des échantillons d'apprentissage corrompus, et celles optimales qui sont obtenues en évaluant les performances de classification avec des échantillons test indépendants et non-bruités.

La Figure 5.9 montre les résultats de la procédure de validation croisée pour l'algorithme SVM-RBF pour les niveaux de bruit à 0, 20, 40 %. Pour chaque graphique, les lignes horizontale et verticale représentent les valeurs des paramètres C et γ respectivement en échelle logarithmique. La première ligne montre les valeurs d'OA obtenues pour la procédure de validation croisée sur les échantillons d'apprentissage, tandis que la seconde ligne montre les valeurs d'OA obtenues pour des échantillons test indépendants. La croix rouge souligne la valeur optimale trouvée par la procédure de validation croisée, *i.e.* la valeur d'OA la plus forte obtenue avec les échantillons d'apprentissage.

Les figures de la première ligne (Figures 5.9a à 5.9c) montrent que les valeurs optimales des paramètres C et γ trouvées par la procédure de validation croisée (croix rouges) restent similaires pour les trois niveaux de bruit. Ces valeurs optimales correspondent à la meilleur moyenne d'OA calculée sur les partitions de validation qui contiennent aussi des échantillons mal étiquetés. Par ailleurs, les Figures 5.9a à 5.9c montrent que la valeur optimale du paramètre C est obtenue pour la valeur maximale de la grille de recherche. Néanmoins, l'augmentation de la taille de la grille de recherche pour le paramètre C conduisait à l'obtention des mêmes valeurs optimales de C.

Les Figures de la seconde ligne (Figures 5.9d à 5.9f) montrent que les valeurs optimales des paramètres, *i.e.* lorsque la valeur d'OA est maximale, sont modifiées lorsque le niveau de bruit augmente. Pour les trois niveaux de bruit, les valeurs optimales de C et γ obtenues avec les échantillons test indépendants sont plus petites que les valeurs sélectionnées par la procédure de validation croisée. Dans ce cas, le SVM-RBF fait du sur-apprentissage.

L'ensemble de ces résultats montrent que les classifieurs SVM sont sensibles à la configuration des hyper-paramètres. Les résultats montrent aussi que la procédure de validation croisée est impactée négativement par la présence de données mal étiquetées.



(a) 0 % – ensemble d'apprentissage

(b) 20 % – ensemble d'apprentissage

(c) 40 % – ensemble d'apprentissage

(d) 0 % – ensemble test

(e) 20 % – ensemble test

(f) 40 % – ensemble test

FIGURE 5.9 – Valeurs d'*Overall Accuracy* (OA) obtenues sur la grille de recherche lors de l'optimisation des paramètres du SVM sur les données SPOT-Landsat à cinq classes décrites par les profils de NDVI pour des niveaux de bruit de 0, 20 et 40 %. La première ligne montre les résultats de la validation croisée obtenues pour les échantillons d'apprentissage, tandis que la seconde ligne montre les valeurs d'OA obtenues pour des échantillons test indépendant et ne contenant pas de données mal étiquetées. La croix rouge montre la meilleure configuration obtenue pour la validation croisée. (**a**) pour un niveau de bruit de 0 % sur l'ensemble d'apprentissage ; (**b**) pour un niveau de bruit de 20 % sur l'ensemble d'apprentissage ; (**c**) pour un niveau de bruit de 40 % sur l'ensemble d'apprentissage ; (**d**) pour un niveau de bruit de 0 % sur l'ensemble test ; (**e**) pour un niveau de bruit de 20 % sur l'ensemble test ; (**f**) pour un niveau de bruit de 40 % sur l'ensemble test.

5.4 Conclusion

L'influence de la présence d'échantillons d'apprentissage mal étiquetés sur les performances du processus de classification supervisée a été analysée. À notre connaissance, les effets d'un tel bruit n'ont jamais été étudiés dans le contexte de la cartographie de l'occupation des sols à partir de séries temporelles d'images satellitaires. Seul les travaux récents de Foody et al. [2016] réalisent une étude similaire mais pour un espace à trois dimensions avec les algorithmes du SVM et des variantes.

Dans ces travaux, les trois algorithmes RF, SVM-RBF et SVM-linéaire sont comparés en présence de différents niveaux de bruit et différentes configurations de classification. Pour ce faire, les études ont été menées sur des données simulées et des données réelles. Une méthodologie spécifique a été proposée pour générer les données simulées. La méthodologie génère des profils de végétation sur une année. Par ailleurs, une nouvelle stratégie pour

la génération du bruit, basée sur les polygones, a aussi été proposée.

Dans ces travaux, différentes configurations de classification ont été testées. Les principaux résultats montrent que le RF et le SVM sont des algorithmes de classification robustes à la présence d'un faible nombre d'échantillons d'apprentissage mal étiquetés.

Premièrement, l'influence du nombre de classes a été analysée. Les jeux de données à deux classes correspondent à des problèmes difficiles ou les performances décroissent plus rapidement que pour les jeux de données à cinq et dix classes. Cette première étude a montré la cohérence des données simulées pour lesquelles des résultats similaires à ceux des données réelles ont pu être observés. De plus, l'algorithme SVM-RBF fait du sur-apprentissage lorsque le vecteur de variables utilisé décrit seulement le profil de NDVI sur une année.

Ainsi, une seconde étude avec l'ajout de variables plus complexes a été réalisée. L'addition des bandes spectrales en plus du NDVI améliore généralement les performances de classification. Contrairement à la première étude, le SVM-RBF est plus robuste que le SVM-linéaire.

Ensuite, la complexité des algorithmes de classification a été étudiée, mettant en évidence l'importance de la qualité des échantillons d'apprentissage sur les temps de calcul et les besoins en mémoire. Comme attendu, la complexité des algorithmes augmente lorsque le nombre de données mal étiquetées augmente.

Puis, il a été observé qu'un bruit aléatoire systématique impacte plus fortement les performances de classification qu'un bruit aléatoire non-systématique.

Enfin, les performances des algorithmes de classification ont été comparées montrant que l'algorithme du RF est plus robuste à la présence de données mal étiquetées. Au contraire, les performances du SVM sont plus sensibles à la présence de données mal étiquetées dues à la difficulté de paramétrer correctement les algorithmes. Plus précisément, il a été observée que la procédure de validation croisée est aussi impactée par la présence de données mal étiquetées, conduisant au sur-apprentissage des modèles construits.

Quatrième partie

Détection des données mal étiquetées

Chapitre 6

Détection des données mal étiquetées

Sommaire

6.1 Détection de données mal étiquetées 120
6.1.1 Méthodes basées sur les algorithmes de classification 121
6.1.2 Méthodes basées sur la proximité entre les échantillons 123
6.1.3 *Clustering*, méthodes de graphes et arbre de décision 130
6.1.4 Limitations 131
6.2 Détection de données mal étiquetées avec le *Random Forest* 132
6.2.1 Score d'*outliers* du *Random Forest* 132
6.2.2 Nouvelles mesures de proximité 134
6.3 Présentation des expérimentations 135
6.3.1 Données satellitaires et données de référence 135
6.3.2 Méthodes étudiées 136
6.3.3 Évaluation 137
6.4 Résultats des expérimentations 139
6.4.1 Paramétrage des méthodes basées sur la distance 140
6.4.2 Comparaison des performances 146
6.4.3 Performances des méthodes pour les données Sentinel-2 151
6.5 Conclusion 155

La présence de données mal étiquetées est un problème récurrent lors de la classification de données satellitaires. En particulier, le Chapitre 5 a montré que l'utilisation d'échantillons d'apprentissage mal étiquetés est fortement pénalisante pour l'apprentissage supervisé. C'est pourquoi, des méthodes permettant de détecter les données mal étiquetées sont nécessaires. Si les échantillons mal étiquetés sont identifiés, il est ensuite possible de proposer un cadre méthodologique qui prenne en compte cette information afin d'améliorer les performances de classification [57].

Dans le contexte de la cartographie de l'occupation des sols sur de grandes étendues, l'existence de données mal étiquetées est due à l'utilisation de données de référence imparfaites. En effet, les échantillons d'apprentissage sont bien souvent extraits d'anciennes bases de données pour obtenir un nombre suffisant d'échantillons sur l'ensemble de la surface à cartographier. Cependant, l'utilisation de ces données anciennes pour classer

57. L'étude d'une telle méthodologie fait l'objet du Chapitre 7.

des images satellitaires plus récentes conduit à la présence de nombreuses données mal étiquetées parmi les échantillons d'apprentissage.

Une stratégie précise pour identifier les données mal étiquetées est de recourir à un expert sur le terrain ou à un opérateur SIG. Ces derniers peuvent en théorie vérifier et corriger manuellement chaque échantillon de la donnée. Sans connaissance *a priori* sur la localisation et le type d'erreurs présentes, ces approches sont coûteuses en temps et en argent.

Ce chapitre s'intéresse aux méthodes de détection automatiques de données mal étiquetées. Dans un premier temps, les méthodes de détection des données mal étiquetées de l'état-de-l'art sont présentées. Les limitations de ces approches sont identifiées, et une méthode s'appuyant sur la structure des arbres de décision binaire apprise par l'algorithme du RF est détaillée dans une deuxième partie. Ensuite, l'intérêt de cette méthode est démontré en l'évaluant et la comparant à quelques méthodes plus traditionnelles de l'état-de-l'art. Enfin, les conclusions du chapitre sont données.

6.1 Détection de données mal étiquetées

Cette section présente les méthodes de l'état-de-l'art ayant pour objectif d'identifier les données mal étiquetées dans un ensemble d'échantillons. Récurrent dans de nombreux domaines d'application, ce problème a pourtant peu été étudié en télédétection [Frénay and Verleysen, 2014]. Ainsi, les méthodes présentées ici viennent de divers domaines, et ne sont pas spécifiques aux données satellitaires.

Comme défini dans le Chapitre 5, une donnée mal étiquetée est une donnée pour laquelle la classe de la référence ne correspond pas à la réalité du terrain. Le vecteur de variables d'une telle donnée sera donc différent de celui des échantillons ayant la même classe de référence. Ainsi, les données mal étiquetées peuvent être vues comme des données aberrantes, *i.e.* des exceptions à une règle générale. Dans la littérature, ces exceptions sont généralement appelées « *outliers* » – « *an observation which deviates so much from other observations as to arouse suspicious that it was generated by a different mechanism* » [Grubbs, 1969; Hawkins, 1980]. Un *outlier* est donc une observation qui se comporte différemment des autres observations. Cette notion subjective [Liu et al., 2002] admet de nombreuses traductions en fonction du domaine d'application : exceptions, données aberrantes, données extrêmes, données atypiques ou encore anomalies [58].

La notion d'*outlier* est parfois permutée avec celle de données mal étiquetées. Bien qu'une majorité des *outliers* soit des données mal étiquetées, tous les *outliers* ne sont pas des données mal étiquetées et inversement. Par exemple, il est possible que le tournesol sur une parcelle ait des difficultés à se développer dues à une mauvaise qualité de sol, la sécheresse, *etc.* La signature spectrale sur cette parcelle va donc différer de celles des autres parcelles de tournesol. Cette parcelle peut être alors vue comme un *outlier* alors que ce n'est pas une donnée mal étiquetée. À l'inverse, si des erreurs d'étiquetage existent entre deux classes très similaires, par exemple le blé et l'orge, ces données mal étiquetées ne seront jamais vues comme des *outliers*.

Cependant, la similarité entre les deux notions a conduit à utiliser des méthodes de détection d'*outliers* pour détecter des données mal étiquetées [Xiong et al., 2006]. Plusieurs travaux passent en revue l'ensemble des méthodes de détection d'*outliers* [Aggarwal, 2013;

58. Dans la littérature anglaise, les termes *outlier* et *anomaly* sont parfois inter-changés même si techniquement il existe une différence entre les deux termes [Janssens, 2013]. Une anomalie est une observation qui agit de manière différente par rapport à un comportement attendu selon un expert du domaine, tandis qu'un *outlier* est une donnée qui est significativement différente des autres données.

Chandola et al., 2009; Han et al., 2011; Hodge and Austin, 2004; Pimentel et al., 2014]. En suivant les travaux de Chandola et al. [2009] et Han et al. [2011] sur les *outliers*, quatre catégories de méthodes de détection de données mal étiquetées sont identifiées :

1. méthodes basées sur les statistiques,
2. méthodes basées sur des algorithmes de classification,
3. méthodes basées sur la notion de proximité,
4. méthodes basées sur le *clustering*.

Les méthodes statistiques sont les plus anciennes. Elles sont traditionnellement divisées entre les approches paramétriques et non-paramétriques. Les approches paramétriques supposent que les données correctement étiquetées suivent une distribution spécifique. En l'absence d'information, des distributions normales ou des GMM [Yamanishi et al., 2000] sont souvent considérées. Les échantillons qui s'éloignent de ces lois sont alors identifiés comme des *outliers* [Barnett and Lewis, 1974; Hawkins, 1980]. Malheureusement, ces méthodes sont très sensibles au modèle choisi, et au nombre de paramètres. Si les hypothèses sur le modèle sont trop restrictives, de nombreux échantillons peuvent être détectés comme des *outliers*. Au contraire, si elles sont trop générales, le modèle peut sur-apprendre les données et manquer des *outliers*.

En revanche, les méthodes statistiques non-paramétriques font aucune hypothèse sur la distribution suivie par les données. Par exemple, l'analyse d'histogramme évite de choisir une distribution en utilisant la probabilité d'occurrences des échantillons [Goldstein and Dengel, 2012]. Les échantillons qui ne suivent pas la distribution donnée par l'histogramme sont alors considérés comme des *outliers*. Ces approches sont généralement peu performantes lorsque les échantillons sont décrits dans des espaces de grande dimension et qu'ils ont une forte variance [Chandola et al., 2009].

Ainsi, ce chapitre se focalise sur les trois autres types de méthodes non-paramétriques : 1) les méthodes basées sur les algorithmes de classification (Section 6.1.1), 2) les méthodes basées sur la proximité entre échantillons (Section 6.1.2), et 3) les méthodes basées sur le *clustering* et d'autres types de structure (Section 6.1.3).

L'objectif ici n'est pas de présenter une liste exhaustive des méthodes, mais de répertorier les différentes stratégies existantes dans la littérature pour la détection de données mal étiquetées. Pour plus détails, une revue des méthodes de détection de données mal étiquetées est réalisée par Frénay and Verleysen [2014].

6.1.1 Méthodes basées sur les algorithmes de classification

Afin de détecter les données mal étiquetées, il est possible d'utiliser les algorithmes de classification supervisée pour résoudre le problème binaire « correctement étiqueté *versus* mal étiqueté ». Cependant, l'obtention d'une donnée de référence exhaustive répertoriant les données correctement étiquetées et les données mal étiquetées est longue, coûteuse et difficile à obtenir. De plus, la nature fortement déséquilibrée du problème, les données mal étiquetées étant minoritaires, complique l'apprentissage du modèle de classification.

Ainsi, seules les méthodes de classification d'*outliers* non-supervisée sont décrites ici, *i.e.* aucune connaissance *a priori* sur les données mal étiquetées n'est nécessaire. Parmi ces méthodes, deux stratégies sont identifiées.

La première stratégie repose sur les algorithmes de classification à une classe. En supposant que les échantillons appartiennent à la même classe « correctement étiqueté », l'objectif est de construire des modèles capables d'identifier les échantillons appartenant

à cette classe. Les échantillons en désaccord avec cette règle sont alors considérés comme mal étiquetés.

La seconde stratégie consiste à entraîner un algorithme de classification supervisée pour le problème « occupation des sols ». Dans ce cas, les données mal étiquetées seront les échantillons mal prédits par le classifieur, ou les échantillons pour lesquels le classifieur est peu confiant ou encore les échantillons qui font augmenter la complexité du modèle.

Algorithmes de classification à une classe

Les méthodes de classification à une classe sont utilisées pour caractériser les données appartenant à une même classe. L'apprentissage du modèle est réalisé à partir de l'ensemble de ces données. Ces méthodes sont régulièrement utilisées pour la détection des *outliers*. Parmi les approches les plus utilisées, les méthodes *One Class - Support Vector Machine* (OC-SVM) [Schölkopf et al., 2000, 2001] et le *Support Vector Data Description* (SVDD) [Tax and Duin, 1999, 2004] s'appuient sur l'algorithme des SVM et le *Replicator Neural Network* (RNN) sur un réseau de neurones. Ces trois méthodes sont décrites dans la suite.

Dans l'algorithme OC-SVM, le principe du SVM est étendu au cas de données non-étiquetées. L'objectif est de construire une frontière de décision, toujours un hyper-plan, pour laquelle la majorité des échantillons d'apprentissage sont du même côté. Un échantillon est jugé comme mal étiquetée si il n'est pas du même côté de la frontière de décision que la majorité des échantillons d'apprentissage.

De façon similaire, l'algorithme SVDD se base sur le principe du SVM. L'objectif est d'englober un maximum des échantillons d'apprentissage dans une hypersphère de plus petit volume possible (*minimum enclosing ball* en anglais). Un échantillon qui se trouve à l'extérieur de cette hypersphère est vu comme un *outlier*.

L'utilisation de ces deux méthodes présente deux difficultés principales [Manevitz and Yousef, 2001] : 1) le réglage des paramètres permettant de contrôler la position de l'hyperplan ou le rayon de l'hypersphère est complexe, et 2) ces méthodes ne sont pas performantes dans des espaces de grande dimension où les échantillons sont très éparpillés [Manevitz and Yousef, 2001].

Un dernier exemple de méthode de classification à une classe est l'algorithme RNN spécialement implémenté pour la détection d'*outliers* [Hawkins et al., 2002; Williams et al., 2002]. Ce réseau de neurones comportant trois couches cachées a pour spécificité d'avoir les mêmes données sur les couches d'entrée et de sortie. L'objectif est d'obtenir un réseau équivalent à la fonction identité. Les couches cachées du réseau de neurones sont chargées de « compresser » l'information des données d'entrée. Ainsi, l'erreur de reconstruction de l'algorithme du réseau est utilisée pour détecter les données mal étiquetées.

Méthodes basées sur la prédiction, la mesure de confiance et la complexité

Dans les méthodes décrites ici, un algorithme de classification supervisée est appris avec l'ensemble des échantillons. Ces échantillons sont étiquetés par la classe fournie par la donnée de référence. Les données mal étiquetées sont alors identifiées en analysant l'algorithme de classification, *e.g.* les mauvaises prédictions ou l'augmentation de la complexité du modèle.

Une première stratégie consiste à évaluer la probabilité que l'échantillon appartienne à la classe fournie par la donnée de référence en se basant sur les prédictions de l'algorithme de classification. Plus elle est faible, plus l'échantillon est susceptible d'être mal étiqueté. Par exemple, la méthode *Pair-Wise Expectation Maximization* (PWEM) estime cette

probabilité en utilisant l'algorithme de *clustering* Espérance-Maximisation (EM) [Rebbapragada and Brodley, 2007]. La spécificité de cette méthode est d'appliquer l'algorithme EM entre chaque paire de classes.

Dans l'approche de Büschenfeld and Ostermann [2012], les *outliers* sont les échantillons qui ont des faibles valeurs de marge. La marge correspond ici à la différence entre les deux plus fortes valeurs de probabilité d'appartenance aux classes, calculées par un SVM.

Un deuxième stratégie consiste à analyser la complexité de la règle de décision apprise par l'algorithme de classification [Gamberger and Lavrač, 1997; Gamberger et al., 1999]. Par hypothèse, la présence d'*outliers* augmente la complexité du classifieur. Par exemple, le nombre de variables nécessaire pour construire la règle de décision peut augmenter. Dans ce contexte, le filtre de saturation proposé par Gamberger and Lavrač [1997] identifie comme mal étiquetés les échantillons qui complexifient le plus la règle de décision.

Il est aussi possible d'ajouter de l'information contextuelle dans le modèle appris par l'algorithme de classification. Si un échantillon a une étiquette différente de celle de ses voisins dans l'image, il aura une plus grande probabilité d'être mal étiqueté [Jia et al., 2014a].

Une dernière stratégie consiste à analyser les désaccords entre plusieurs algorithmes de classification. Ces méthodes sont détaillées dans le Chapitre 7.

6.1.2 Méthodes basées sur la proximité entre les échantillons

Les méthodes décrites dans cette partie s'appuient sur la notion de proximité entre échantillons, aussi appelée similarité [59]. La proximité entre deux échantillons est une mesure qui quantifie leur similitude. Généralement, elle est mesurée en analysant le voisinage des échantillons. Un échantillon est alors identifié comme étant un *outlier* s'il est isolé, *i.e.* peu d'échantillons sont présents dans son voisinage.

Parmi les méthodes basées sur la proximité, deux types de méthodes existent [Han et al., 2011] : 1) les méthodes basées sur la distance, et 2) les méthodes basées sur la densité. Elles sont toutes les deux décrites dans la suite.

Basée sur la distance

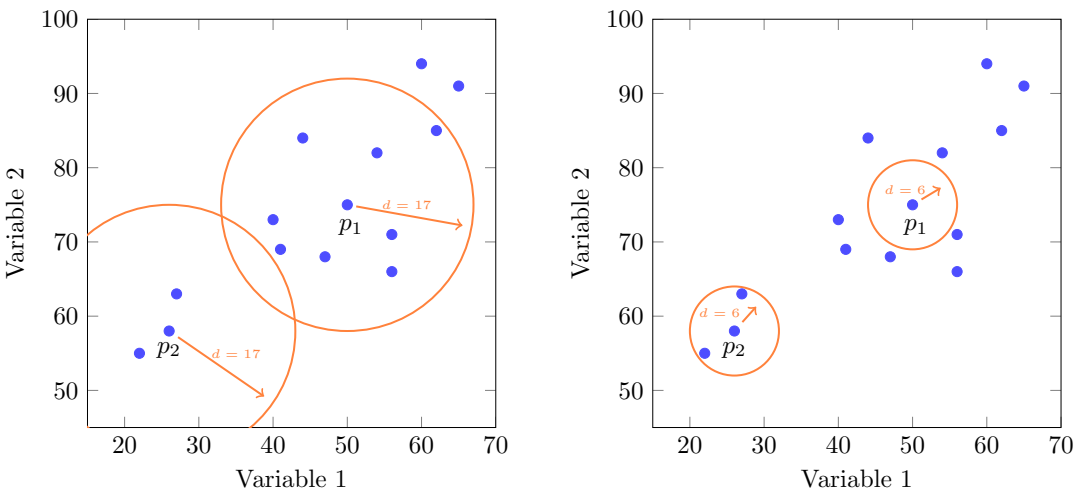
Les méthodes de détection d'*outlier* basées sur la distance définissent le voisinage d'un échantillon en fonction d'une distance. Formellement, Knorr and Ng [1998] proposent la définition suivante : un échantillon p dans une base de données T est un *outlier* si au moins n échantillons [60] dans T sont à une distance supérieure à d de p. Autrement dit, l'ensemble $DB(n, d)$ des *outliers* est défini tel que :

$$DB(n, d) = \{p \mid |\{q \in T \mid dist(p, q) < d\}| \leq n\}, \tag{6.1}$$

avec $dist(p, q)$ la distance entre les échantillons p et q, et $|.|$ le cardinal.

La Figure 6.1 illustre cette définition dans un espace à deux dimensions pour des échantillons représentés par les points bleus. Deux configurations différentes pour les paramètres n et d sont étudiées. L'objectif de ces deux exemples est de déterminer si les échantillons p_1 et p_2 sont des *outliers*. Les échantillons à l'intérieur des cercles orange représentent les voisins des échantillons p_1 et p_2 qui sont à une distance de p_1 et p_2 inférieure à d.

59. La terminologie dépend du domaine d'application.
60. Dans la définition originale, le paramètre utilisé est τ qui représente une fraction d'échantillons



(a) $d = 17$ et $n = 4$

(b) $d = 6$ et $n = 3$

FIGURE 6.1 – Définition d'un *outlier* basée sur la distance selon Knorr and Ng [1998].

Dans le cas de la Figure 6.1a, le paramètre n est fixé à 4 échantillons et la distance d est égale à 17. En appliquant la définition de Knorr and Ng [1998], l'échantillon p_1 n'est pas un *outlier* car neuf échantillons sont à l'intérieur du cercle orange, *i.e.* plus de quatre échantillons sont à une distance inférieure à 17 de p_1. En revanche, l'échantillon p_2 est un *outlier* puisque seulement trois échantillons sont à une distance inférieure à 17 de p_2.

Dans le cas de la Figure 6.1b, la distance d est fixée à une plus petite valeur, *i.e.* $d = 6$. Ainsi, le rayon des cercles orange est diminué. De plus, la valeur du nombre d'échantillons n est fixée à 3. Avec cette nouvelle configuration, l'échantillon p_1 n'a plus aucun plus proche voisin à une distance inférieure à 6. Il est donc identifié comme un *outlier*. Au contraire, l'échantillon p_2 n'est pas un *outlier* car trois échantillons sont à l'intérieur du cercle orange.

La définition de Knorr and Ng [1998] donnée par l'équation (6.1) ne permet pas de quantifier à quel point un échantillon est un *outlier*. En effet, la mesure obtenue est binaire : chaque échantillon est identifié *outlier* ou non. Dans cette définition, la configuration du paramètre n permet de jouer sur le nombre d'*outliers* identifiés. Comme le niveau de bruit présent dans les données est *a priori* inconnu, la valeur optimale de n est difficile à configurer. La Figure 6.1 montre que la configuration du paramètre d est également difficile sans connaissance sur la dispersion des valeurs de distance. La valeur optimale de d va principalement dépendre de la définition de la distance utilisée pour calculer la similarité entre échantillons ainsi que des valeurs des variables décrivant les échantillons.

Afin d'éviter la configuration du paramètre de distance d, une solution consiste à utiliser la notion de distance aux plus proches voisins. Dans ce contexte, le voisinage de l'échantillon p, noté $N_k(p)$, est l'ensemble des k échantillons qui ont les plus petites distances avec p.

Intuitivement, le cardinal de $N_k(p)$, noté $|N_k(p)|$, doit être égal à k. Cependant, dans certains cas particuliers, $|N_k(p)|$ peut être supérieur à k. Considérons un échantillon p avec six voisins dont un voisin est à une distance d'une unité, deux voisins à deux unités, et trois voisins à trois unités. Dans ce cas là, le calcul du cardinal de l'ensemble $|N_k(p)|$ pour différentes valeurs de k donne

— $|N_1(p)| = 1$,

— $|N_2(p)| = |N_3(p)| = 3$,

— $|N_4(p)| = |N_5(p)| = |N_6(p)| = 6$.

Lorsque k est égal à 2, 4 ou 5, le cardinal de $N_k(p)$ est donc supérieur à la valeur de k. Dans des cas spécifiques, les k plus proches voisins de p peuvent donc être plus nombreux que k. Pour simplifier le discours, ces cas spécifiques ne seront pas considérés dans la suite.

Chaque échantillon peut alors être caractérisé par la distance avec ces plus proches voisins. Dans la littérature, la valeur de la distance $D^k(p)$ au k-ième plus proche voisin d'un échantillon p est couramment utilisée. Pour chaque échantillon, la valeur de la distance $D^k(p)$ est donc différente. Avec la notation $D^k(p)$, le voisinage de l'échantillon p s'exprime de la manière suivante :

$$N_k(p) = \left\{ q \in T \setminus \{p\} \mid dist(p,q) \leq D^k(p) \right\}. \tag{6.2}$$

Le principal avantage de la distance $D^k(p)$ est qu'elle ne nécessite pas une connaissance de la dispersion des valeurs de distance entre échantillons. De plus, le paramètre k est plus facile à interpréter que le paramètre d utilisée dans la définiton de Knorr and Ng [1998]. Ainsi, la majorité des méthodes de détection d'*outliers* basées sur la distance s'appuient sur la notion du k-ième plus proche voisin.

Par exemple, Ramaswamy et al. [2000] propose la définition suivante : un échantillon p est un *outlier* s'il n'y a pas plus de $n-1$ autres échantillons de la base de données T dont les distances au k-ième plus proche voisin sont strictement supérieures à $D^k(p)$. Autrement dit, l'ensemble des *outliers* $DB(n,k)$ correspond aux n échantillons ayant les plus grandes distances au k-ième plus proche voisin :

$$DB(n,k) = \left\{ p \;\middle|\; \left| \left\{ q \in T \mid D^k(q) > D^k(p) \right\} \right| < n-1 \right\}. \tag{6.3}$$

La Figure 6.2 illustre la définition de Ramaswamy et al. [2000] pour huit échantillons. Dans cet exemple, la valeur du paramètre k est fixée à 2. Tous les cercles représentent la distance au deuxième plus proche voisin pour l'ensemble des échantillons. Dans cet exemple, la valeur du paramètre n est fixée à 3. Par conséquent, les échantillons détectés comme *outlier* correspondent aux trois échantillons p_1, p_2 et p_3 avec les plus grands cercles, qui sont représentés en rouge.

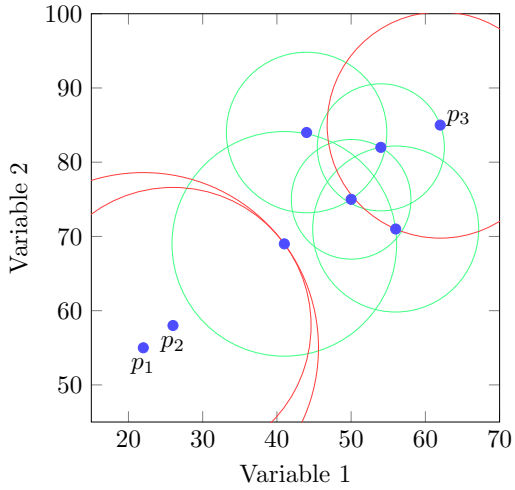
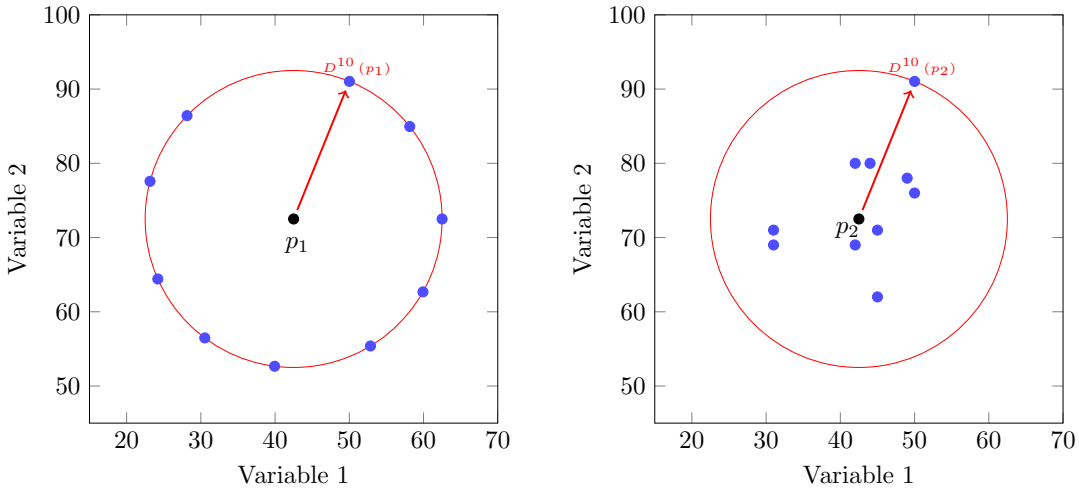


FIGURE 6.2 – Définition d'un *outlier* selon Ramaswamy et al. [2000]. ($k = 2$ et $n = 3$)

La définition proposée par Ramaswamy et al. [2000] permet de calculer un score d'*outlier* $O_{k\text{NN}}$ pour l'échantillon p. La valeur du score d'*outlier* indique alors le degré d'anomalie[61] de l'échantillon p. Ce score d'*outlier* correspond ici à la distance au k-ième plus proche voisin :

$$O_{k\text{NN}}(p) = D^k(p). \tag{6.4}$$

La définition d'*outlier* de l'équation (6.3) ne prend pas en compte la densité locale de p. La notion de densité locale est précisée par la Figure 6.3. L'objectif des Figures 6.3a et 6.3b est d'évaluer si les échantillons p_1 et p_2 sont des *outliers* lorsque la distance au dixième plus proche voisin est utilisée. Pour ce faire, les voisins de p_1 et p_2 sont représentés par les points bleus et la distance au dixième plus proche voisin D^{10} est représentée par un cercle rouge.



(a) Faible densité autour de l'échantillon p_1 (b) Forte densité autour de l'échantillon p_2

Source : inspiré de Angiulli and Pizzuti [2005]

FIGURE 6.3 – Illustration de la notion de densité pour deux échantillons p_1 et p_2 pour lesquels les distances au dixième plus proche voisin sont identiques ($D^10(p_1) = D^10(p_2)$).

Pour cet exemple, si le score d'*outlier* est calculé avec l'équation, alors les scores d'*outlier* sont identiques pour p_1 et p_2. Cependant, la densité locale de l'échantillon p_1 est plus faible que celle de p_2 : les voisins de p_1 sont équidistants de p_1, tandis que ceux de p_2 sont plus proches. Afin de prendre en compte cette différence, la méthode kNNW propose de calculer un nouveau score d'*outlier* $O_{k\text{NNW}(p)}$ [Angiulli and Pizzuti, 2005]. Ce score calcule la somme des distances de p à ses k plus proches voisins :

$$O_{k\text{NNW}(p)} = \sum_{q \in N_k(p)} dist(p, q). \tag{6.5}$$

Si l'équation (6.5) est utilisée dans l'exemple de la Figure 6.3, l'échantillon p_1 aura un score kNNW plus élevé que l'échantillon p_2. En effet, l'ensemble des dix plus proches voisins de p_1 sont plus éloignés.

Bien que la méthode kNNW prenne en compte la densité locale de p, elle n'utilise pas la densité locale des plus proches voisins de p. L'importance de la densité locale du voisinage est expliqué par la Figure 6.4. La Figure 6.4 montre un ensemble d'échantillons décrits

61. Anomalie est utilisée ici pour la traduction du terme *outlierness*.

dans un espace à deux dimensions. Les échantillons verts sont composés de l'échantillon p et de deux *clusters* principaux n'ayant pas la même densité. Dans cet exemple, on souhaiterait détecter l'échantillon p comme étant un *outlier*. Même s'il est proche du *cluster* de droite, sa densité locale est différente de celle de ses plus proches voisins.

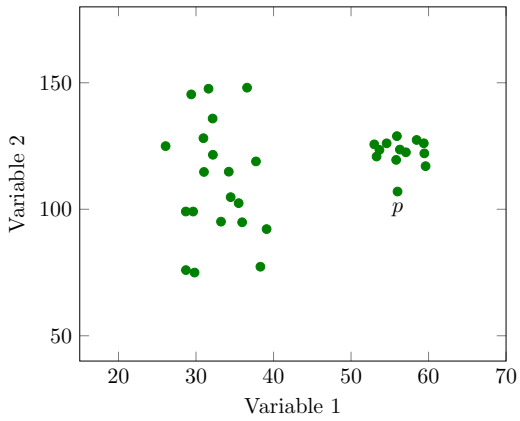


FIGURE 6.4 – Distribution d'échantillons décrits par deux *clusters* de différentes densité.

Dans l'exemple de la Figure 6.4, les méthodes kNN et kNNW peuvent détecter p comme étant un *outlier* en utilisant une petite valeur pour le paramètre k. Mais dans ce cas là, certains échantillons du *cluster* de gauche seront aussi vus comme des *outlier* car les échantillons de ce *cluster* sont plus dispersés. Nous verrons dans la suite comment les méthodes basées sur la densité permettent de palier à cette problématique.

Par ailleurs, les scores d'*outlier* calculés par les méthodes kNN et kNNW ont aussi pour inconvénient de ne pas être bornés. Les valeurs de ces scores dépendent de la distance utilisée pour le calcul de similarité et des variables décrivant les échantillons. Ainsi, les scores d'*outlier* sont difficilement comparables entre différents jeux de données. Une solution proposée consiste à calculer un score d'*outlier* borné entre 0 et 1 : c'est la méthode *Stochastic Outlier Selection* (SOS) [Janssens, 2013].

Dans la méthode SOS, le score d'*outlier* $O_{sos}(p)$ d'un échantillon p correspond à la probabilité que p soit mal étiqueté. Plus précisément, il est calculé comme le produit des dissimilarités entre l'échantillon p et les autres échantillons q dans T. La mesure de dissimilarité entre les échantillons p et q est le complément à la mesure d'affinité $\text{aff}(p, q)$. Formellement, la probabilité $O_{sos}(p)$ s'exprime de la manière suivante :

$$O_{sos}(p) = \prod_{\substack{q \in T \\ q \neq p}} \left(1 - \frac{\text{aff}(p, q)}{\sum_{\substack{r \in T \\ r \neq p}} \text{aff}(p, r)} \right). \tag{6.6}$$

La mesure d'affinité est proportionnelle à la mesure de similarité $dist(p, q)$, et est bornée entre 0 et 1 [62]. Elle est définie de la manière suivante :

$$\text{aff}(p, q) = \begin{cases} exp\left(\frac{-dist(p,q)^2}{2\sigma_p^2}\right) & \text{si } p \neq q, \\ 0 & \text{sinon,} \end{cases} \tag{6.7}$$

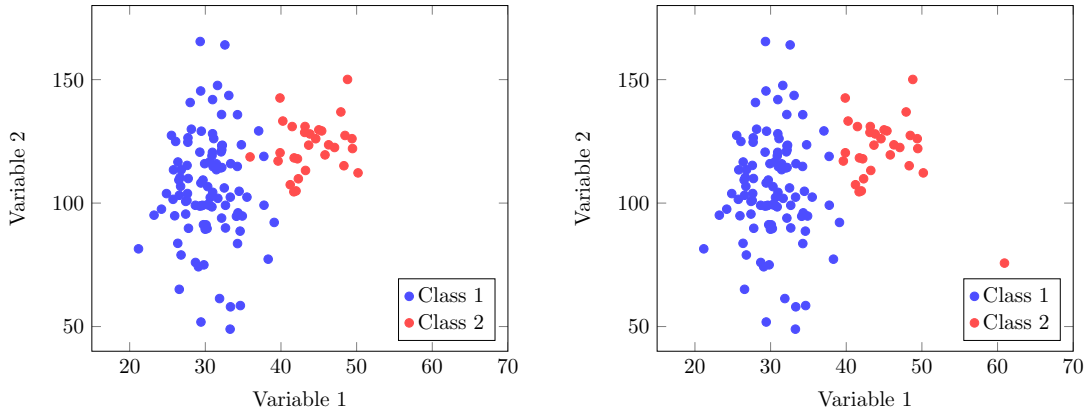
62. La mesure d'affinité a notamment été utilisée dans des algorithmes de *clustering* [Frey and Dueck, 2007].

où σ_p représente la variance de l'échantillon p. Cette variance est calculée pour chaque échantillon en utilisant la notion de perplexité h, issue de la théorie de l'information. Pour rappel, la perplexité h est liée à la notion d'entropie de Shannon H par la formule $h = 2^H$. Dans cette méthode, la perplexité est un paramètre défini par l'utilisateur qui va permettre de déduire la variance σ_p de chaque échantillon. L'estimation de la valeur de σ_p est détaillé dans les travaux de Janssens [2013]. La valeur de la variance s'adapte à la densité locale de chaque observation : une forte densité implique une faible variance. Théoriquement, l'affinité d'un échantillon p est donnée par 90 % de ses h plus proche voisins. La perplexité h est donc un équivalent du paramètre k des méthodes kNN et kNNW.

Les méthodes kNN, kNNW et SOS ont été initialement développées pour la détection d'*outliers*, *i.e.* pour des échantillons appartenant à la même population. Cependant, l'objectif de ces travaux est la détection de données mal étiquetées parmi les échantillons d'apprentissage qui appartiennent à différentes classes. Ainsi, les échantillons apportent une information supplémentaire qui est leur classe d'appartenance fournie par la donnée de référence. Ainsi, il est nécessaire d'adapter les méthodes décrites précédemment aux problèmes « multi-classes ».

La stratégie la plus simple consiste à appliquer les méthodes kNN, kNNW et SOS séparément pour chaque classe. Cependant, cette stratégie ne garantit pas l'obtention de scores d'*outliers* ou de probabilités comparables entre chaque classe [Kriegel et al., 2011]. De plus, les valeurs optimales des paramètres k et h peuvent dépendre de la classe considérée.

Afin d'exploiter l'étiquette des échantillons, une autre stratégie consiste à utiliser des méthodes basées sur la distance qui tirent profit de l'information fournie par la donnée de référence (*class outlier mining* en anglais). Dans la littérature, ces méthodes cherchent à identifier deux types d'*outliers* : 1) les *semantic outliers* [He et al., 2002], et 2) les *cross-outliers* [Papadimitriou and Faloutsos, 2003]. La Figure 6.5 illustre ces deux concepts pour des données appartenant à deux classes (bleue et rouge) dans un problème à deux dimensions.



(a) *Semantic outlier* (b) *Cross-outlier*

Source : inspiré de Nezvalová et al. [2015]

FIGURE 6.5 – Illustration des notions de *semantic* et de *cross-outlier*.

Dans cet exemple, la Figure 6.5a montre un échantillon rouge proche des deux classes, mais encerclé par des échantillons bleus. Il est donc très probable que la véritable étiquette de l'échantillon rouge soit bleue, et que donc cet échantillon soit un *semantic outlier*. La Figure 6.5b illustre le principe de *cross-outlier* avec l'échantillon rouge éloigné (en bas

à droite) de l'ensemble des échantillons. Dans ce cas là, les plus proches voisins de cet échantillon sont rouges, mais ils sont tous très éloignés. Typiquement un *semantic outlier* peut être une donnée mal étiquetée, tandis qu'un *cross-outlier* peut être une donnée mal étiquetée ou représenter une autre apparence minoritaire de la classe.

Afin de prendre en compte ses deux aspects, Hewahi and Saad [2007] proposent le score *Class Outlier Factor* (COF) pour un échantillon donné p. Le score d'*outlier* O_{COF} combine : 1) la distance aux échantillons appartenant à la même classe, 2) la classe d'appartenance de ses plus proches voisins, et 3) son degré d'isolement.

Dans la littérature, d'autres méthodes cherchent à détecter uniquement les *semantic outlier* en utilisant la classe fournie par la donnée de référence. Ces méthodes sont connues sous le nom de méthodes d'édition. Elles comparent la classe de référence de l'échantillon p avec celles de ses k plus proches voisins. Plus précisément, la méthode d'édition ENN identifie comme *outliers* les échantillons pour lesquels une majorité de leurs k plus proches voisins appartiennent à une autre classe [Wilson, 1972]. Comme dans la définition proposée par [Knorr and Ng, 1998], le score d'*outliers* des méthodes d'édition est binaire : 0 si l'échantillon est correctement étiqueté, 1 sinon.

Deux variantes itératives de cet algorithme ont été proposées [Tomek, 1976] :

1. L'algorithme *Repeated Edited Nearest Neighbor* (RENN) qui réitère ENN pour un k fixé en éliminant à chaque itération les échantillons détectés comme étant des *outlier*. Il s'arrête automatiquement lorsque plus aucun échantillon n'est détecté comme étant un *outlier*.
2. L'algorithme All$k$NN qui applique itérativement ENN en incrémentant le paramètre k de 1 jusqu'à une valeur définie par l'utilisateur. Pour chaque k, les échantillons détectés comme *outlier* sont éliminés de l'ensemble d'apprentissage.

Les méthodes d'édition réalisent la détection d'*outlier* localement en supposant que la majorité des voisins d'un échantillon mal étiqueté soit correctement étiquetée. Malheureusement, cette hypothèse n'est pas toujours vérifiée lorsque le nombre de données mal étiquetées est important.

Basée sur la densité

Les méthodes basées sur la densité se proposent d'adresser une des limitations des méthodes basées sur la distance, à savoir que les *outlier* auront des densités locales différentes des densités locales de leurs plus proches voisins (Figure 6.4). L'idée est de calculer le score d'*outlier* d'un échantillon en fonction des densités locales de ses plus proches voisins. Le score d'*outlier* d'un échantillon est alors évalué en comparant sa densité locale avec les densités locales de ses plus proches voisins. Les échantillons ayant des densités sensiblement inférieures à celles de leurs voisinages sont identifiés comme des *outliers*.

Parmi ces méthodes, la plus connue est le *Local Outlier Factor* (LOF) [Breunig et al., 2000]. Elle calcule le score d'*outlier* $O_{\text{LOF}}(p)$ d'un échantillon p en deux étapes. La première étape consiste à calculer une densité locale pour l'échantillon p et pour ses k plus proches voisins, tandis que la seconde étape consiste à comparer les différentes densités locales.

Pour ces calculs, une définition de la densité locale d'un échantillon p en fonction de ses k plus proches voisins est nécessaire. Pour ce faire, la distance $rdist_k(p,q)$ (*reachability distance* en anglais) entre les deux échantillons p et q est introduite :

$$rdist_k(p,q) = \max\left(D^k(q), dist(p,q)\right). \tag{6.8}$$

Si l'échantillon p est situé à une distance supérieure à $D^k(q)$ de l'échantillon q, alors la distance $rdist_k(p,q)$ correspondra à la distance réelle entre p et q. Par contre, si l'échan-

tillon p est proche q, alors la distance $rdist_k(p, q)$ sera bornée et égale à $D^k(q)$. Cette notion de $rdist(p, q)$ est illustrée par la Figure 6.6.

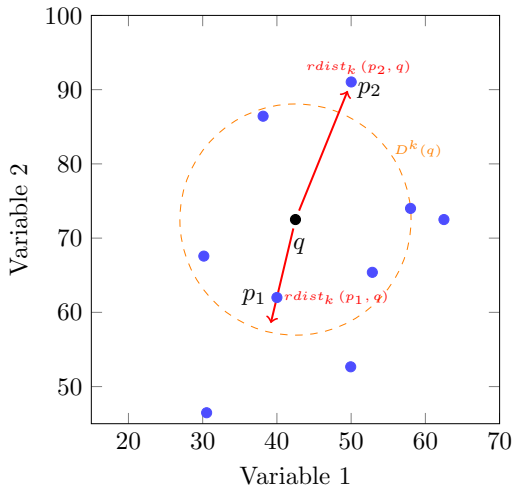


FIGURE 6.6 – Illustration de la distance $rdist$ ($k = 5$). Le cercle orange représente la distance au k-ième plus proche voisin de l'échantillon q.

Dans la Figure 6.6, la distance $rdist$ est évaluée entre l'échantillon q et les deux échantillons p_1 et p_2. La valeur du paramètre de voisinage k est fixée à 5. La distance $rdist_k(p_1, q)$ est égale à la distance $D^k(q)$ car l'échantillon p_1 est à une distance de q inférieure à $D^5(q)$. En revanche, la distance $rdist_k(p_2, q)$ est égale à la distance réelle $dist(p_2, q)$ entre les deux échantillons car l'échantillon p_2 est à une distance de q supérieure à $D^5(q)$.

Pour chaque échantillon p, la distance $rdist_k$ est calculée entre p et l'ensemble de ses k plus proches voisins (noté $N_k(p)$). Ces calculs permettent d'estimer $lrd_k(p)$ qui correspond à la densité locale de l'échantillon p :

$$lrd_k(p) = \frac{1}{\dfrac{\sum\limits_{q \in N_k(p)} rdist_k(p, q)}{|N_k(p)|}}. \tag{6.9}$$

Le score d'*outlier* de l'échantillon p est ensuite calculé en comparant la densité locale de p avec celle de ses k plus proches voisins. Plus précisément, le calcul du score d'*outlier* $O_{LOF}(p)$ s'exprime de la manière suivante :

$$O_{LOF}(p) = \frac{\sum\limits_{q \in N_k(p)} \dfrac{lrd_k(q)}{lrd_k(p)}}{|N_k(p)|}. \tag{6.10}$$

Par ailleurs, l'algorithme LOF connaît de nombreuses variantes avec par exemple la modification du calcul de densité. Une revue de ces variantes est proposée par Schubert et al. [2014].

6.1.3 *Clustering*, méthodes de graphes et arbre de décision

Outre les approches basées sur les algorithmes de classification et la proximité, des méthodes de détection d'*outliers* calculent les scores d'*outliers* en se basant sur des algorithmes de *clustering*, de méthodes de graphes ou d'arbres de décision.

L'objectif des méthodes de *clustering* est de regrouper les échantillons similaires au sein d'un même *cluster*. Pour ces méthodes, différentes stratégies sont proposées pour calculer le score d'*outlier*. Par exemple, la méthode *Cluster-Based Local Outlier Factor* (CBLOF), proposée par He et al. [2003], considère que les *outliers* sont soit les échantillons appartenant aux petits *clusters*, soit les échantillons isolés au sein des grands *clusters*. Outre le choix de l'algorithme de *clustering*, cette méthode requiert donc la configuration d'un seuil pour différencier les petits des grands *clusters*. He et al. [2004] proposent une variante de CBLOF qui prend en compte l'étiquette des échantillons donnée par la donnée de référence.

Une autre stratégie pour détecter les *outliers* est de calculer la proximité entre échantillons en utilisant une structure de type graphe. Dans ces approches, chaque échantillon représente un nœud du graphe où les nœuds sont connectés en fonction d'un critère. Par exemple, dans les ϵ-graphes, deux échantillons sont connectés si leur distance est inférieure à un paramètre ϵ pré-défini. La détection d'*outliers* se fait alors en cherchant les échantillons isolés dans le graphe. La méthode proposée par Garcia et al. [2015] se base sur un ϵ-graphe où toutes les connexions entre échantillons n'appartenant pas à la même classe sont supprimées. Dans ce cas, le score d'*outlier* correspond au degré d'isolement des échantillons dans le graphe, qui est calculé en comptant le nombre de connexions pour chaque échantillon.

Au lieu du graphe, il est aussi possible d'utiliser une structure de type arbre de décision binaire afin de visualiser les relations entre les échantillons. Ainsi, Liu et al. [2012] proposent l'algorithme iForest (*isolation Forest* en anglais) qui construit un ensemble d'arbres de type *Extremely Randomized Trees* (Section 2.4.2). Contrairement au principe utilisé dans le RF, chaque arbre est construit avec seulement une sous-partie des échantillons (tirage aléatoire sans remise). Par hypothèse, un *outlier* est un échantillon écarté rapidement lors de la construction d'un arbre de décision binaire. Les *outliers* correspondent donc aux échantillons qui ont des chemins courts dans les arbres. Ainsi, le score d'*outlier* de la méthode iForest est fonction du chemin moyen parcouru par les échantillons sur l'ensemble des arbres.

6.1.4 Limitations

Cet état-de-l'art a répertorié trois grandes familles de détection de données mal étiquetées. Elles sont non-paramétriques et ne nécessitent pas de connaissances *a priori* sur les échantillons mal étiquetés. La première catégorie s'appuie sur des algorithmes de classification, la deuxième exploite la proximité entre les échantillons, et la troisième se base sur des algorithmes de *clustering*, de méthodes de graphes ou d'arbres de décision.

Parmi les méthodes basées sur la proximité, une majorité s'appuie sur la distance des échantillons à leurs k plus proches voisins. La valeur de k est très difficile à déterminer et dépend principalement du niveau de bruit présent dans les données. Généralement une grande valeur de k implique la prise en compte d'un grand voisinage qui assure la stabilité de l'algorithme à défaut de sa précision. Au contraire, une plus petite valeur de k peut rendre l'algorithme très instable.

De plus, l'efficacité de ces méthodes dépend de la définition de la distance utilisée pour mesurer la similarité entre échantillons. Dans notre contexte, la définition d'une distance pertinente est difficile. Cette distance doit prendre en compte toutes les complexités des séries temporelles d'images satellitaires, *i.e.* les fortes variabilités des variables utilisées (réflectance, indices normalisés, primitives temporelles, *etc.*), les espaces de grandes dimension et aussi la redondance de l'information entre variables.

Ainsi, l'approche du iForest est attractive puisqu'aucune définition de distance est nécessaire. L'utilisation de la structure des arbres de décision binaire pour identifier les échantillons mal étiquetées permet théoriquement de travailler dans des espaces de grandes dimensions avec un grand nombre d'échantillons [Liu et al., 2012]. Cependant, l'algorithme iForest nécessite de configurer le nombre d'échantillons dans les sous-ensembles utilisés pour la construction de chaque arbre. Or ce paramètre n'est pas trivial à configurer [Bandaragoda, 2015]. Par ailleurs, cette méthode fait l'hypothèse que les *outliers* sont des échantillons qui ont des chemins courts dans les arbres. Malheureusement, cette hypothèse n'est pas toujours vérifiée puisque les échantillons faciles à classer peuvent aussi avoir des chemins courts dans les arbres.

Par ailleurs, les méthodes utilisant les résultats des algorithmes de classification semblent prometteuses pour une faible présence de données mal étiquetées. Cependant, pour des forts niveaux de bruit, les méthodes semblent peu robustes puisque l'apprentissage de l'algorithme de classification, généralement un SVM, est alors fait à partir d'échantillons mal étiquetés qui perturbent la phase d'apprentissage (Chapitre 5).

Dans notre problématique, la méthode de détection de données mal étiquetées doit être 1) robuste, 2) facile à paramétrer, et 3) nécessiter le moins possible d'hypothèse sur la distribution suivie par les échantillons. Pour ce faire, nous utilisons la méthode de détection d'*outliers* définie par Breiman pour détecter les données mal étiquetées. Bien que rarement utilisé en télédétection, cette méthode présente plusieurs avantages. Comme pour l'algorithme iForest, la définition d'une distance mesurant la similarité entre des échantillons décrits par des vecteurs de variables de grande dimension n'est pas nécessaire. La section suivante est dédiée à la description de cette méthode.

6.2 Détection de données mal étiquetées avec le *Random Forest*

La méthode de détection d'*outliers* présentée ici s'appuie sur la notion de similarité entre échantillons. Ainsi, un échantillon sera considéré comme un *outlier* s'il est dissimilaire aux autres échantillons de sa classe. Afin de définir la notion de similarité, Breiman propose d'utiliser la structure des arbres construits par le RF. Dans ce cas là, deux échantillons sont similaires s'ils suivent le même parcours dans les arbres du RF. Breiman généralise cette idée en proposant une mesure de proximité calculée sur l'ensemble des arbres de la forêt.

Une première partie détaille le calcul du score d'*outlier* proposé par Breiman, et une seconde partie introduit deux nouvelles mesures de proximité proposées dans le cadre de ces travaux.

6.2.1 Score d'*outliers* du *Random Forest*

Afin de définir la notion de similarité définie par Breiman, considérons deux échantillons p et q qui tombent dans les nœuds terminaux $n_k(p)$ et $n_k(q)$ respectivement pour le k-ième arbre de la forêt. Breiman propose de calculer une mesure de similarité $sim_k(p, q)$ entre les deux échantillons p et q qui s'exprime de la manière suivante :

$$sim_k(p, q) = \begin{cases} 1 & \text{si } n_k(p) = n_k(q), \\ 0 & \text{sinon.} \end{cases} \tag{6.11}$$

La mesure de similarité est utilisée pour calculer la proximité $prox(p, q)$ entre deux échantillons p et q dans l'ensemble des arbres de la forêt. Plus précisément, la proximité correspond à la moyenne des mesures de similarité sur l'ensemble des K arbres de la forêt :.

$$prox(p, q) = \frac{1}{K} \sum_{k=1}^{K} sim_k(p, q). \tag{6.12}$$

Deux échantillons qui tombent souvent dans les mêmes nœuds terminaux ont une proximité proche de 1, au contraire des échantillons qui ne finissent jamais ensemble ont une proximité proche de 0.

La mesure de proximité a été utilisée pour différentes applications dont le positionnement multidimensionnel (*multidimensional scaling* en anglais) [Kruskal, 1964] et le *gap-filling* de données manquantes. Par ailleurs, cette mesure de proximité a aussi été utilisée comme mesure de distance dans les algorithmes de *clustering* [Shi and Horvath, 2006], de PPV ou encore du SVM pour la définition du noyau [Englund and Verikas, 2012]. Dans le contexte de la détection d'*outlier*, la mesure de proximité décrite à l'équation (6.12) a aussi été utilisée pour le calcul d'un score d'*outlier*[63].

Pour calculer le score d'*outlier* $O_{raw}(p)$ de l'échantillon p ayant pour classe de référence c_r, les proximités entre p et tous les échantillons appartenant à la classe c_r sont calculées. Soit $N_{c_r}(p)$ l'ensemble des échantillons à la même classe c_r que p : $N_{c_r}(p) = \{q \in T \setminus \{p\} \mid c_r(q) = c_r(p)\}|$. Avec ces notations, le score $O_{raw}(p)$ est donné par l'équation suivante :

$$O_{raw}(p) = \frac{|N_{c_r}(p)|}{\sum\limits_{q \in N_{c_r}(p)} prox(p, q)^2}, \tag{6.13}$$

Ainsi, plus les mesures de proximité sont faibles entre l'échantillon p et les échantillons appartenant à la même classe, plus la mesure $O_{raw}(p)$ sera forte. Dans ce cas là, l'échantillon p est tombé peu souvent dans les mêmes nœuds que les autres échantillons de sa classe.

La mesure $O_{raw}(p)$ est ensuite normalisée afin de pouvoir comparer le score d'*outlier* entre toutes les classes. Cette normalisation est donc effectuée pour chacune des classes. Pour ce faire, la valeur médiane *med* et la déviation moyenne absolue autour de la médiane MAD des scores d'*outlier* O_{raw} sont calculées pour chaque classe. Soit $\boldsymbol{\Omega^p_{raw}}$ l'ensemble des valeurs de O_{raw} pour les échantillons appartenant à la même classe que p (p inclus). Le score d'*outlier* $O_{RF}(p)$ de l'échantillon p peut alors s'écrire de la manière suivante[64] :

$$O_{RF}(p) = \frac{O_{raw}(p) - med(\boldsymbol{\Omega^p_{raw}})}{MAD(\boldsymbol{\Omega^p_{raw}})}. \tag{6.14}$$

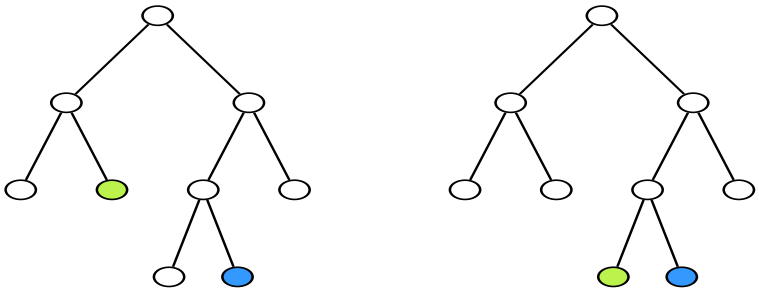
63. La méthode décrite ici, proposée par Breiman, n'a pas fait l'objet d'une publication avant son décès. Le détail est cependant disponible dans un rapport technique : www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf. Par ailleurs, le code Fortran développé par Breiman est aussi disponible à www.stat.berkeley.edu/~breiman/RandomForests/cc_software.htm

64. Dans le code initial de Breiman (Fortran90), la valeur de $MAD(\boldsymbol{\Omega^p_{raw}})$ est calculée telle que : $MAD(\boldsymbol{\Omega^p_{raw}}) = \frac{1}{N_{c_r}(p)+1} \sum\limits_{q|cl(q)=cl(p)} \min\left(|O_{raw}(q) - med(\boldsymbol{\Omega^p_{raw}})|, 5med(\boldsymbol{\Omega^p_{raw}})\right)$. Ce raffinement a été gardé dans l'implémentation proposée.

6.2.2 Nouvelles mesures de proximité

Comme montré dans la partie précédente, le calcul du score d'*outlier* à partir de la structure des arbres construits par le RF repose sur la mesure de similarité *sim* définie dans l'équation (6.11). Or la mesure de similarité est « stricte » : le résultat est 1 ou 0. Son utilisation peut donc mener à une perte d'information.

Afin d'illustrer cette affirmation, considérons les deux exemples (cas **A** et **B**) de la Figure 6.7. Pour chaque cas, deux échantillons sont considérés : un qui tombe dans le nœud vert, et un autre dans le nœud bleu. Pour les deux cas, les mesures de similarité calculées entre les deux échantillons par l'équation (6.11) sont nulles. Cependant, les deux échantillons sont plus proches dans le cas **B**. Leurs chemins se séparent seulement à l'avant dernier nœud. Alors que les échantillons du cas **A** paraissent plus éloignés, ils se séparent dès la racine.



(a) Cas **A** (b) Cas **B**

FIGURE 6.7 – Exemples d'arbres de décision. Les nœuds verts et bleus représentent les nœuds terminaux où deux échantillons sont tombés. Pour ces deux cas, la mesure de proximité telle que proposée par Breiman est nulle.

Basée sur cette observation, il est possible de proposer de nouvelles mesures. La plus simple consiste à compter le nombre de branches qui séparent les deux échantillons. Dans l'exemple de la Figure 6.7, cinq branches séparent les deux échantillons pour le cas **A**, et deux branches pour le cas **B**. Intuitivement, cette mesure permet bien de caractériser l'éloignement des échantillons dans l'arbre. Cependant, cette distance est difficilement comparable entre deux arbres de la forêt. En effet, les arbres construits par le RF n'ont pas tous la même profondeur en fonction des échantillons *bootstrap* utilisés. Ainsi un plus grand nombre de branches séparera les échantillons lorsque les arbres seront construits plus en profondeur. Il parait donc important de proposer des mesures normalisées entre 0 et 1 pour chaque arbre.

Par ailleurs, la profondeur du plus petit ancêtre commun, *lowest common ancestor* (LCA) en anglais, entre deux nœuds est d'intérêt pour caractériser la distance entre deux échantillons tombés dans ses nœuds. En effet, le plus petit ancêtre commun est le nœud à partir desquels les deux échantillons considérés se séparent. Si deux échantillons tombent dans le même nœud terminal, alors ce nœud est aussi le plus petit ancêtre commun. Ainsi, des échantillons dont le plus petit ancêtre commun est haut dans l'arbre se ressemblent moins que si la séparation a lieu plus tardivement (Figure 6.7).

En prenant en compte ces considérations, nous proposons deux nouvelles mesures pour caractériser la similarité entre deux échantillons dans un arbre. Afin d'expliciter ces mesures, les notations suivantes sont introduites :

— $lca(n_k(p), n_k(q))$ le plus petit ancêtre commun pour les échantillons p et q tombés dans les noœuds terminaux $n_k(p)$ et $n_k(q)$ du k-ième arbre,

— $root_k$ le nœud racine du k-ième arbre,

— $g(n_k, m_k)$ le nombre de branches qui séparent les nœuds n_k et m_k dans le k-ième arbre.

La première mesure est basée sur la distance au plus petit ancêtre commun $sim_k^{\text{DistanceLCA}}$. Elle est calculée de la manière suivante :

$$sim_k^{\text{DistanceLCA}}(p,q) = \begin{cases} \dfrac{g(root_k, lca\,(n_k(p), n_k(q))}{\max\,(g(root_k, n_k(p)), g(root_k, n_k(q)))} & \text{si } c_p(p) = c_p(q), \\ 0 & \text{sinon.} \end{cases} \tag{6.15}$$

La seconde mesure est basée sur la pureté du plus petit ancêtre commun $sim_k^{\text{PuretyLCA}}$. Comme expliqué dans la Section 2.4.2, un nœud est pur si les échantillons d'apprentissage qui sont tombés dans ce nœud appartiennent à la même classe. Si la mesure de pureté pour le plus petit ancêtre commun est très élevée, alors les échantillons qui sont dans ce nœud appartiennent pour la majorité à la même classe. La mesure de pureté du nœud n_k, notée $Gini(n_k)$, est calculée avec l'équation (2.13) en considérant les échantillons d'apprentissage présents dans le nœud n_k. La valeur $sim_k^{\text{PuretyLCA}}$ s'exprime alors de la manière suivante :

$$sim_k^{\text{PuretyLCA}}(p,q) = \begin{cases} 1 - Gini\,(lca(n_k(p), n_k(q))) & \text{si } c_p(p) = c_p(q), \\ 0 & \text{sinon.} \end{cases} \tag{6.16}$$

En considérant les différentes mesures de similarité, la proximité et le score d'*outlier* O_{RF} sont calculés à l'aide des équations (6.12) et (6.14) respectivement. Afin de visualiser les différences entre les trois mesures de similarité, des exemples de valeurs de proximité pour un ensemble d'échantillons d'apprentissage sont visibles dans l'Annexe B.2.

6.3 Présentation des expérimentations

Ces travaux s'intéressent aux performances de différentes méthodes de détection des données mal étiquetées. L'objectif est de mettre en évidence les bonnes performances de la stratégie proposée dans cette thèse, qui est basée sur l'utilisation des scores d'*outlier* du RF. Plus spécifiquement, la précision de détection de ces méthodes est évaluée quantitativement et comparée avec des méthodes de l'état-de-l'art.

Dans un premier temps, les données satellitaires et données de référence utilisées dans les expérimentations sont présentées. Dans un deuxième temps, les méthodes de détection de données mal étiquetées utilisées dans les évaluations sont listées. Finalement, les mesures d'évaluation utilisées sont décrites.

6.3.1 Données satellitaires et données de référence

Les différentes expérimentations menées dans ce chapitre sont basées sur trois jeux de données : 1) les données simulées, 2) les données SPOT-Landsat, et 3) les données Sentinel-2.

Données simulées et SPOT-Landsat

Les données simulées et SPOT-Landsat correspondent aux jeux de données, composés de cinq classes, utilisés dans le Chapitre 5. Pour ces deux jeux de données, le vecteur de variables utilisé correspond aux profils de NDVI. Pour rappel, la dimension du vecteur de variables est de 15 et 23 pour les données simulées et SPOT-Landsat respectivement. Le nombre d'échantillons est de 500 par classe (10 échantillons prélevés dans 50 polygones).

Les deux jeux de données sont bruités en ajoutant des données mal étiquetées. La procédure de génération de bruit a été décrite à la Section 5.2.2. Comme au Chapitre 5, le bruit aléatoire est ajouté pour des niveaux de 5 à 95 % par pas de 5 %.

Données Sentinel-2

Concernant les données Sentinel-2, la série temporelle utilisée est celle décrite dans la Section 3.1.2. Après avoir appliqué la procédure pour harmoniser les dates des images sur les six tuiles (Section 3.2.3), la série temporelle est composée de trente dates. Pour chaque date, les images sont composées de dix bandes spectrales (celles à 10 mètres de résolution spatiale, et celles à 20 mètres ré-échantillonnées à 10 mètres, Tableau 3.3).

Dans le cas de ces données, deux données de référence sont disponibles : 1) le RPG 2014, et 2) des données terrain acquises en 2016. Comme les images satellitaires sont acquises en 2016, l'utilisation du RPG 2014 comme donnée de référence conduit à la présence de données mal étiquetées. Cette situation est proche de la réalité puisque le RPG 2014 était la donnée de référence décrivant la végétation la plus récente disponible pour l'ensemble de la France fin 2016. Dans ces expérimentations, les données terrain 2016 sont utilisées comme vérité terrain pour déterminer le niveau de bruit présent dans les données RPG 2014. L'utilisation de procédures rigoureuses de collecte des occupations des sols permet de considérer qu'aucune donnée mal étiquetée n'est présente dans ces données. Pour ces données, six classes de végétation sont gardées : céréales à paille (blé et orge), maïs, colza, tournesol, vignes et prairies (temporaires et permanentes).

Le détail sur le nombre d'échantillons utilisés pour les évaluations sera présenté dans la Section 6.4.3.

6.3.2 Méthodes étudiées

Plusieurs méthodes de détection de données mal étiquetées de la littérature sont étudiées. En suivant les récents travaux sur la détection de données mal étiquetées [Garcia et al., 2015; Sáez et al., 2016; Smith and Martinez, 2015], les méthodes suivantes sont analysées : kNN, kNNW, LOF, ENN, RENN, et All$k$NN. Par ailleurs, les performances des méthodes iForest et SOS sont aussi évaluées. Ces deux méthodes sont sélectionnées pour leur principe de fonctionnement différent des six autres méthodes de la littérature. En effet, la méthode iForest est basée sur l'utilisation d'un ensemble d'arbres pour caractériser la proximité entre échantillons et la méthode SOS permet d'obtenir un score d'*outlier* borné.

L'objectif des évaluations est de comparer toutes ces méthodes avec celles décrites dans la Section 6.2. Ces dernières basées sur la structure des arbres du RF n'ont jamais été utilisées pour la problématique de détection de données mal étiquetées. Par ailleurs, les évaluations permettront aussi de déterminer l'intérêt des deux mesures de similarité – DistanceLCA et PuretyLCA – proposées dans ces travaux.

Les méthodes kNN, kNNW, LOF, et SOS sont des méthodes de détection d'*outliers* proposées pour des échantillons appartenant à une seule classe. Elles doivent donc être

adaptées à la problématique de détection de données mal étiquetées pour laquelle les échantillons appartiennent à différentes classes. Pour ces méthodes, le score d'*outlier* d'un échantillon p est évalué en utilisant pour les calculs d'*outlier* uniquement les échantillons qui appartiennent à la même classe que p.

Les implémentations Python disponibles dans la bibliothèque Scikit-Learn sont utilisées pour les méthodes SOS et iForest. Concernant les six autres méthodes basées sur la distance (kNN, kNNW, LOF, ENN, RENN, et All$k$NN) et les trois méthodes basées sur la structure des arbres du RF, les implémentations ont été réalisées en Python et C++ respectivement.

Dans la littérature, ces méthodes sont développées pour traiter des échantillons indépendants. Cependant, cette hypothèse est souvent transgressée dans le contexte de la télédétection à cause de la nature des données de référence. Comme indiqué à la Section 5.2.2, les données de référence sont généralement composées de polygones qui représentent des zones homogènes ayant la même occupation des sols. Les échantillons extraits d'un même polygone sont alors très similaires. Ainsi, les k plus proches voisins de chaque échantillon sont des échantillons appartenant au même polygone. Dans notre problématique, les données mal étiquetées correspondent principalement à des polygones mal étiquetés. Si cette considération n'est pas prise en compte, la distance entre un échantillon donné et ses plus proches voisins est toujours faible. Les scores d'*outlier* calculés pour les méthodes basées sur la distance au k plus proches voisins sont alors biaisés. Ce problème est d'autant plus important lorsque les échantillons sont extraits de polygones de taille différentes. En effet, les échantillons qui appartiennent à de gros polygones auront plus de voisins similaires que ceux appartenant à de petits polygones. Les scores d'*outlier* seront alors plus élevés pour les échantillons appartenant à de petits polygones.

Afin de mieux prendre en compte la corrélation qui existe entre les échantillons, une modification du calcul du score d'*outlier* est proposée. L'idée est de calculer le score d'*outlier* d'un échantillon p sans utiliser les échantillons qui appartiennent au même polygone que p. L'impact positif de ce raffinement a été analysé, mais n'est pas présenté dans ce manuscrit. À noter que les implémentations des méthodes SOS et iForest n'ont pas été modifiées pour prendre en compte ce raffinement.

Par ailleurs, une majorité des méthodes de la littérature utilise la distance euclidienne pour calculer la proximité entre chaque paire d'échantillons. Or, la distance euclidienne est très sensible à la dispersion des différentes variables utilisées. C'est pourquoi, l'ensemble des données sont standardisées en soustrayant par la moyenne et en divisant par l'écart-type.

6.3.3 Évaluation

Les méthodes étudiées dans ces travaux peuvent être divisées en deux catégories en fonction du type de score d'*outlier* calculé. La première catégorie comprend les méthodes d'édition qui calculent un score binaire : 1 si l'échantillon est identifié comme mal étiqueté, 0 sinon. La seconde catégorie comprend toutes les autres méthodes qui calculent un score d'*outlier* non-binaire. Pour les méthodes calculant un score non-binaire, il est donc nécessaire de définir un seuil pour identifier les données mal étiquetées.

La définition du seuil n'est pas évidente. Intuitivement, si les méthodes sont précises, les échantillons ayant un fort score d'*outlier* doivent correspondre à des échantillons mal étiquetés. Au contraire, les échantillons ayant de faibles score d'*outlier* doivent correspondre à des échantillons correctement étiquetés. Ainsi, une stratégie courante consiste à ordonner les échantillons en fonction de leur score d'*outlier* et de considérer les n échan-

tillons avec les plus forts scores d'*outliers* comme mal étiquetés.

Une fois les données mal étiquetées identifiées, l'évaluation des méthodes de détection de données mal étiquetées peut se faire comme pour un problème de classification binaire. Dans le cas de cette étude, les échantillons mal étiquetés sont parfaitement connus pour les différents jeux de données. Pour les données simulées et SPOT-Landsat, un bruit généré est utilisé. Pour les données Sentinel-2, les échantillons mal étiquetés sont identifiés en utilisant une vérité terrain. Ainsi, la connaissance des données mal étiquetées permet de calculer une matrice de confusion en considérant deux classes : mal étiquetée *versus* correctement étiquetée. Le Tableau 6.1 montre cette matrice de confusion où N représente le nombre total d'échantillons.

Tableau 6.1 – Exemple de matrice de confusion pour un problème de détection de données mal étiquetées.

Réel \ Prédit	Mal étiqueté	Correctement étiqueté	Total
Mal étiqueté	vrais positifs (VP)	faux négatifs (FN)	VP+FN
Correctement étiqueté	faux positifs (FP)	vrai négatifs (VN)	FP+VN
Total	VP+FP	FN+VN	N

De cette matrice de confusion, il est possible de définir plusieurs métriques, notamment :

— taux de vrais positifs ou sensibilité ou rappel : $\mathrm{PA} = \frac{\mathrm{VP}}{\mathrm{VP+FN}}$

— taux de vrais négatifs ou spécificité : $\mathrm{TVN} = \frac{\mathrm{VN}}{\mathrm{FP+VN}}$

— valeur prédictive positive ou précision : $\mathrm{UA} = \frac{\mathrm{VP}}{\mathrm{VP+FP}}$

— valeur prédictive négative : $\mathrm{VPN} = \frac{\mathrm{VN}}{\mathrm{FN+VN}}$

Ces mesures permettent notamment de calculer le F-Score. Comme pour un problème de classification, le F-Score correspond à la moyenne harmonique de la précision (UA) et du rappel (PA) :

$$\text{F-Score} = 2\frac{\mathrm{PA} \times \mathrm{UA}}{\mathrm{PA} + \mathrm{UA}}. \tag{6.17}$$

La valeur du F-Score peut être évaluée pour les mesures de détection calculant des scores d'*outlier* binaires et non-binaires. Dans le cas d'un score non-binaire, le F-Score est évalué pour une valeur de n fixée.

En suivant les recommandations de la récente *review* de Campos et al. [2015], $P@n$ la précision à n et la courbe *Receiver Operating Characteristic* (ROC) sont aussi utilisées pour l'évaluation des différentes méthodes. La métrique $P@n$ et la courbe ROC sont décrites dans la suite.

La précision à n correspond au pourcentage de données mal étiquetées parmi les n échantillons ayant les plus fortes mesures d'*outliers*. Cette métrique peut donc être évaluée uniquement pour les méthodes de détection qui calculent un score d'*outlier* non-binaire.

La courbe ROC est couramment utilisée pour étudier la sensibilité d'une méthode à ses paramètres. Dans notre contexte, elle permet d'étudier l'influence du seuil n pour les méthodes calculant un score d'*outlier* non-binaire. Plus spécifiquement, la courbe ROC évalue la sensibilité (PA) en fonction du taux de faux positifs (1-UA) pour différentes valeurs de n [Fawcett, 2006]. De plus, l'évaluation de l'aire sous la courbe (*Area Under the Curve* (AUC) en anglais) permet de résumer la courbe ROC en une valeur. Elle est généralement approximée avec une méthode d'intégration par trapèze. La Figure 6.8

montre différents exemples de courbes ROC avec les valeurs de ROC-AUC associées (dans la légende).

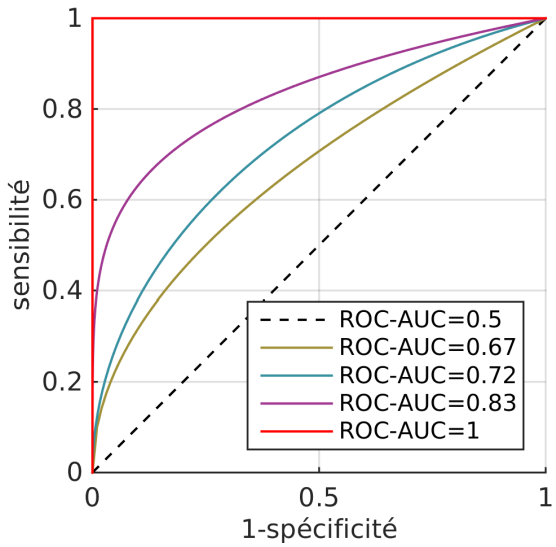


FIGURE 6.8 – Exemple de courbes *Receiver Operating Characteristic* (ROC) et de valeur d'*Area Under the Curve* (AUC). En rouge, la courbe idéale. En pointillé noir, le cas aléatoire.

La courbe rouge représente le cas idéal pour lequel la méthode identifie toutes et uniquement les données mal étiquetées. La courbe en pointillé noir correspond alors au cas aléatoire pour lequel la méthode identifierait une fois sur deux les données mal étiquetées.

Malgré une interprétation facile, la courbe ROC présente deux défauts. Le premier est qu'elle ne prend pas en compte les valeurs de scores d'*outlier* des échantillons. Pourtant, cette information permet de caractériser le degré d'isolement des échantillons. Afin de prendre en compte la valeur des scores d'*outlier*, des mesures de corrélation, Spearman et Kendall, peuvent être utilisées [Schubert et al., 2012]. Le second défaut de la courbe ROC est qu'elle n'est pas adaptée dans des problèmes peu équilibrés où peu d'*outliers* sont présents. Dans ce cas spécifique, l'évaluation peut être réalisée avec la courbe précision-rappel moins sensible aux problèmes déséquilibrés [Goldstein and Uchida, 2016].

6.4 Résultats des expérimentations

Dans cette partie, les performances des méthodes de détection de données mal étiquetées sélectionnées à la Section 6.3.2 sont évaluées. Comme mentionné à la Section 6.3.1, trois jeux de données sont utilisés : les données simulées et SPOT-Landsat pour lesquelles différents niveaux de bruit sont générés et les données Sentinel-2 pour lesquelles la présence de données mal étiquetées est due à une situation réelle.

Parmi les méthodes étudiées, une majorité nécessite la configuration du paramètre k ou du seuil n. Ainsi, une première étude analyse la sensibilité des méthodes basées sur la distance vis-à-vis de ces paramètres. Ensuite, une deuxième étude vise à évaluer et comparer toutes les méthodes pour différents niveaux de bruit. Enfin, une dernière étude est consacrée à l'évaluation des méthodes sur les données Sentinel-2.

6.4.1 Paramétrage des méthodes basées sur la distance

L'objectif de cette première étude est d'étudier les méthodes basées sur la distance. Les méthodes étudiées sont les suivantes : kNN, kNNW, LOF, ENN, RENN et All$k$NN. La sensibilité de ces méthodes à leurs paramètres est analysée pour différents niveaux de bruit. Plus spécifiquement, les données simulées et SPOT-Landsat décrites dans la Section 6.3.1 sont utilisées. Les niveaux de bruit étudiés sont ici restreints à 10, 20, 30 et 40 %.

Pour toutes les méthodes, la configuration du paramètre de voisinage k est nécessaire. La valeur de k qui permet d'obtenir la meilleure précision de détection, dépend de la méthode étudiée, du jeu de données et du niveau de bruit présent dans les données. Dans ces études, les valeurs de k testées vont de 1 à 371 par pas de 10. Par ailleurs, les méthodes kNN, kNNW et LOF nécessitent de définir une valeur seuil n afin d'identifier les données mal étiquetées. Pour ces méthodes, une étude sur la valeur de n sera réalisée.

Dans cette partie, les évaluations sont réalisées pour les deux critères suivants : ROC-AUC et F-Score. Une première étude est dédiée aux méthodes qui calculent un score d'*outlier* non-binaire, *i.e.* aux méthodes kNN, kNNW et LOF. Pour ces méthodes, les valeurs de ROC-AUC obtenues pour différentes valeurs de k sont analysées. Une seconde étude est ensuite consacrée à l'analyse des performances de toutes les méthodes avec le F-Score.

Analyse des courbes *Receiver Operating Characteristic*

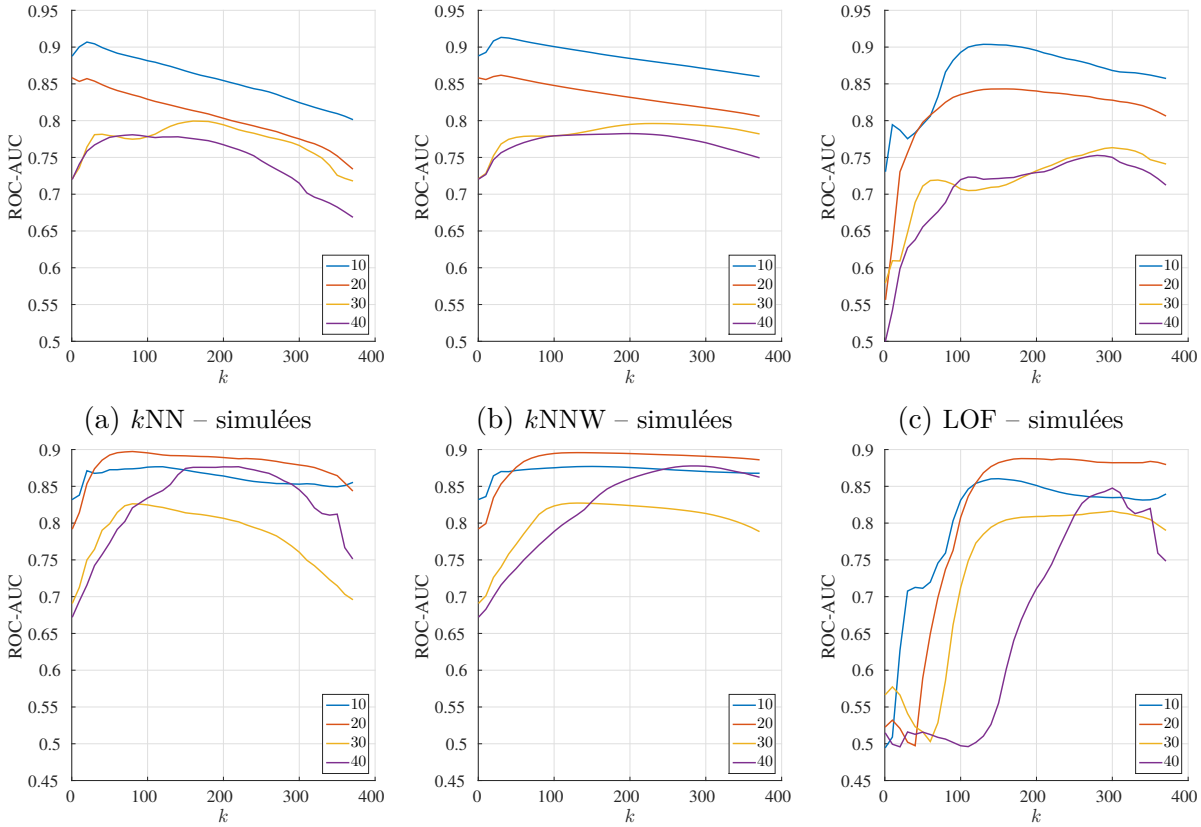
Les performances des méthodes calculant un score non-binaire (kNN, kNNW et LOF) sont ici évaluées. La Figure 6.9 montre l'influence du paramètre k sur la valeur de ROC-AUC. Les colonnes représentent les résultats obtenus pour ces différentes méthodes. La première ligne correspond aux résultats obtenus pour les données simulées à cinq classes, tandis que la seconde ligne correspond aux résultats obtenus pour les données SPOT-Landsat. Chaque courbe représente un niveau de bruit : 10 % en bleu, 20 % en rouge, 30 % en jaune, et 40 % en violet. Afin de comparer les résultats entre les méthodes, les échelles sont identiques pour chaque jeu de données.

Dans la suite, la valeur $\hat{k}$ désigne la valeur optimale de k pour laquelle le maximum de F-Score ou de ROC-AUC est atteint. Pour un même jeu de données et un même niveau de bruit, les méthodes k-NN et k-NNW obtiennent des valeurs de ROC-AUC plus élevées que la méthode LOF. De manière générale, les méthodes kNN et kNNW ont des comportements très similaires. Une comparaison plus précise sera réalisée par la suite.

Pour une même valeur de k, les valeurs de ROC-AUC sont élevées pour des faibles niveaux de bruit. Les données mal étiquetées sont donc plus facilement identifiées lorsque le niveau de bruit est petit.

Par ailleurs, la comparaison des résultats entre les données simulées et les données SPOT-Landsat montre que les valeurs $\hat{k}$ sont fortement dépendantes du jeu de données étudié. En effet, la valeur $\hat{k}$ ne sera pas la même entre les données simulées et SPOT-Landsat pour un même niveau de bruit. Parallèlement, la Figure 6.9 montre que la valeur $\hat{k}$ est influencée par le niveau de bruit pour les trois méthodes étudiées. Si le niveau de bruit augmente, la voisinage d'un échantillon p est contaminé par un plus grand nombre de données mal étiquetées. Ainsi, augmenter la valeur de k permet de prendre en compte un plus grand nombre de voisins, et d'améliorer la précision du calcul du score d'*outlier*.

De plus, il existe une plage pour laquelle les valeurs de ROC-AUC sont stables pour toutes les configurations. Ce plateau présente des longueurs variables en fonction des jeux de données, du niveau de bruit et de la méthode utilisée. Par exemple, il est très long



(a) kNN – simulées (b) kNNW – simulées (c) LOF – simulées

(d) kNN – SPOT-Landsat (e) kNNW – SPOT-Landsat (f) LOF – SPOT-Landsat

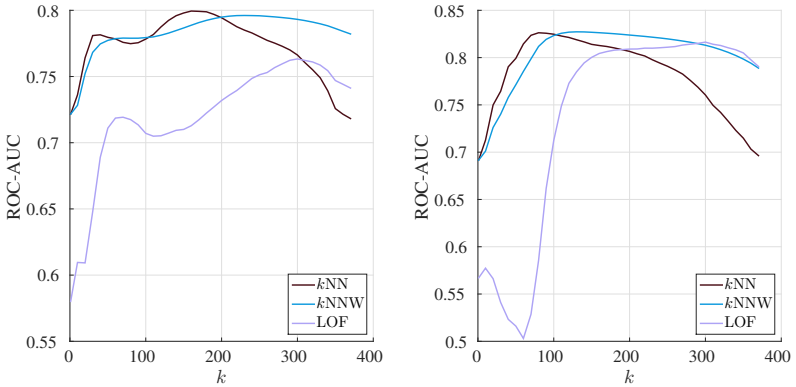
FIGURE 6.9 – Évolution du ROC-AUC en fonction du paramètre k pour les méthodes kNN, kNW et LOF pour les données simulées et SPOT-Landsat à cinq classes pour quatre niveaux de bruit 10, 20, 30 et 40 %.

pour les méthodes kNN et kNNW sur les données SPOT-Landsat. En revanche, il est quasi-inexistant pour les mêmes méthodes sur les données simulées avec des niveaux de bruit faibles de 10 et 20 %.

Pour la méthode LOF basée sur la densité, les résultats de la Figure 6.9 (troisième colonne) montre l'obtention d'optimums locaux pour de petites valeurs de k. Dans ce cas là, des échantillons mal étiquetés sont correctement identifiés localement.

Afin de mieux comparer les résultats de la Figure 6.9, la Figure 6.10 montre les valeurs de ROC-AUC obtenues sur les données simulées et SPOT-Landsat pour un niveau de bruit de 30 %. En fait, la Figure 6.10 montre l'ensemble des courbes jaunes de la Figure 6.9. Plus précisément, la Figure 6.10a montre les résultats pour les données simulées, tandis que la Figure 6.10b montre les résultats pour les données SPOT-Landsat. Chaque courbe représente une méthode : kNN en marron, kNNW en bleu et LOF en violet clair.

Pour un même niveau de bruit, la Figure 6.10 met en évidence les résultats précédents : les méthodes kNN et kNNW obtiennent des valeurs de ROC-AUC plus élevées que la méthode LOF. Bien que non montré-ici, ce résultat a été vérifié pour tous les niveaux de bruit. De plus, ce résultat est en accord avec ceux de Goldstein and Uchida [2016]. Dans ce dernier travail, les auteurs recommandent de privilégier la méthode kNN ou ses variantes lorsque les *outliers* à détecter sont éloignés de tous les autres échantillons (*cross-outlier*, Figure 6.5b). En revanche, ils recommandent d'utiliser la méthode LOF lorsque les *outliers* à détecter sont proches de d'autres échantillons mais avec une densité différente (échantillon p de la Figure 6.4). Dans notre contexte, le premier cas est plus fréquent,



(a) 30 % – simulées (b) 30 % – SPOT-Landsat

FIGURE 6.10 – Évolution du ROC-AUC en fonction du paramètre k pour les méthodes kNN, kNW et LOF pour les données simulées et SPOT-Landsat à cinq classes pour un niveau de bruit de 30 %.

car les échantillons mal étiquetés sont généralement éloignés des autres échantillons de la classe à laquelle ils appartiennent.

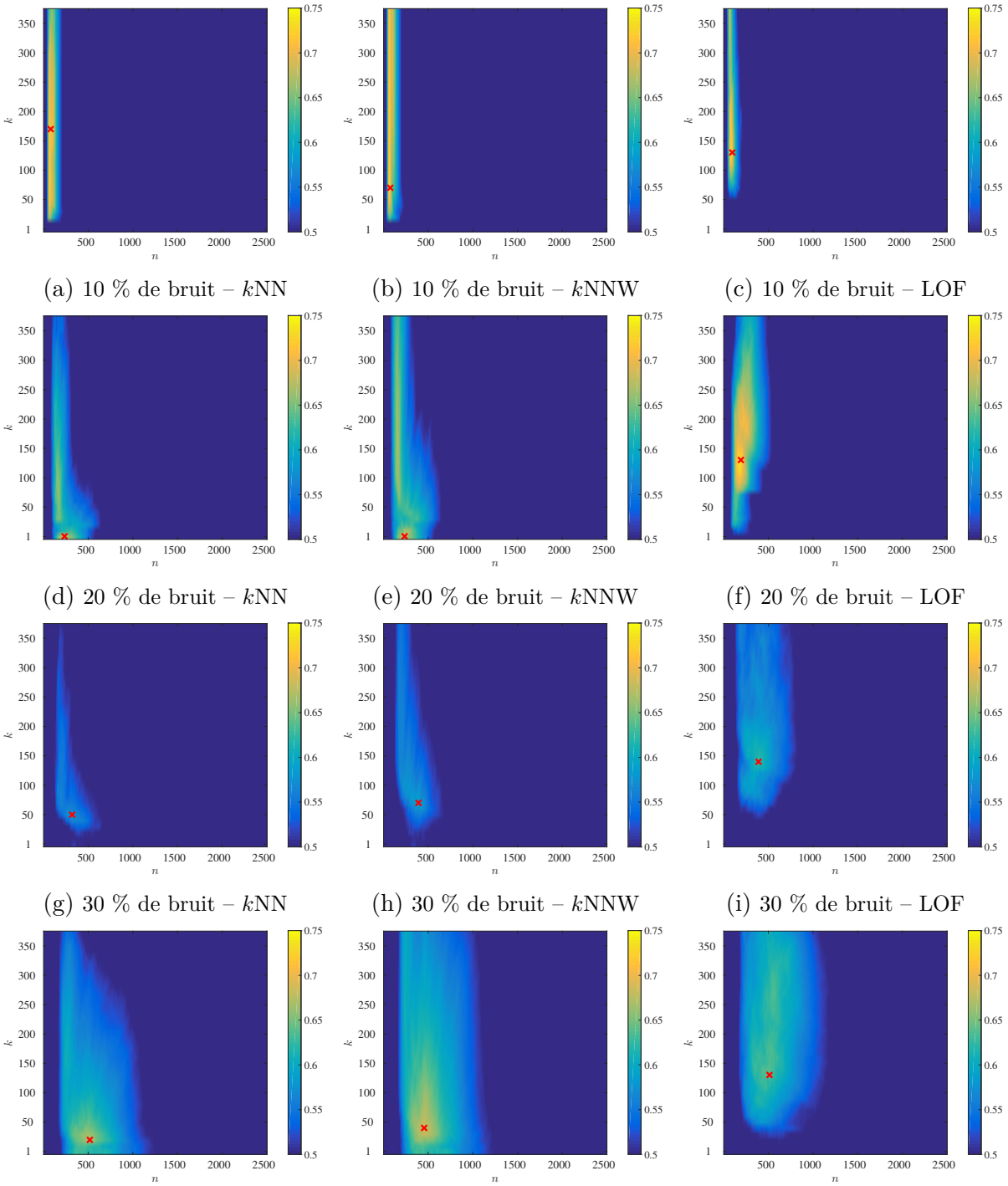
Analyse des valeurs de F-Score

La seconde étude est dédiée à l'analyse des F-Scores pour l'ensemble des méthodes basées sur la distance. Comme précédemment, les données simulées et SPOT-Landsat sont utilisées pour quatre niveaux de bruit (10, 20, 30 et 40 %). Les trois méthodes évaluées dans l'étude précédente sont étudiées et comparées avec les méthodes d'édition ENN, RENN et All$k$NN. Ces dernières calculent un score d'*outlier* binaire qui permet de diviser directement les échantillons en deux sous-ensembles – correctement étiqueté *versus* mal étiqueté. Le F-Score des méthodes d'édition est donc étudié en fonction du paramètre k. Par contre, les méthodes kNN, kNNW et LOF nécessitent de définir la valeur de n, en plus de celle de k. Le F-Score de ces méthodes est alors étudié en fonction de ces deux paramètres. La Figure 6.11 montre les F-Scores obtenus pour les données simulées à cinq classes.

Les colonnes de droite à gauche correspondent aux méthodes kNN, kNNW et LOF. Chaque ligne représente un niveau de bruit allant de 10 à 40 % par pas de 10 %. La croix rouge indique la configuration (k,n) qui permet d'obtenir le meilleur F-Score. Pour améliorer la visibilité, les valeurs de F-Score inférieures à 0,5 ne sont pas détaillées. Elles apparaissent toutes en bleu foncé.

La Figure 6.11 montre que les valeurs optimales de n et k dépendent à la fois de la méthode utilisée et du niveau de bruit. La tendance générale montre que les valeurs de F-Score diminuent lorsque le niveau de bruit augmente. Par ailleurs, la valeur optimale de n augmente avec le niveau de bruit. Ce qui est attendu puisque le nombre de données mal étiquetées à détecter est plus important pour ces niveaux de bruit plus élevé.

Comme montré lors de la première étude, la valeur $\hat{k}$ varie en fonction du niveau de bruit pour les méthodes kNN et kNNW. Pour la méthode LOF, la valeur $\hat{k}$ est plus stable. En effet, l'optimum est toujours obtenu aux alentours de 125 quelque soit le niveau de bruit. Par ailleurs, la Figure 6.11 montre que l'influence du paramètre k est moins importante que celle du paramètre n. Ce résultat confirme ceux obtenus lors de l'étude du ROC-AUC, qui faisait apparaître un plateau pour lequel les valeurs de ROC-AUC étaient stables.



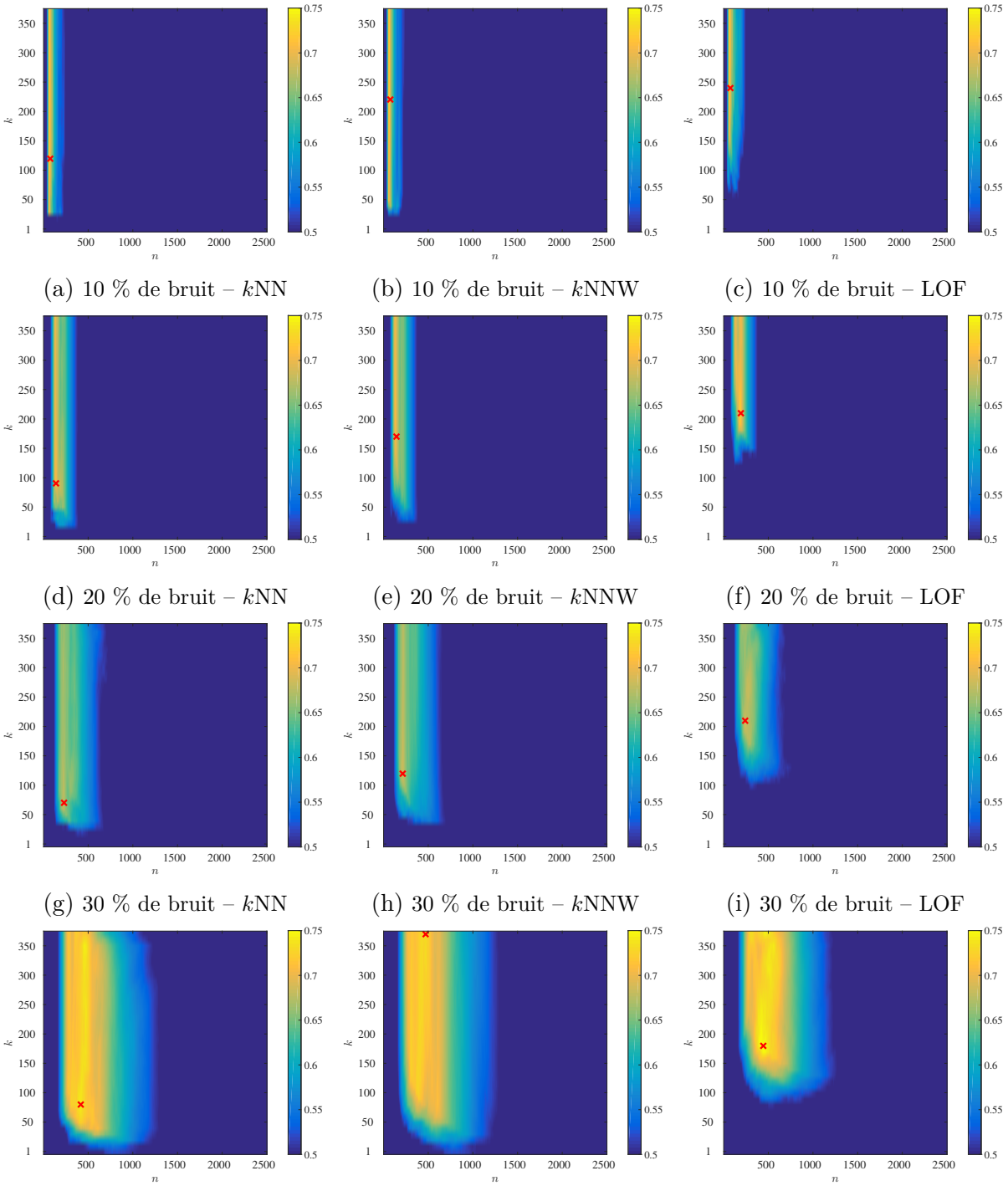
(a) 10 % de bruit – kNN (b) 10 % de bruit – kNNW (c) 10 % de bruit – LOF

(d) 20 % de bruit – kNN (e) 20 % de bruit – kNNW (f) 20 % de bruit – LOF

(g) 30 % de bruit – kNN (h) 30 % de bruit – kNNW (i) 30 % de bruit – LOF

(j) 40 % de bruit – kNN (k) 40 % de bruit – kNNW (l) 40 % de bruit – LOF

FIGURE 6.11 – Évolution du F-Score en fonction des paramètres k et n pour les données simulées à cinq classes en fonction de quatre niveaux de bruit 10, 20, 30 et 40 %. La croix rouge représente le meilleur F-Score.



(a) 10 % de bruit – kNN (b) 10 % de bruit – kNNW (c) 10 % de bruit – LOF

(d) 20 % de bruit – kNN (e) 20 % de bruit – kNNW (f) 20 % de bruit – LOF

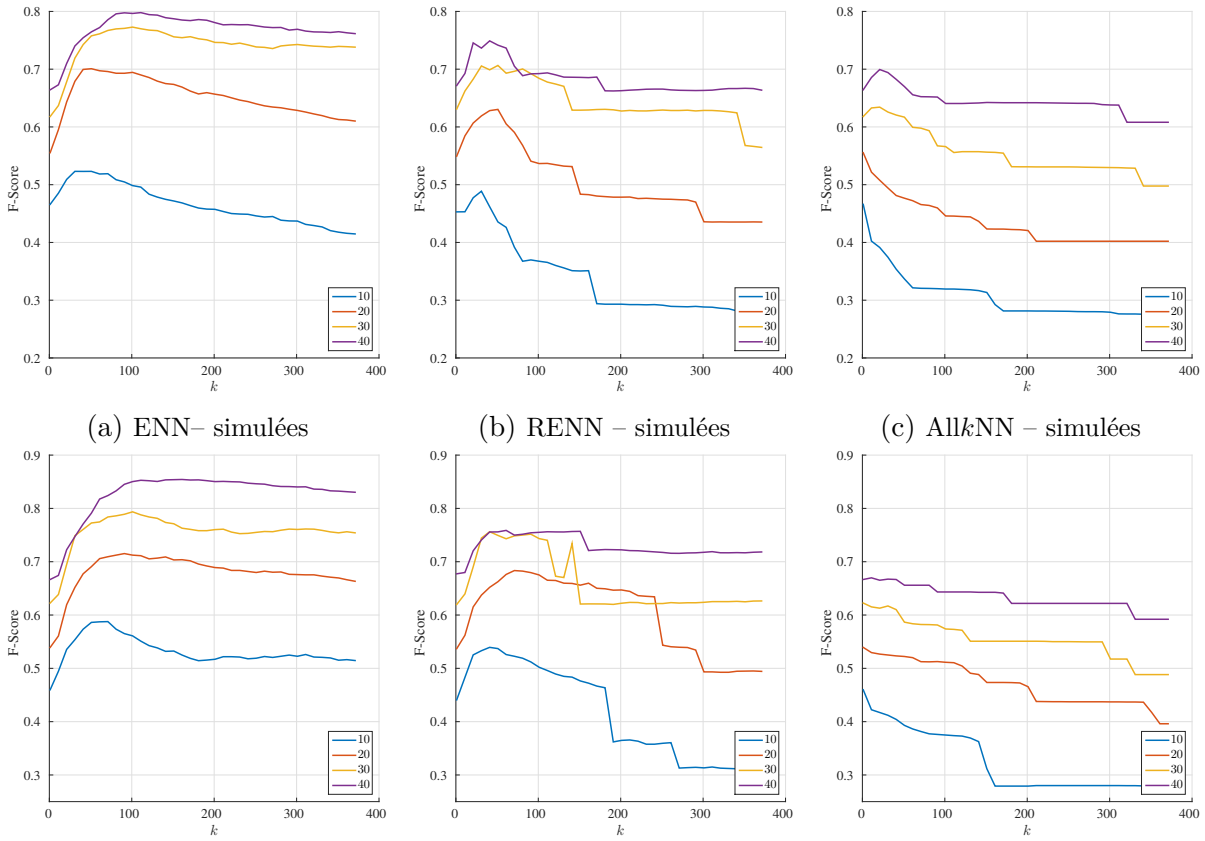
(g) 30 % de bruit – kNN (h) 30 % de bruit – kNNW (i) 30 % de bruit – LOF

(j) 40 % de bruit – kNN (k) 40 % de bruit – kNNW (l) 40 % de bruit – LOF

FIGURE 6.12 – Évolution du F-Score en fonction des paramètres k et n pour les données SPOT-Landsat à cinq classes en fonction de quatre niveaux de bruit 10, 20, 30 et 40 %. La croix rouge représente le meilleur F-Score.

En suivant la même configuration, la Figure 6.12 montre les F-Scores obtenus en fonction des paramètres k et n pour les données SPOT-Landsat. L'analyse de la Figure 6.12 donne des conclusions similaires à celles faites sur les données simulées pour le la Figure 6.11. Néanmoins, les valeurs de F-Score correspondant à l'optimum (k,n) sont plus grandes pour les données SPOT-Landsat.

Afin de comparer les méthodes kNN, kNNW et LOF aux méthodes d'édition, il est nécessaire d'étudier la sensibilité des méthodes d'édition au paramètre k. Pour ces méthodes, la Figure 6.13 montre l'influence du paramètre k sur la valeur du F-Score. Les colonnes représentent de droite à gauche les résultats obtenus pour les méthodes ENN, RENN et All$k$NN. La première ligne correspond aux résultats obtenus pour les données simulées à cinq classes, tandis que la seconde colonne correspond aux résultats obtenus pour les données SPOT-Landsat. Chaque courbe représente un niveau de bruit : 10 % en bleu, 20 % en rouge, 30 % en jaune, et 40 % en violet. Afin de comparer les résultats entre les méthodes, les échelles sont identiques pour chaque jeu de données.



(a) ENN– simulées (b) RENN – simulées (c) All$k$NN – simulées

(d) ENN – SPOT-Landsat (e) RENN – SPOT-Landsat (f) All$k$NN – SPOT-Landsat

FIGURE 6.13 – Évolution du F-Score en fonction du paramètre k pour les méthodes ENN, RENN et All$k$NN pour les données simulées et SPOT-Landsat à cinq classes pour quatre niveaux de bruit 10, 20, 30 et 40 %.

La Figure 6.13 montre que les plus fortes valeurs de F-Score sont obtenues globalement pour des valeurs de k inférieures à 100. Les valeurs $\hat{k}$ sont assez similaires pour les différents niveaux de bruit. De plus, le paramétrage pour ces méthodes parait simple puisqu'il existe un plateau pour lequel les valeurs de F-Score sont stables. Plus spécifiquement, la valeur $\hat{k}$ pour la méthode ENN est comprise entre 50 et 100. Pour les méthodes RENN et All$k$NN, elle est généralement située aux alentours de 50.

La Figure 6.13 permet aussi de mettre en évidence les différences entre les trois mé-

thodes d'édition. Ainsi, la méthode ENN obtient des meilleures performances suivie de la méthode RENN, et de la méthode All$k$NN. Les plus faibles performances de la méthode All$k$NN s'expliquent par l'utilisation itérative de la méthode ENN pour des valeurs de k allant de 1 jusqu'à la valeur définie par l'utilisateur. Comme les résultats de la méthode ENN ne sont pas bons pour la première itération ($k = 1$), la méthode All$k$NN obtient de mauvais résultats.

Comme pour les méthodes kNN, kNNW et LOF, les valeurs de F-Score obtenues pour les méthodes d'édition sur la Figure 6.13 sont plus faibles pour les données simulées que pour les données SPOT-Landsat. Une comparaison plus approfondie entre les six méthodes sera réalisée à la Section 6.4.2.

Tous les résultats présentés montrent que le paramètre k est difficile à configurer. Il dépend principalement des données analysées et du niveau de bruit présent dans ces données. Cette étude montre aussi que les méthodes d'édition ENN, RENN et All$k$NN sont moins sensibles au paramétrage de k que les méthodes kNN, kNNW et LOF. Dans la littérature, une solution est souvent utilisée pour diminuer l'influence du paramètre k dans les méthodes kNN, kNNW et LOF. Elle consiste à moyenner les scores d'*outlier* pour plusieurs valeurs de k [Breunig et al., 2000; Goldstein and Uchida, 2016]. Cette solution n'est pas explorée dans ce manuscrit.

6.4.2 Comparaison des performances

L'objectif de cette partie est de comparer les performances de différentes méthodes de détection de données mal étiquetées. Plus spécifiquement, le score d'*outlier* O_{RF}, qui n'a jamais été étudié dans le contexte de la détection de données mal étiquetées, est évalué. Les performances de ces méthodes sont comparées avec les méthodes basées sur la distance étudiées dans la Section 6.4.1, et les méthodes iForest et SOS présentées à la Section 6.3.2. Afin de comparer les performances des différentes méthodes, les trois critères d'évaluation présentés à la Section 6.3.3 sont utilisés : F-Score, $P@n$ et ROC-AUC. Pour rappel, la comparaison avec les méthodes d'édition qui calculent des scores d'*outlier* binaires peut être réalisée uniquement avec le F-Score. Les évaluations sont réalisées sur les données simulées et SPOT-Landsat pour lesquelles le niveau de bruit varie de 5 à 95 % par pas de 5 %.

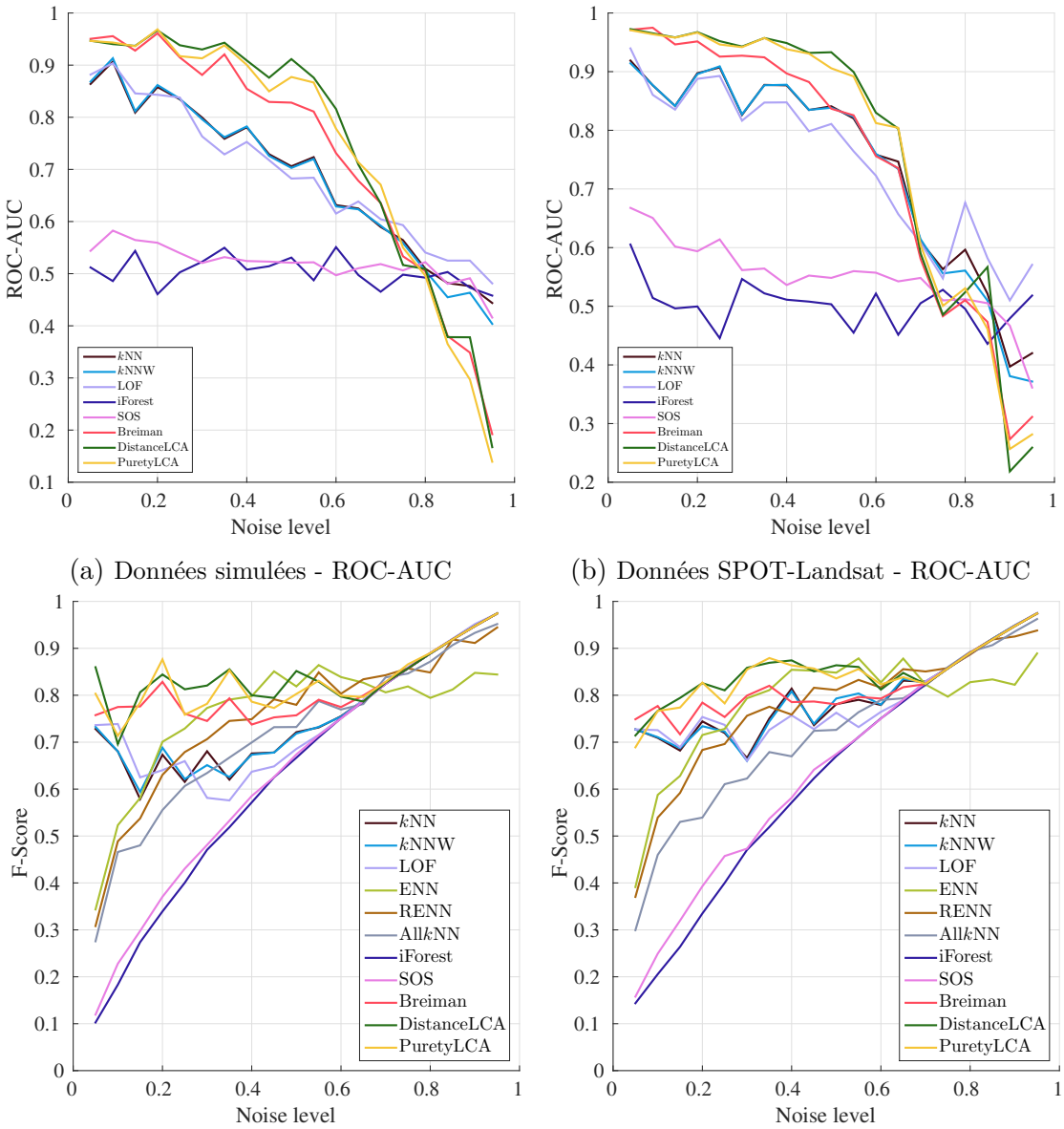
Comme expliqué dans la Section 6.3.2, les implémentations de la librairie *Scikit-Learn* sont utilisées pour les méthodes SOS et iForest. Pour ces deux approches, les paramètres par défaut sont gardés. Ainsi, la perplexité h est égale à 30 pour la méthode SOS, et 100 arbres sont utilisés dans l'algorithme iForest.

Pour les neuf autres méthodes, les meilleurs résultats sont présentés en se plaçant dans la configuration la plus favorable. L'optimisation des paramètres est réalisée pour chaque niveau de bruit. Pour les méthodes d'édition, la valeur $\hat{k}$ est celle permettant d'obtenir la meilleure valeur de F-Score. Pour les méthodes kNN, kNNW et LOF, le choix des valeurs des paramètres dépend du critère d'évaluation. Pour le calcul des valeurs de F-Score, la meilleure configuration (k,n) est choisie (*i.e.* les croix rouges sur les Figures 6.11 et 6.12). Pour le calcul des valeurs de ROC-AUC et de $P@n$, la valeur $\hat{k}$ permettant d'obtenir la valeur de ROC-AUC la plus élevé est choisie.

Pour les méthodes basées sur le score d'*outlier* du RF, l'évaluation du F-Score nécessite la configuration du seuil n. La valeur de n permettant d'obtenir le meilleur F-Score est sélectionnée. Par ailleurs, le calcul des scores d'*outlier* du RF nécessite l'apprentissage d'un modèle de RF. Pour toutes les expérimentations, la configuration suivante est utilisée : le nombre d'arbres K est de 100, le nombre de variables aléatoires sélectionnées à chaque

nœud m est égal à $\sqrt{p}$ avec p la dimension du vecteur de variables, la profondeur maximale *max_depth* est de 25 et le nombre minimal d'échantillons par nœud *min_samples* est de 10.

Une première étude évalue les valeurs de ROC-AUC et de F-Score. La Figure 6.14 montre ces résultats en fonction du niveau de bruit. La première ligne montre les valeurs de ROC-AUC obtenues, la seconde celles du F-Score. La première colonne montre les résultats obtenus pour les données simulées, tandis que la seconde colonne montre les résultats pour les données SPOT-Landsat. Chaque courbe représente une méthode différente.



(a) Données simulées - ROC-AUC

(b) Données SPOT-Landsat - ROC-AUC

(c) Données simulées - F-Score

(d) Données SPOT-Landsat - F-Score

FIGURE 6.14 – Évaluation de la précision des méthodes de détection de données mal étiquetées pour les données simulées et SPOT-Landsat à cinq classes.

Concernant les valeurs de F-Score (montrées dans seconde ligne), il peut paraître surprenant qu'elles augmentent avec le niveau de bruit. Néanmoins, ce résultat est attendu puisque pour des faibles niveaux de bruit obtenir une précision forte, *i.e.* détecter peu de faux positifs, est très difficile. Or, une faible précision fait diminuer le F-Score même si le rappel est élevé.

La Figure 6.14 montre que les méthodes SOS et iForest obtiennent des résultats bien en-deçà. Les valeurs de ROC-AUC sont très proches de 0,5. Elles font à peine mieux qu'un algorithme de détection qui déterminerait les scores d'*outlier* au hasard. Pour ces méthodes, la corrélation entre les échantillons appartenant à un même polygone n'est pas prise en compte dans le calcul des scores d'*outlier*. Les échantillons ont donc toujours des voisins très similaires. Comme expliqué à la Section 6.3.2, les scores d'*outlier* pour ces méthodes sont alors faibles pour tous les échantillons.

Par ailleurs, la méthode iForest fait l'hypothèse que les échantillons mal étiquetés sont ceux qui ont les chemins les plus courts dans les arbres. Cependant, des échantillons correctement étiquetés peuvent aussi avoir des chemins courts dans les arbres s'ils sont différents des autres échantillons, et donc faciles à classer.

Concernant les autres méthodes, la Figure 6.14 montre que les méthodes basées sur le score d'*outlier* O_{RF} sont plus précises que les autres méthodes. De plus, les méthodes kNN, kNNW et LOF ont des performances très similaires. De même, les méthodes ENN et RENN obtiennent des résultats assez identiques. Ces résultats confirment les similarités observées entre ces méthodes dans la Section 6.4.1. La méthode All$k$NN obtient de moins bonnes performances que les autres méthodes en particulier pour les données SPOT-Landsat. Ces résultats sont en accord avec ceux obtenus à la Section 6.4.1.

Les faibles performances des méthodes basées sur les scores $O_{k\text{NN}}$, $O_{k\text{NNW}}$ et O_{LOF} étaient attendues puisque ces méthodes utilisent la distance euclidienne pour calculer la similarité entre les échantillons. Or, la distance euclidienne n'est pas adaptée à la complexité des vecteurs de variables extraits des séries temporelles d'images satellitaires (Section 6.1.4). Il est possible d'utiliser d'autres types de distance pour calculer la proximité entre échantillons. Cependant, peu de travaux en télédétection ont été menés dans ce sens. À noter tout de même que la distance *Dynamic Time Warping* (DTW), adaptée pour l'étude des séries temporelles, a été utilisée avec succès pour la classification de données satellitaires à partir d'un k-PPV [Petitjean, 2012]. Pour améliorer les performances de ces méthodes, cette distance pourrait ainsi remplacer la distance euclidienne.

En comparant les trois méthodes basées sur le calcul du score O_{RF}, les résultats sont similaires pour les trois mesures de proximité (Breiman, DistanceLCA et PuretyLCA). Cependant, pour un niveau de bruit supérieur à 20 %, les mesures proposées (DistanceLCA et PuretyLCA) sont plus précises que la mesure initiale de Breiman.

Afin de mieux comprendre les résultats obtenus pour le F-Score, les Tableaux 6.2 et 6.3 montrent les valeurs de rappel (PA) et précision (UA) pour les données simulées et SPOT-Landsat respectivement. Les résultats sont montrés pour les quatre niveaux de bruit suivants : 10, 20, 30 et 40 %. Les valeurs en gras montrent les meilleurs résultats obtenus pour chaque niveau de bruit pour chaque critère d'évaluation. La configuration des paramètres utilisée est identique à celle des Figures 6.14c et 6.14d.

Outre les différences déjà observées avec la Figure 6.14, les Tableaux 6.2 et 6.3 permettent de mettre en évidence plusieurs comportements. Ainsi, les méthodes iForest et SOS obtiennent de fortes valeurs de rappel mais de très faibles valeurs de précision. Elles identifient donc trop d'échantillons comme étant mal étiquetés. Les méthodes d'édition sont aussi sujettes au problème de sur-détection surtout pour les petits niveaux de bruit. Au contraire, les méthodes basées sur le score O_{RF} et les méthodes kNN, kNNW et LOF montrent un bon équilibre entre la précision et le rappel.

Afin de compléter les résultats précédents, les mesures de $P@n$ de précision à n sont évaluées. Ces mesures indiquent le nombre de données mal étiquetées correctement identifiées parmi les n échantillons ayant les plus forts scores d'*outlier*. Comme mentionné dans la Section 6.3.3, cette évaluation ne peut pas être mesurée pour les méthodes d'édition

Tableau 6.2 – Valeurs de rappel (PA) et précision (UA) obtenues pour les différentes méthodes de détection de données mal étiquetées sur les données simulées. Les valeurs en gras représentent les meilleurs résultats.

	PA	**UA**	**PA**	**UAP**	**PA**	**UA**	**PA**	**UA**
Bruit (%)	10		20		30		40	
kNN	66,0	70,2	68,6	66,1	60,5	77,7	75,4	61,3
kNNW	66,0	70,2	65,8	72,1	57,6	74,9	80,9	57,7
LOF	65,6	84,5	64,8	63,3	62,9	54,0	70,9	57,8
ENN	80,0	28,1	66,7	35,4	93,6	64,9	89,9	71,9
RENN	80,0	24,4	66,7	29,9	93,6	59,0	89,0	60,7
All$k$NN	83,3	27,6	60,8	27,0	93,6	41,7	91,0	52,3
iForest	84,0	10,3	**99,2**	20,5	**98,8**	30,9	**100**	40,0
SOS	**86,8**	13,1	88,0	23,5	91,2	32,7	98,3	41,7
Breiman	79,2	75,9	81,2	84,6	71,3	78,0	80,8	67,8
DistanceLCA	66,4	73,1	82,4	**86,6**	83,6	**80,6**	82,2	77,9
PuretyLCA	60,8	**86,4**	90,0	85,4	80,5	76,0	79,1	**78,2**

Tableau 6.3 – Valeurs de rappel (PA) et précision (UA) obtenues pour les différentes méthodes de détection de données mal étiquetées sur les données SPOT-Landsat. Les valeurs en gras représentent les meilleurs résultats.

	PA	**UA**	**PA**	**UA**	**PA**	**UA**	**PA**	**UA**
Bruit (%)	10		20		30		40	
kNN	55,6	**97,9**	68,4	81,6	58,9	76,6	69,2	**98,9**
kNNW	56,0	97,2	71,2	75,7	59,6	74,0	76,8	85,3
LOF	64,0	83,8	71,0	80,3	65,6	66,5	73,0	78,7
ENN	28,6	19,4	**100**	70,3	98,9	77,4	95,8	83,0
RENN	28,6	18,5	**100**	67,2	**100**	76,9	95,8	77,7
All$k$NN	28,6	14,0	88,9	36,0	**100,0**	44,8	95,8	53,5
iForest	38,8	14,0	99,6	20,1	98,7	30,9	**100**	40,0
SOS	63,6	15,5	85,2	25,5	84,9	32,8	97,1	41,5
Breiman	76,0	79,5	77,8	79,1	81,7	78,2	93,9	67,5
DistanceLCA	73,6	80,0	78,2	**87,5**	87,3	**84,4**	87,3	87,6
PuretyLCA	**78,8**	74,6	82,2	83,2	89,5	81,8	89,3	83,6

qui calculent un score d'*outlier* binaire.

Les Tableaux 6.4 et 6.5 montrent les précisions à 10, 50 et 100 pour les données simulées et les données SPOT-Landsat respectivement. Pour chaque $P@n$ étudiée, quatre niveaux de bruit sont étudiés : 10, 20, 30 et 40 %.

TABLEAU 6.4 – Précisions à n, $P@n$ pour $n = 10$, $n = 50$ et $n = 100$ pour quatre niveaux de bruit (10, 20, 30 et 40 %) pour les données simulées à cinq classes.

	$P@10$				$P@50$				$P@100$			
Bruit (%)	10	20	30	40	10	20	30	40	10	20	30	40
kNN	100	100	100	100	98	100	100	100	87	96	99	100
kNNW	100	100	100	100	100	100	100	100	91	99	99	100
LOF	100	100	100	100	100	100	100	100	96	99	98	99
iForest	0	0	0	0	0	0	0	0	0	0	2	3
SOS	10	10	10	30	8	12	16	26	8	12	16	22
Breiman	100	100	100	100	100	100	100	100	93	99	100	99
DistanceLCA	100	100	100	100	100	100	100	98	95	100	100	97
PuretyLCA	100	100	100	100	100	100	100	100	95	100	100	99

TABLEAU 6.5 – Précisions à n, $P@n$ pour $n = 10$, $n = 50$ et $n = 100$ pour quatre niveaux de bruit (10, 20, 30 et 40 %) pour les données SPOT-Landsat à cinq classes.

	$P@10$				$P@50$				$P@100$			
Bruit (%)	10	20	30	40	10	20	30	40	10	20	30	40
kNN	100	100	100	100	100	100	100	100	100	100	100	100
kNNW	100	100	100	100	100	100	100	100	100	100	99	100
LOF	100	100	100	100	94	100	100	100	87	99	100	100
iForest	0	0	20	50	0	0	16	18	0	0	13	17
SOS	0	10	30	50	6	18	30	34	10	12	31	38
Breiman	100	100	100	100	100	100	100	100	100	100	99	97
DistanceLCA	100	100	100	100	100	100	98	100	95	99	99	100
PuretyLCA	100	100	100	100	100	100	100	94	100	99	99	97

Les Tableaux 6.4 et 6.5 montrent une nouvelle fois des résultats catastrophiques pour les méthodes iForest et SOS. Ainsi, les plus forts scores d'*outlier* sont attribués à des échantillons correctement étiquetés pour les méthodes iForest et SOS.

Concernant les autres méthodes, des résultats similaires sont obtenus entre les méthodes kNN, kNNW et LOF et les méthodes basées sur le score O_{RF} (Breiman, DistanceLCA et PuretyLCA). Pour ces six méthodes, les plus forts scores d'*outlier* représentent donc bien des données mal étiquetées.

L'ensemble des résultats présenté dans cette partie montre globalement de bonnes performances pour les différentes méthodes étudiées. Néanmoins, les deux méthodes iForest et SOS ne semblent pas adaptées au contexte spécifique des données mal étiquetées. Par ailleurs, les résultats obtenus avec les méthodes de détection basées sur le RF sont très prometteurs.

6.4.3 Performances des méthodes pour les données Sentinel-2

Les résultats présentés dans cette partie correspondent à l'évaluation des méthodes précédentes sur les données Sentinel-2. L'intérêt de cette étude est de confronter les méthodes de détection à un bruit non-généré. En effet, les données mal étiquetées présentes dans le jeu de données Sentinel sont dues à l'utilisation d'une donnée de référence obsolète (2014) par rapport à l'acquisition des données satellitaires (2016). Ainsi, l'objectif des méthodes utilisées est de détecter quelles sont les étiquettes 2014 qui ne correspondent pas à la vérité terrain 2016.

Les études de la Section 6.4.2 montrent que les méthodes iForest et SOS obtiennent de faibles performances. Dans cette partie, ces méthodes ne sont donc pas étudiées. Ainsi, les études sur les données Sentinel-2 sont réalisées sur les méthodes basées sur la distance, la densité ou encore le score d'*outlier* du RF.

Le nombre d'échantillons disponibles par classe est affiché dans le Tableau 6.6. Les lignes indiquent l'occupation des sols en 2014 (classe fournie par la donnée de référence c_r), tandis que les colonnes donnent l'occupation des sols en 2016 (classe vérité terrain c_{vt}). Les valeurs sur la diagonale, en gras, correspondent donc aux nombres d'échantillons correctement étiquetés. Au contraire, les valeurs hors diagonale représentent les échantillons mal étiquetés. Par exemple, 7971 échantillons sont étiquetés comme du tournesol en 2014, mais sont en réalité du maïs en 2016. De plus, la dernière colonne montre le pourcentage de données mal étiquetées contenues dans la donnée 2014.

Tableau 6.6 – Nombre d'échantillons pour les données Sentinel-2 en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt})

2016 (c_{vt}) / **2014** (c_r)	**CP**	**M**	**C**	**T**	**V**	**P**	**Total**	**% ME**
CP	**126 932**	11 141	16 853	19 457	420	4473	179 276	29,2
M	24 877	**101 151**	3 182	3 555	0	1 429	134 194	24,6
C	14 752	2297	**5569**	18 769	0	1487	42 874	87,0
T	20 290	7971	16 442	**73 819**	0	2542	121 064	39,0
V	0	0	0	0	11 536	0	**11 536**	0,0
P	11 121	3046	1353	0	0	**62 433**	77 953	19,9

CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.
% ME : pourcentage de données mal étiquetées

Le Tableau 6.6 montre que le nombre d'échantillons varie fortement entre les classes. La classe vigne est minoritaire, tandis que la classe céréale à pailles est sur-représentée. Par ailleurs, le pourcentage de données mal étiquetées (dernière colonne) varie aussi en fonction des classes. Des classes comme la vigne ne contiennent aucune donnée mal étiquetée. En revanche, la classe colza contient 87 % de données mal étiquetées. En deux années, la quasi-totalité des parcelles de colza en 2014 change d'occupation des sols. Ce fort changement d'occupation des sols s'explique principalement par la rotation des cultures qui est imposée par les politiques pour la régénération des sols.

Dans cette étude, le bruit est limité à 30 % pour l'ensemble des classes. Si le niveau de bruit présent est supérieur à 30 %, il est diminué à 30 % en sélectionnant aléatoirement 70 % de données correctement étiquetées. Cela arrive notamment pour les classes colza et tournesol. Sinon, le niveau de bruit n'est pas modifié.

Afin de se placer dans un cas similaire aux données simulées et SPOT-Landsat, seulement 500 échantillons par classe sont sélectionnés. L'objectif des expérimentations est alors de détecter les données mal étiquetées présentes dans ce sous-ensemble d'échantillons. La répartition des échantillons utilisés pour les évaluations de cette partie est montrée dans

le Tableau 6.7. Les lignes indiquent l'occupation des sols en 2014, tandis que les colonnes donnent l'occupation des sols en 2016. Ainsi, le niveau de bruit moyen pour les données Sentinel-2 est de 22 %.

TABLEAU 6.7 – Nombre d'échantillons utilisés pour les expérimentations sur les données Sentinel-2 en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt})

2016 (c_{vt}) / **2014** (c_r)	**CP**	**M**	**C**	**T**	**V**	**P**	**Total**	**% ME**
CP	**354**	20	45	66	1	14	500	29,2
M	89	**378**	8	16	0	9	500	24,4
C	56	12	**350**	72	0	10	500	30,0
T	59	32	47	**351**	0	11	500	29,8
V	0	0	0	0	**500**	0	500	0,0
P	76	16	8	0	0	**400**	500	20,0

CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.
% ME : pourcentage de données mal étiquetées

En utilisant ces données, la première évaluation consiste à analyser les courbes ROC. La Figure 6.15 montre ces résultats pour les méthodes kNN, kNNW, LOF, Breiman, DistanceLCA, et PuretyLCA. Les valeurs du paramètre k pour les méthodes kNN, kNNW et LOF sont optimisées comme dans la Section 6.4.2. La meilleure courbe ROC est obtenue pour les méthodes basées sur le calcul du score O_{RF}, tandis que la moins bonne est obtenue pour la méthode LOF. Les courbes ROC similaires des méthodes kNN et kNNW sont entre ces deux types d'approches.

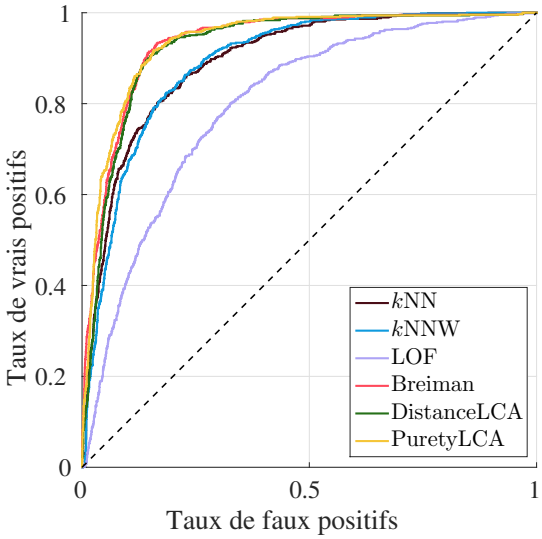


FIGURE 6.15 – Courbe ROC pour les données Sentinel-2

Pour corroborer ces résultats, une deuxième évaluation est réalisée avec le F-Score et les précisions à 10, 50 et 100. Le Tableau 6.8 montre ces résultats pour l'ensemble des méthodes permettant de calculer ses critères. Les valeurs en gras correspondent aux meilleures performances.

Pour les méthodes kNN, kNNW, LOF et les scores O_{RF}, les résultats obtenus pour les valeurs de F-Score sont similaires à ceux obtenus précédemment en analysant les courbes ROC. Par ailleurs, l'analyse des valeurs de F-Score montre aucune différence significative entre les trois mesures de similarité proposées Breiman, DistanceLCA, et PuretyLCA. Les précisions à 10 sont très bonnes pour les méthodes de détection basées sur le score O_{RF}.

Tableau 6.8 – F-Scores et précisions à n (avec n égal à 10, 50 et 100) obtenus pour les différentes méthodes de détection de données mal étiquetées sur les données Sentinel-2. Les valeurs en gras représentent les meilleures performances.

Méthode	**F-Score**	P@10	P@50	P@100
kNN	68,9	10	60	78
kNNW	67,6	10	58	76
LOF	55,4	10	42	52
ENN	**79,7**	-	-	-
RENN	79,0	-	-	-
All$k$NN	67,7	-	-	-
Breiman	75,5	**100**	**96**	**90**
DistanceLCA	75,3	90	86	81
PuretyLCA	75,9	**100**	94	**90**

Lorsque la précision à 10 est de 100 %, les dix échantillons ayant les plus forts scores d'*outlier* correspondent à des données mal étiquetées. Ainsi, les méthodes Breiman et PuretyLCA sont très précises. Les valeurs de précision à 50 et 100 sont aussi élevées pour ces méthodes.

Concernant les méthodes kNN, kNNW et LOF, les précisions à n sont moins bonnes que celles observées dans le cas d'un bruit généré. Comme expliqué à la Section 6.3.2, ces méthodes calculent les scores d'*outlier* pour chaque échantillon p en utilisant uniquement les échantillons qui appartiennent à la même classe que p. Or, la Section 6.4.1 a montré que la valeur $\hat{k}$ pour ces trois méthodes dépend fortement du niveau de bruit dans les données. Pour les données simulées et SPOT-Landsat, le niveau de bruit est identique par classe. La même valeur $\hat{k}$ peut donc convenir pour l'ensemble des classes. Cependant, les données Sentinel-2 contiennent différents niveaux de bruit par classe. Dans cette situation, une valeur de k unique pour toutes les classes n'est donc pas l'idéal.

Par ailleurs, le Tableau 6.8 montre les très bonnes valeurs de F-Score obtenues par les méthodes d'édition. Pour ces données, les méthodes d'édition sont donc plus performantes que les méthodes basées sur le RF. Afin de compléter ces résultats et mieux comprendre ces différences, une troisième évaluation est ici proposée.

Dans l'évaluation précédente, le F-Score est calculé en considérant un problème de classification binaire correctement *versus* mal étiqueté. Cette troisième évaluation consiste à calculer le F-Score, de rappel (PA) et de précision (UA) seulement pour les échantillons qui appartiennent à la même classe. Ces résultats par classe sont montrés dans les Tableaux 6.9, 6.10 et 6.11. Le F-Score pour la classe vigne ne peut pas être calculé puisque cette classe ne contient pas d'échantillons mal étiquetés. Les valeurs en gras représentent les meilleurs résultats. Par ailleurs, la configuration du paramètre k est identique à celle utilisée pour le Tableau 6.8.

Ces trois tableaux montrent que les performances de détection ne sont pas identiques par classe. Pour l'ensemble des méthodes, les valeurs de F-Score de la classe prairies sont généralement plus faibles que celles des autres classes. La classe prairies est connue pour présenter de grande variabilité avec des pratiques agricoles très différentes d'une parcelle à l'autre.

Bien que la méthode ENN ait obtenu une valeur de F-Score plus élevée lors d'une évaluation sur l'ensemble des classes (Tableau 6.8), le Tableau 6.9 montre que les méthodes basées sur les scores O_{RF} obtiennent des valeurs de F-Score par classe plus élevées. La dynamique des scores d'*outlier* O_{RF} est donc probablement différentes entre les classes.

Tableau 6.9 – Valeurs de F-Score pour les méthodes de détection de données mal étiquetées. Les valeurs en gras représentent les meilleurs résultats.

Méthode	**Céréales à paille**	**Maïs**	**Colza**	**Tournesol**	**Prairies**
kNN	71,3	75,7	77,0	74,6	54,6
kNNW	72,2	74,5	73,4	71,7	56,3
LOF	67,2	60,3	62,8	63,8	53,2
ENN	77,1	84,3	85,7	79,2	71,6
RENN	76,5	84,8	85,3	77,8	71,0
All$k$NN	66,7	77,7	65,3	65,0	**72,3**
Breiman	**80,5**	**89,1**	**92,2**	83,7	69,0
DistanceLCA	78,5	86,2	91,5	85,8	66,1
PuretyLCA	78,6	90,1	91,5	**88,2**	66,1

Tableau 6.10 – Valeurs de rappel (PA) pour les méthodes de détection de données mal étiquetées. Les valeurs en gras représentent les meilleurs résultats.

Méthode	**Céréales à paille**	**Maïs**	**Colza**	**Tournesol**	**Prairies**
kNN	76,7	91,0	78,0	83,9	78,0
kNNW	78,1	88,5	74,7	83,2	78,0
LOF	81,5	77,1	82,7	83,9	83,0
ENN	91,1	96,7	94,0	93,3	68,0
RENN	95,9	98,4	96,7	**94,0**	71,0
All$k$NN	**98,6**	**100**	**97,3**	**94,0**	**86,0**
Breiman	86,3	97,5	94,7	91,3	70,0
DistanceLCA	86,3	97,5	**97,3**	91,3	79,0
PuretyLCA	91,8	96,7	**97,3**	90,6	79,0

Tableau 6.11 – Valeurs de précision (UA) pour les méthodes de détection de données mal étiquetées. Les valeurs en gras représentent les meilleurs résultats.

Méthode	**Céréales à paille**	**Maïs**	**Colza**	**Tournesol**	**Prairies**
kNN	66,7	64,9	76,0	67,2	41,9
kNNW	67,1	64,3	72,3	62,9	44,0
LOF	57,2	49,5	50,6	51,4	39,2
ENN	66,8	74,7	78,7	68,8	**75,6**
RENN	63,6	74,5	76,3	66,4	71,0
All$k$NN	50,4	63,5	49,2	49,7	62,3
Breiman	**75,5**	82,1	**89,9**	77,3	68,0
DistanceLCA	72,0	77,3	86,4	80,1	56,8
PuretyLCA	68,7	**84,3**	86,4	**86,0**	56,8

Ainsi, le calcul du F-Score sur l'ensemble des échantillons n'est pas favorables à ces méthodes.

Par ailleurs, les Tableaux 6.10 et 6.11 montrent que la méthode All$k$NN a des valeurs de rappel élevées mais de faibles valeurs de précision. Comme pour les données simulées et SPOT-Landsat, cette méthode privilégie la sur-détection afin d'identifier l'ensemble des données mal étiquetées.

Les résultats présentés ici sur un cas réel montre une nouvelle fois le potentiel des méthodes d'édition et des méthodes basées sur le RF pour la détection des données mal étiquetées.

6.5 Conclusion

La détection de données mal étiquetées à partir de méthodes de détection d'*outlier* a été analysée dans ce chapitre. À notre connaissance, peu de travaux ont été menés sur cette thématique pour des échantillons 1) décrits par des variables extraits de séries temporelles d'images satellitaires, et 2) contenant des données de référence issues d'une situation réelle.

Dans ces travaux, un état-de-l'art sur les méthodes de détection d'*outlier* a été réalisé. Cette étude bibliographique montre que la majorité des méthodes nécessitent la configuration d'au moins un paramètre et la définition d'une distance pour définir la similarité entre les échantillons. Par ailleurs, la majorité de ces méthodes ne sont pas spécifiques au problème de la détection de données mal étiquetées. Ainsi, plusieurs méthodes ne prennent pas en compte la classe fournie la donnée de référence.

Afin de s'affranchir de ces limitations, ces travaux ont proposé l'utilisation de scores d'*outlier* basés sur la structure des arbres du RF. Initialement proposée par Breiman, cette approche n'avait jamais été évaluée dans le contexte de la détection de données mal étiquetées. Le contexte spécifique de la classification de séries temporelles d'images satellitaires a conduit à utiliser cette approche pour deux raisons principales. D'une part, le Chapitre 4 a montré l'intérêt du RF pour la classification de séries temporelles dans des espaces de grande dimension. Ainsi, la structure des arbres construits par le RF permet de bien caractériser les relations entre les échantillons. D'autre part, le Chapitre 5 a montré que les performances du RF étaient peu influencées en présence de peu de données mal étiquetées. Ainsi, les échantillons mal étiquetés utilisés pour la construction des arbres influenceront peu le calcul du score O_{RF}.

Suivant la mesure de similarité proposé par Breiman, deux nouvelles mesures de similarité ont été proposées. À notre connaissance, seulement deux autres études proposent des modifications du score O_{RF} nécessitant la configuration d'au moins un nouveau paramètre [Englund and Verikas, 2012; Nezvalová et al., 2015]. Dans ce chapitre, les méthodes basées sur les scores d'*outlier* du RF sont évaluées et comparées avec les méthodes couramment utilisées dans la littérature. Pour ce faire, trois jeux de données qui contiennent différents niveaux de bruit sont utilisés.

Une première étude a été consacrée aux méthodes basées sur la distance. Plus spécifiquement, la configuration des paramètres de ces méthodes sont étudiées. Les résultats montrent que le paramètre k est difficile à configurer pour les méthodes kNN, kNNW et LOF. En effet, la valeur optimale de k dépend des jeux de données et du niveau de bruit présent dans ces données. Les résultats de cette première étude montrent que les méthodes d'édition sont moins sensibles à ce paramètre.

Une deuxième étude a comparé la précision de l'ensemble des méthodes de détection étudiées pour différents niveaux de bruit. Quelque soit le critère d'évaluation choisi (ROC-

AUC, F-Score ou $P@n$), les méthodes basées sur les scores d'*outlier* et les méthodes d'édition obtiennent les précisions les plus élevées. En revanche, les méthodes iForest et SOS obtiennent des résultats bien en-deçà.

Une dernière étude a évalué les performances des méthodes de détection sur les données Sentinel-2 pour lesquelles les données mal étiquetées représentent une situation réelle. La bonne performance des méthodes de détection montre qu'il est possible de détecter des changements du sol en identifiant les échantillons mal étiquetés. En particulier, cette évaluation a montré les bons résultats des méthodes basées sur le score O_{RF} et des méthodes d'édition. Elle permet aussi de mettre en évidence les problèmes de sur-détection de la méthode All$k$NN.

Dans ces travaux, l'influence du paramétrage du RF sur les scores d'*outlier* n'a pas été évaluée. Par ailleurs, plusieurs processus aléatoires (*bootstrap* et *random feature selection*) sont utilisés pour construire le modèle du RF. Ainsi, deux modèles construits avec les mêmes échantillons peuvent présenter des dissimilarités. Les scores d'*outlier* peuvent donc changer entre les deux modèles. Une solution non étudiée ici pour obtenir des scores O_{RF} plus stables est d'utiliser uniquement les arbres où l'échantillon est *bootstrap* dans les calculs de proximité. Ainsi, cet échantillon n'a pas influencé la structure de l'arbre.

Enfin, l'ensemble des approches évaluent individuellement les scores d'*outlier* des échantillons en utilisant les valeurs des vecteurs de variables. Dans notre problématique, les données mal étiquetées correspondent bien souvent à des polygones mal étiquetés. Il est donc rare qu'un échantillon isolé spatialement soit une donnée mal étiquetée. Ainsi, les méthodes de détection présentées ici pourraient être améliorées en introduisant de l'information spatiale dans le calcul des scores d'*outlier*.

Chapitre 7

Prise en compte des données mal étiquetées

Sommaire

7.1 Filtrage des données mal étiquetées 158
7.1.1 Score d'*outlier* et seuillage 159
7.1.2 Ensemble d'algorithmes de classification et vote 160
7.1.3 Suppression, correction ou pondération 163
7.2 Filtrage itératif des données mal étiquetées 164
7.2.1 Filtrage itératif dans la littérature 165
7.2.2 Filtrage itératif avec le *Random Forest* 166
7.3 Présentation des expérimentations 171
7.3.1 Évaluation 171
7.3.2 Données satellitaires et données de référence 172
7.4 Résultats des expérimentations 175
7.4.1 Étude de filtrages non-itératifs 175
7.4.2 Étude de filtrages itératifs 185
7.4.3 Étude des données Sentinel-2 196
7.5 Conclusion 202

Dans de nombreux problèmes de classification réels, les échantillons d'apprentissage sont entachés de données mal étiquetées. En télédétection, cette présence a diverses sources, comme l'utilisation de données de référence obsolètes par rapport à la réalité du terrain. Or, les algorithmes d'apprentissage supervisé se fient aux étiquettes fournies par la donnée de référence afin de construire leur règle de décision. La présence de données mal étiquetées peut donc conduire à une diminution des performances des algorithmes (Chapitre 5). Cependant, le Chapitre 6 a montré l'efficacité de certaines méthodes de détection d'*outliers* pour identifier les données mal étiquetées, notamment les scores d'*outlier* calculés à partir de la structure des arbres du RF.

Afin de prendre en compte les données mal étiquetées dans le processus de classification, nous proposons un cadre méthodologique. Ce dernier consiste à filtrer les données mal étiquetées avant de réaliser l'apprentissage de l'algorithme de classification. Cette étape de filtrage s'appuie sur les méthodes de détection de données mal étiquetées introduites au Chapitre 6.

Ce chapitre se focalise sur les méthodes de filtrage. Dans un premier temps, différentes stratégies de filtrage existantes dans la littérature sont décrites. Dans un deuxième temps, le principe de filtrage itératif est abordé et de nouvelles stratégies, dans le contexte spécifique de la classification de séries temporelles d'images satellitaires, sont proposées. L'originalité repose sur l'utilisation de l'algorithme du RF pour soit calculer les scores d'*outlier*, soit combiner les prédictions des arbres du modèle. Ensuite, les expérimentations sont présentées, et les résultats détaillés. Enfin, les conclusions de ces expérimentations sont données.

7.1 Filtrage des données mal étiquetées

Parmi les approches proposées dans la littérature, deux stratégies de filtrage sont couramment utilisées pour réduire l'impact des erreurs d'étiquetage dans l'apprentissage [Frénay and Verleysen, 2014].

La première stratégie consiste à utiliser des algorithmes robustes à la présence de données mal étiquetées comme ORBoost et CN2 (Section 5.1.2). Ces méthodes peuvent être une solution lorsque le nombre de données mal étiquetées est faible. Cependant, elles sont peu efficaces en présence d'un grand nombre de données mal étiquetées [Frénay and Verleysen, 2014]. La seconde stratégie plus utilisée consiste alors à filtrer les données mal étiquetées avant l'étape d'apprentissage du classifieur. Si le filtrage est réussi, alors l'algorithme de classification est entraîné avec des échantillons d'apprentissage « propres ». Cette seconde stratégie est étudiée dans ces travaux.

La Figure 7.1 illustre à nouveau le principe général de la classification supervisée en indiquant en rouge le positionnement du filtrage dans la chaîne de traitement. Cette étape de filtrage s'appuie donc sur les données de référence et le vecteur de variables décrivant les échantillons.

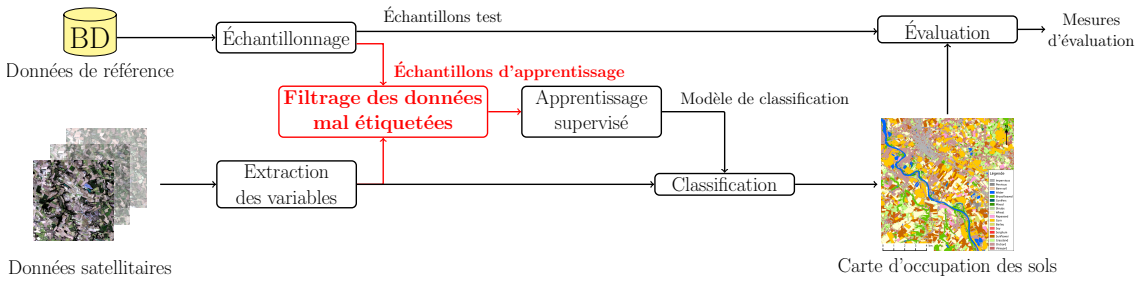


FIGURE 7.1 – Processus de classification supervisée avec une étape de filtrage des données mal étiquetées (en rouge) avant l'étape d'apprentissage supervisé.

L'ajout de l'étape de filtrage permet de faciliter l'étape d'apprentissage puisque idéalement les données mal étiquetées ne sont plus présentes. Ainsi, la fiabilité de la règle de décision est augmentée. De plus, un tel filtrage permet d'identifier les données potentiellement mal étiquetées qui peuvent être ensuite analysées, par exemple par un expert. Cette connaissance peut permettre de corriger les données mal étiquetées [Gamberger et al., 2000].

Plus précisément, l'étape de filtrage se déroule en deux étapes principales comme le montre la Figure 7.2. Tout d'abord, la détection des données mal étiquetées est réalisée afin de diviser en deux sous-ensembles les échantillons d'apprentissage : 1) les données correctement étiquetées qui seront utilisées telles quelles, et 2) les données mal étiquetées. Concernant les données identifiées comme étant mal étiquetées, elles peuvent subir

différents traitements. Elles sont soit supprimées définitivement de l'ensemble d'apprentissage, soit réinjectées avec les données correctement étiquetées après modification (*e.g.* changement d'étiquette). Dans un premier temps, nous considérons le cas le plus fréquent qui consiste à supprimer les échantillons identifiés comme étant mal étiquetés. La Section 7.1.3 est dédiée aux autres approches qui permettent de réinjecter les données mal étiquetées.

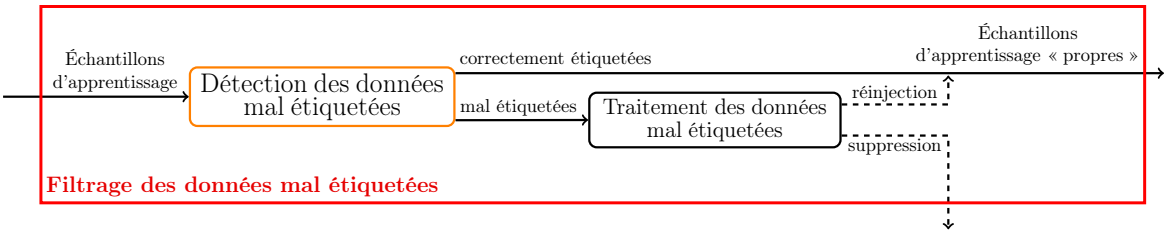
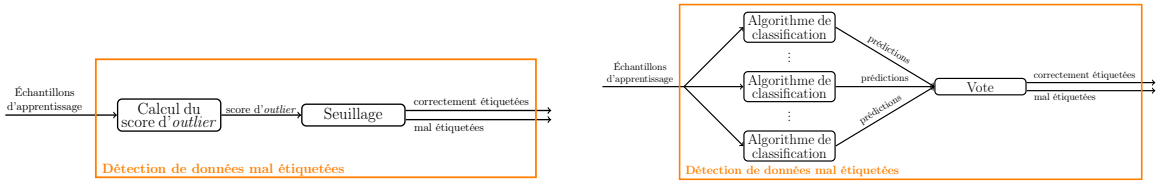


FIGURE 7.2 – Illustration du principe général du filtrage des données mal étiquetées.

Concernant la détection des données mal étiquetées, deux approches sont possibles. La première, illustrée par la Figure 7.3a, consiste à utiliser un algorithme de détection de données mal étiquetées comme ceux introduits au Chapitre 6. Tout d'abord, un score d'*outlier* est calculé pour chaque échantillon. Puis, un seuil est utilisé comme critère de décision afin de séparer les échantillons en deux sous-ensembles : correctement étiqueté et mal étiqueté. Cette approche est détaillée dans la Section 7.1.1. La seconde approche, illustrée par la Figure 7.3b, consiste à utiliser les prédictions d'un ensemble d'algorithmes de classification appris sur les données d'apprentissage. Ainsi les données mal étiquetées seront les données pour lesquelles la classe prédite par l'ensemble de classifieurs est différente de celle fournie par la donnée de référence. Cette approche est détaillée dans la Section 7.1.2.



(a) Score d'*outlier* et seuillage

(b) Méthodes d'ensemble et vote

FIGURE 7.3 – Stratégies possibles pour la détection des données mal étiquetées.

Les deux stratégies présentées à la Figure 7.3 ont un principe de fonctionnement similaire. Le degré d'anomalie des échantillons est tout d'abord mesuré en calculant un score d'*outlier* ou en utilisant un ensemble d'algorithmes de classification. Puis, une étape de décision – le seuillage ou le vote – permet de diviser les échantillons en deux catégories : correctement étiquetée *versus* mal étiquetée.

7.1.1 Score d'*outlier* et seuillage

Les techniques de filtrage décrites dans cette partie reposent sur l'utilisation de méthodes de détection d'*outliers*. Comme vu à la Section 6.1.2, les méthodes de détection d'*outliers* calculent un score d'*outlier* qui représente le degré d'anomalie supposé des échantillons. Ensuite, une étape de division permet de séparer en deux sous-ensembles indépendants les échantillons en se basant sur les valeurs des scores d'*outlier*.

Pour faire cette séparation, la stratégie adoptée dépend principalement de la distribution des scores d'*outlier*, et donc de la méthode de détection d'*outlier* utilisée. Par exemple, les méthodes d'édition de type ENN ont des scores d'*outlier* binaires. Dans ce cas là, la séparation des échantillons en deux sous-ensembles est automatique. Tous les échantillons identifiés comme étant des *outliers* par l'algorithme seront considérés comme des données mal étiquetées. Néanmoins, une majorité des méthodes de détection d'*outlier* calculent des scores d'*outlier* non-binaires. La définition d'un seuil est alors nécessaire pour l'étape de décision.

Dans la littérature, deux stratégies sont utilisées pour définir la valeur de ce seuil. La première consiste à déterminer, la valeur du score d'*outlier* au-delà de laquelle les échantillons seront considérés comme étant mal étiquetés. Cette stratégie est difficile à mettre en place car la dispersion des scores d'*outlier* dépend de la méthode utilisée, du jeu de données étudié, et des classes étudiées (le nombre de classes et la variabilité intra-classes).

Par exemple, pour le score d'*outlier* calculé avec le RF, Breiman propose 10 comme valeur de seuil. Ainsi, tous les échantillons ayant un score d'*outlier* supérieur à 10 sont considérés comme des *outliers*. Suivant cette recommandation, certains travaux suppriment de leurs données les échantillons dont le score d'*outlier* est supérieur à 10 [Pang et al., 2006; Rodríguez-Galiano et al., 2012; Touw et al., 2012]. Cependant, d'autres travaux ont proposé des valeurs différentes pour ce seuil [Tsuji et al., 2012; Zhou and Zhang, 2016].

Une seconde stratégie pour définir le seuil consiste à identifier comme mal étiquetés les n échantillons avec les scores d'*outlier* les plus forts [Hewahi and Saad, 2007; Ramaswamy et al., 2000]. Cette stratégie ne nécessite donc pas de connaissance sur la dispersion des valeurs d'*outlier*. Si les scores d'*outlier* représentent bien le degré d'anomalie des échantillons, alors la valeur optimale de n correspond au nombre réel d'échantillons mal étiquetés présents dans les données.

En pratique, le nombre de données mal étiquetées est inconnu. Le niveau de bruit doit donc être estimé afin de trouver la valeur optimale de n. À notre connaissance peu de travaux cherchent à faire cette estimation. Néanmoins, Garcia et al. [2015] proposent de considérer que le niveau de bruit présent dans les données correspond à l'erreur commise par l'algorithme 1-PPV appris sur les échantillons bruités. Malheureusement, cette estimation du nombre de données mal étiquetées est biaisée puisque l'algorithme des 1-PPV est entraîné à partir de données mal étiquetées.

Quelque soit la stratégie utilisée, la configuration dans l'étape de décision de la valeur du seuil n est une tâche complexe : soit peu d'échantillons mal étiquetés sont éliminés, soit trop d'échantillons correctement étiquetés sont éliminés.

7.1.2 Ensemble d'algorithmes de classification et vote

Les méthodes de détection de données mal étiquetées décrites ici reposent sur l'utilisation des prédictions d'un ensemble d'algorithmes de classification. Pour ces méthodes, les échantillons extraits des données de référence, qui contiennent des données mal étiquetées, sont utilisés pour l'apprentissage de différents modèles de classification. Ces modèles sont ensuite appliqués pour prédire la classe des échantillons d'apprentissage. L'utilisation de plusieurs modèles de classification permet de diversifier les points de vue. Si pour un échantillon donné les classes prédites par les modèles de classification sont souvent différentes de la classe fournie par la donnée de référence, alors l'échantillon est probablement mal étiqueté [Brodley and Friedl, 1999]. La Figure 7.4 montre le principe de fonctionnement de ces approches.

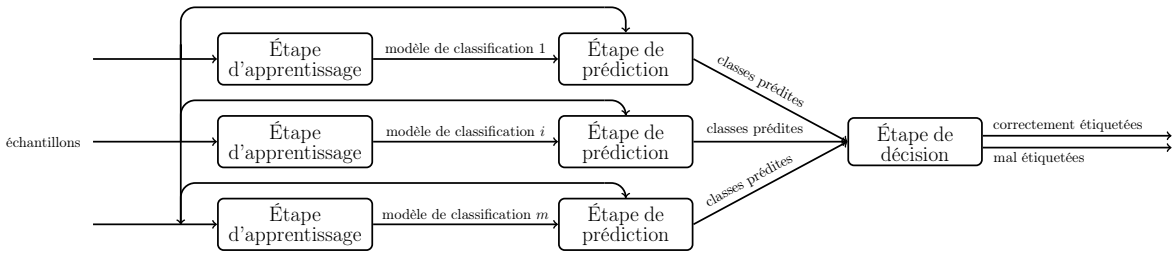


FIGURE 7.4 – Principe simplifié des méthodes de détection de données mal étiquetées basées sur un ensemble d'algorithmes de classification.

Dans le schéma simplifié de la Figure 7.4, tous les échantillons sont utilisés pour l'apprentissage des algorithmes de classification. L'objectif est alors d'identifier les données mal étiquetées présentes dans cet ensemble d'échantillons. Idéalement, les classes prédites par les modèles de classification doivent être identiques à celles fournies par la donnée de référence. Si cette hypothèse n'est pas vérifiée, les échantillons doivent être identifiés comme étant mal étiquetés par l'étape de décision.

Dans l'exemple de la Figure 7.4, les algorithmes de classification sont tous appris avec les mêmes échantillons. Ainsi, la diversité de l'ensemble des modèles de classification est probablement pauvre. Les modèles vont alors être similaires et prédire les mêmes classes. De plus, les performances de classification vont être impactées négativement si les échantillons utilisés en apprentissage contiennent des données mal étiquetées. Par ailleurs, les modèles de classification prédisent la classe d'échantillons utilisés en apprentissage. Ces prédictions peuvent donc être biaisées.

Afin de remédier à ces problèmes, une solution possible consiste à apprendre l'ensemble des algorithmes de classification sur différents sous-ensembles d'échantillons. Cette solution a pour avantage d'augmenter la diversité dans l'ensemble des algorithmes de classification, et d'utiliser les modèles de classification sur des échantillons qui n'ont pas participé à l'apprentissage.

En prenant en compte ces considérations, les approches de détection de données mal étiquetées basées sur un ensemble d'algorithmes de classification nécessitent de définir :

1. la stratégie d'échantillonnage qui sépare les échantillons en sous-ensembles,
2. le type et le nombre d'algorithmes de classification à utiliser,
3. le critère permettant de combiner les votes des algorithmes de classification lors de l'étape de décision.

Dans la littérature, le type et le nombre d'algorithmes de classification à utiliser sont directement liés à la stratégie d'échantillonnage choisie. En général, deux stratégies d'échantillonnage sont proposées dans la littérature.

La première stratégie consiste à apprendre un ensemble de m classifieurs sur des partitions des échantillons. Le partitionnement des échantillons consiste à diviser en q partitions indépendantes les échantillons disponibles. L'apprentissage des classifieurs est ensuite réalisé selon deux approches 1) sur une partition [Khoshgoftaar et al., 2007], soit 2) sur l'union de $q-1$ partitions [Brodley and Friedl, 1999; Smith and Martinez, 2015]. Dans les deux types d'approches, les modèles de classification sont ensuite utilisés pour prédire les classes des échantillons non-utilisés pour l'apprentissage. Pour le premier type d'approche, la classe de chaque échantillon est prédite par $(q-1) \times m$ algorithmes de classification. Pour le second type d'approche, la classe de chaque échantillon est prédite seulement par m algorithmes de classification.

Parmi les approches qui réalisent l'apprentissage sur l'union de $q-1$ partitions, la méthode la plus connue est celle de Brodley and Friedl [1999]. Les échantillons sont tout d'abord divisés en dix partitions ($q = 10$). Pour chaque union de neuf partitions, trois algorithmes de classification ($m = 3$) sont entraînés : arbre de décision binaire C4.5, 1-PPV et analyse discriminante linéaire. La stratégie de partitionnement en 10 partitions a aussi été utilisée dans l'approche *Noise Identification using Classifier Diversity* (NICD) proposée par Smith and Martinez [2015]. Dans cette approche, l'ensemble des algorithmes de classification utilisés est sélectionné minutieusement. Pour ce faire, une vingtaine d'algorithmes de classification sur 129 jeux de données est étudiée. L'objectif est de déterminer l'ensemble d'algorithmes de classification qui permet de maximiser la diversité entre classifieurs, en favorisant des comportements différents sur les différents jeux de données. Ces travaux recommandent l'utilisation de neuf algorithmes.

Ces stratégies conduisent à la présence de données mal étiquetées dans chaque partition. En trop grande quantité, cette présence peut conduire à des classifieurs de mauvaise qualité [Ali and Pazzani, 1996; Yuan et al., 2016]. Afin de minimiser ce risque, une solution consiste à nettoyer les données de chaque partition avant l'apprentissage des algorithmes de classification. Par exemple, Gamberger et al. [1999] reprennent la méthode proposée par Brodley and Friedl [1999] en excluant de l'apprentissage les échantillons identifiés comme mal étiquetés par un filtre de saturation (Section 6.1.1).

La seconde stratégie d'échantillonnage consiste à apprendre le même algorithme de classification ($m = 1$) sur q sous-ensembles d'échantillons qui ne sont pas nécessairement indépendants [Gamberger et al., 1999; Sluban et al., 2010; Verbaeten and Van Assche, 2003; Zhu et al., 2003]. Dans ce cas, les q prédictions des algorithmes de classification sont utilisées pour tous les échantillons. Afin de créer les sous-ensembles d'échantillons d'apprentissage, les techniques de *bagging* et de *boosting* (Section 2.4.2) ont été utilisées. Par exemple, Verbaeten and Van Assche [2003] proposent l'utilisation d'un ensemble d'arbres de décision binaire C4.5 appris sur des échantillons *bootstrap*. Une approche similaire a été proposée en remplaçant l'ensemble d'arbres de décision binaire C4.5, par les arbres appris par un modèle de RF [Sluban et al., 2010]. Concernant le *boosting*, Verbaeten and Van Assche [2003] proposent d'entraîner un algorithme d'*AdaBoost* (Section 2.4.2). Les échantillons pour lesquels l'algorithme *AdaBoost* attribue un fort poids sont supprimés au fur et à mesure de l'apprentissage de l'ensemble.

Quelque soit la stratégie utilisée pour construire l'ensemble des algorithmes de classification, il faut combiner les prédictions des algorithmes de classification afin d'identifier les échantillons mal étiquetés. Dans la littérature, trois critères de décision sont généralement proposés : 1) le vote majoritaire, un échantillon est identifié comme mal étiqueté si une majorité des classifieurs votent pour une autre classe que celle de la donnée de référence ; 2) le consensus, un échantillon est identifié comme mal étiqueté si tous les classifieurs votent pour une autre classe que celle de la donnée de référence, et 3) le seuillage, un échantillon est identifié comme mal étiqueté si x % des classifieurs votent pour une autre classe que celle de la donnée de référence. Le vote majoritaire et le consensus sont donc des cas particuliers du seuillage pour lesquels x est égal à 50 et 100 % respectivement.

Le consensus, plus restrictif, peut ne pas détecter certaines données mal étiquetées, tandis que le vote majoritaire peut éliminer un trop grand nombre de données correctement étiquetées. Ainsi, le seuillage peut permettre de trouver un bon compromis entre éliminer trop d'échantillons correctement étiquetés ou garder trop d'échantillons mal étiquetés, mais il nécessite la configuration du seuil x [Smith and Martinez, 2015]. Aucune stratégie n'est donc parfaite, et celle optimale dépend généralement du contexte. Par exemple, dans le cas où un grand nombre d'échantillons d'apprentissage est disponible, le vote majoritaire

donne généralement de meilleurs résultats que le consensus [Brodley and Friedl, 1999].

7.1.3 Suppression, correction ou pondération

Les méthodes de détection de données mal étiquetées décrites précédemment sont généralement utilisées dans un filtrage (Figure 7.2) qui supprime les échantillons identifiés comme étant mal étiquetés. Cependant, la suppression est une étape de traitement des données mal étiquetées assez agressive. D'une part, elle fait diminuer le nombre d'échantillons d'apprentissage et, d'autre part elle supprime bien souvent des échantillons correctement étiquetés. Ainsi, d'autres stratégies ont été testées dans la littérature, notamment la correction et la pondération. Ces autres stratégies sont principalement utilisées avec les approches basées sur les prédictions des algorithmes de classification, mais aussi pour des méthodes de détection de données mal étiquetées qui calculent une probabilité comme la méthode SOS présentée au Chapitre 6.

La correction, ou encore ré-étiquetage [65], consiste à modifier l'étiquette de l'échantillon vers la classe la plus souvent prédite par les classifieurs. Bien que le ré-étiquetage peut être bénéfique pour les performances de l'algorithme de classification final [Teng, 2001; Zeng and Martinez, 2001], cette stratégie donne généralement de moins bons résultats que la suppression [Brodley and Friedl, 1999; Feng et al., 2015]. Ce résultat est assez attendu puisque qu'il est difficile d'obtenir une correction fiable qui n'introduit pas de nouvelles mauvaises étiquettes.

Une solution prometteuse est la combinaison de la suppression et du ré-étiquetage. Dans ce cas, les échantillons ré-étiquetés sont ceux pour lesquels la confiance dans la classe prédite est élevée [Koplowitz and Brown, 1981; Lallich et al., 2002; Muhlenbach et al., 2004]. Les autres échantillons identifiés comme mal étiquetés sont supprimés. Par exemple, Barandela et al. [2003] adaptent les travaux de Brodley and Friedl [1999] pour lesquels un ensemble d'algorithme de classification est utilisé pour identifier les données mal étiquetées. Dans cette variante, les échantillons identifiés comme mal étiquetés sont ré-étiquetés si la classe prédite par plus de la moitié des classifieurs est la même, sinon ils sont supprimés.

Finalement, le dernier traitement des données identifiées comme mal étiquetées est la pondération. Cette dernière consiste à affecter un poids à tous les échantillons qui sera ensuite utilisé par l'algorithme de classification final. Ce poids est directement calculé par la méthode utilisée pour détecter les données mal étiquetées. Il correspond généralement à la probabilité que l'échantillon appartienne à la classe fournie par la donnée de référence. Ce poids traduit donc un degré de confiance que peut avoir l'algorithme de classification final dans l'étiquette de l'échantillon. Dans cette approche, les échantillons ne sont donc pas divisés en deux sous-ensembles correctement étiquetés et mal étiquetés. Tous les échantillons disponibles sont utilisés pour l'apprentissage de l'algorithme de classification final. Cependant, les échantillons avec les poids les plus élevés auront une influence plus forte sur la règle de décision apprise.

Par exemple, l'approche PWEM propose comme poids, le vecteur de probabilité d'appartenance aux classes calculé pour chaque échantillon [Rebbapragada and Brodley, 2007]. La probabilité que l'échantillon appartienne à la classe indiquée par la donnée de référence est alors utilisée pour pondérer les échantillons lors de l'apprentissage du classifieur final. De manière similaire, Smith and Martinez [2015] utilisent la probabilité calculée par la méthode NICD (Section 6.1.1) pour pondérer les échantillons.

65. Dans la littérature, la notion de correction peut aussi faire référence à la correction de la valeur d'un ou plusieurs attributs. Ce cas là n'est pas abordé dans ce manuscrit.

L'inconvénient de la stratégie de pondération est que l'algorithme de classification final doit être capable de prendre en compte le poids des échantillons lors de la construction de sa règle de décision. Dans la littérature, certains travaux ont proposé d'adapter des algorithmes de classification existants [Smith and Martinez, 2015]. Par exemple, une modification de la construction des arbres de décision binaire C4.5 a été introduite. L'idée est de modifier le calcul de l'entropie utilisée pour diviser les échantillons d'apprentissage. Pour le cas du RF, le poids des échantillons peut être utilisé à la fois au moment de la sélection des échantillons *bootstrap* et au moment du calcul du coefficient de Gini [Chen et al., 2004; Smith and Martinez, 2015].

Les résultats de Smith and Martinez [2015] montrent que la pondération est aussi efficace que la suppression. De plus, cette stratégie a pour avantage de ne pas nécessiter la définition d'un critère de décision comme le seuil ou le vote.

7.2 Filtrage itératif des données mal étiquetées

Les stratégies de filtrage décrites dans les parties précédentes reposent sur l'utilisation d'une méthode de détection de données mal étiquetées. Quelque soit la méthode de détection utilisée, la définition d'un critère de décision (seuillage ou vote) est nécessaire pour diviser les données en deux sous ensembles : correctement étiqueté *versus* mal étiqueté. Néanmoins, la valeur du seuil ou le type de vote est un paramètre généralement difficile à configurer. Une valeur de seuil trop faible ou un vote restrictif peut mener à une sous-détection des données mal étiquetées. Au contraire, une valeur de seuil trop élevée ou un vote indulgent peut conduire à la sur-détection.

Afin de réduire l'influence du choix du seuil ou du type de vote, une solution consiste à utiliser un filtrage itératif [Sáez et al., 2016]. Dans cette approche, une petite portion des échantillons mal étiquetés sont identifiés et traités à chaque itération. L'objectif du filtrage itératif est double :

1. L'identification d'une sous-partie des échantillons mal étiquetés permet d'augmenter la précision de la détection des données mal étiquetées. Comme en théorie la valeur du seuil est faible ou le vote est restrictif, peu de données correctement étiquetées sont identifiées comme mal étiquetées à chaque itération.
2. Le traitement au fur et à mesure des échantillons potentiellement mal étiquetés permet d'améliorer l'identification des données mal étiquetées aux itérations suivantes. Théoriquement, les méthodes de détection de données mal étiquetées seront appliquées sur des données qui contiennent de moins en moins d'échantillons mal étiquetés.

La Figure 7.5 illustre le principe du filtrage itératif. Dans cette approche, les données identifiées comme mal étiquetées sont soit supprimées, soit réinjectées (corrigées ou pondérées) avec les données identifiées comme correctement étiquetées. À chaque itération t, un critère d'arrêt est évalué sur l'ensemble des données. Si le critère d'arrêt n'est pas vérifié, la méthode de détection de données mal étiquetées est une nouvelle fois appliquée.

Le filtrage itératif des données mal étiquetées nécessite un critère d'arrêt afin de décider à quel moment le processus s'arrête. De plus, un compromis entre le critère d'arrêt et le critère de décision est nécessaire. Si peu de données sont identifiées comme mal étiquetées à chaque itération, la détection sera probablement très précise. Cependant, le nombre d'itérations nécessaires pour obtenir un ensemble d'échantillons « propres » sera élevé. Comme les temps de calcul sont proportionnels au nombre d'itérations, le choix du critère d'arrêt et de décision est donc critique.

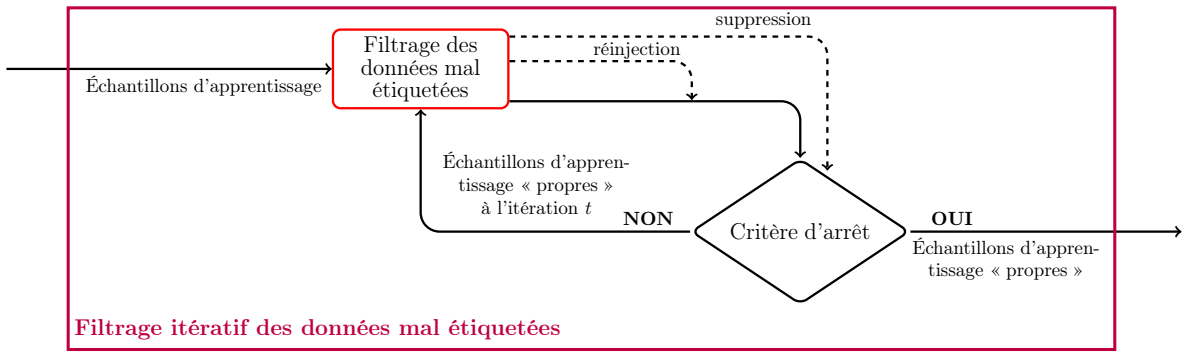


FIGURE 7.5 – Illustration du principe du filtrage itératif.

Dans les parties suivantes, différentes stratégies de filtrage itératif sont présentées. Une première partie est consacrée aux filtrages itératifs utilisés dans la littérature. Puis, une seconde partie décrit les processus de filtrage itératif proposés dans ces travaux de thèse. Les deux nouvelles stratégies de filtrage proposées reposent sur l'utilisation des scores d'*outlier* calculés par le RF et les prédictions faites par l'ensemble des arbres du RF.

7.2.1 Filtrage itératif dans la littérature

Dans la littérature, les deux types de méthodes de détection de données mal étiquetées décrites à la Section 7.1 sont utilisés dans des processus itératifs. Par conséquent, les critères de décision et d'arrêt de ces processus dépendent des méthodes de détection utilisées. Dans la suite, différents critères d'arrêt proposés dans la littérature sont étudiés 1) pour les méthodes basées sur le calcul d'un score d'*outlier*, 2) pour les méthodes basées sur un ensemble d'algorithmes de classification.

Concernant les méthodes basées sur le calcul d'un score d'*outlier*, la définition du critère d'arrêt est plus ou moins évidente. Par exemple, le critère d'arrêt est imposé lorsque le score d'*outlier* est binaire. En effet, le filtrage s'arrête automatiquement lorsque plus aucune donnée n'est identifiée comme étant mal étiquetée. Dans cette catégorie, la méthode plus connue est la version itérative de la méthode ENN : la méthode RENN [Tomek, 1976]. La méthode MULTIEDIT est une variante pour laquelle la méthode RENN est appliquée sur des partitions des données [Devijver and Kittler, 1980]. Un autre exemple est la méthode *Iterated Training Sample Selection* (ITSS) qui consiste à appliquer itérativement la méthode de Jia et al. [2014a] décrite à la Section 6.1.1.

Cependant, la définition du critère d'arrêt n'est pas aussi simple lorsque le score d'*outlier* n'est pas binaire. Dans la littérature, peu de travaux ont été faits dans ce sens. La proposition la plus couramment utilisée repose sur le calcul de la complexité d'un algorithme de classification appris avec les échantillons en sortie de l'étape de filtrage. Une baisse significative de la complexité correspond alors au moment où le filtrage itératif est arrêté [Gamberger and Lavrač, 1997].

Par ailleurs, l'utilisation d'un score d'*outlier* non-binaire nécessite la définition d'un seuil dans l'étape de décision à chaque itération. Comme mentionné à la Section 7.1.1, deux stratégies sont possibles pour définir le seuil. Dans la littérature, le seuil retenu est généralement basé sur le nombre d'échantillons à éliminer. Lors d'un filtrage itératif, la stratégie la plus fiable consiste à supprimer à chaque itération un seul échantillon, celui avec le plus fort score d'*outlier* [Gamberger and Lavrač, 1997]. Cependant, cette approche nécessite un grand nombre d'itérations pour éliminer tous les échantillons mal étiquetés.

Les temps de calcul seront donc augmentés. Au contraire, une valeur de seuil permettant la suppression de plusieurs échantillons à chaque itération permet de réduire le nombre total d'itérations, mais des échantillons correctement étiquetés seront susceptibles d'être éliminés. Un compromis est donc nécessaire entre la rapidité de l'étape de filtrage et sa précision.

De manière similaire, le critère d'arrêt doit être défini pour les filtrages itératifs basés sur un ensemble d'algorithmes de classification (Section 7.1.2). Pour rappel, les échantillons identifiés comme mal étiquetés sont ceux pour laquelle la classe prédite est différente de celle fournie par la donnée de référence. À chaque itération, ces échantillons sont traités (supprimés, corrigés ou pondérés). Puis, l'ensemble des algorithmes de classification est à nouveau appris avec les échantillons encore présents après l'étape de filtrage. Dans ce cas, le processus itératif s'arrête donc automatiquement lorsque la classe prédite pour tous les échantillons est identique à celle fournie par la donnée de référence. Cependant, de nombreuses itérations peuvent être nécessaires avant que le processus s'arrête automatiquement. Ainsi, il a été proposé d'arrêter le processus si au cours des trois dernières itérations moins de 1 % d'échantillons du nombre total d'échantillons est supprimé de l'ensemble d'apprentissage.

Les méthodes de filtrage itératif les plus connues basées sur ce principe sont la méthode *Partitioning Filter* [Zhu et al., 2003] et sa variante *Iterative Partitioning Filter*[66] [Rebours, 2004]. Ces méthodes appliquent de manière itérative l'approche de Khoshgoftaar et al. [2007] décrite dans la Section 7.1.2. Ce critère d'arrêt est aussi utilisé dans la méthode *Iterative Noise Filter based on the Fusion of Classifiers* (INFFC) [Sáez et al., 2016]. Ce processus itératif plus complexe est composé de deux étapes. La première étape détecte les données mal étiquetées en utilisant un ensemble d'algorithmes de classification. La seconde étape utilise les échantillons identifiés comme correctement étiquetés à la première étape pour apprendre un nouvel ensemble d'algorithmes de classification. Cet ensemble est supposé plus fiable que celui appris à la première étape puisque la présence des données mal étiquetées est théoriquement réduite [Gamberger et al., 1999]. Ainsi, les prédictions de cet ensemble sont utilisées sur tous les échantillons d'apprentissage afin d'identifier les données mal étiquetées. La procédure s'arrête lorsque le nombre d'échantillons supprimés au cours des trois dernières itérations représente moins de 1 % du nombre total d'échantillons.

7.2.2 Filtrage itératif avec le *Random Forest*

Dans cette partie, nous proposons de nouvelles méthodes de filtrage itératif basées sur l'utilisation du RF. Suivant les deux stratégies de détection de données mal étiquetées présentées dans les Sections 7.1.1 et 7.1.2, deux filtrages itératifs sont proposés.

Le premier filtrage est basé sur la détection de données mal étiquetées en utilisant les scores O_{RF} présentés à la Section 6.2.1. Dans ce cas, deux critères de décision basés sur les scores d'*outliers* sont proposés. De plus, différents critères d'arrêt sont analysés.

Le second filtrage repose sur la combinaison des prédictions faites par l'ensemble des arbres d'un modèle RF. Dans celui-ci, les prédictions obtenues pour l'ensemble des arbres sont interprétées comme un vecteur d'appartenance aux classes. Ainsi, un échantillon est identifié comme étant mal étiqueté si la valeur maximale de ce vecteur est obtenue pour une autre classe que celle fournie par la donnée de référence. Dans ces travaux, plusieurs vecteurs de probabilités sont utilisés.

66. La différence entre le *Partitioning Filter* et le *Iterative Partitioning Filter* réside seulement dans le choix de l'algorithme de classification.

Basé sur le score d'*outlier* du *Random Forest*

Les premiers filtrages itératifs proposés se basent sur l'utilisation de la méthode de détection d'*outliers* présentée à la Section 6.2. En partant d'un ensemble d'échantillons, cette méthode de détection est appliquée de manière itérative. À chaque itération, les scores O_{RF} sont calculés pour chaque échantillon. Les bonnes performances des scores d'*outlier* du RF – Breiman, DistanceLCA et PuretyLCA – ont été montré au Chapitre 6. Ces trois scores sont donc utilisés dans ces filtrages itératifs.

Le filtrage repose sur l'utilisation de scores d'*outlier* non-binaires. Par conséquent, la définition d'un critère de décision est nécessaire. En évaluant les valeurs O_{RF}, ce critère permet de prendre une décision à chaque itération et pour chaque échantillon : correctement étiqueté ou mal étiqueté. Pour ce filtrage, la détection des données mal étiquetées est répétée tant que le critère d'arrêt n'est pas respecté. Dans ces travaux, la définition de deux critères de décision donne lieu a deux types de filtrages itératifs.

Le premier filtrage itératif proposé considère que les n échantillons avec les plus forts scores d'*outlier* O_{RF} sont des données mal étiquetées. En considérant ce critère de décision, n échantillons sont supprimés à chaque itération [Hewahi and Saad, 2007; Ramaswamy et al., 2000]. Comme indiqué dans la Section 7.2.1, la configuration de n permet de régler le compromis entre la précision de la méthode et le nombre d'itérations nécessaires pour atteindre les performances optimales.

L'utilisation itérative de cette méthode de détection nécessite la définition d'un critère d'arrêt. Dans ces travaux, le critère d'arrêt proposé repose sur l'analyse du score d'*outlier* du n-ième échantillon supprimé. Idéalement, les scores d'*outlier* sont élevés dans les premières itérations à cause de la présence de données mal étiquetées. Lors de la suppression de données mal étiquetées au cours des itérations, les valeurs des scores d'*outlier* doivent théoriquement diminuer. Soit $O^t_{RF}(n)$ la valeur du score d'*outlier* du n-ième échantillon supprimé à l'itération t. Un critère d'arrêt possible consiste à calculer la différence $O^{t-1}_{RF}(n) - O^t_{RF}(n)$ entre deux itérations consécutives. Cette différence doit être quasiment nulle quand quasiment toutes les données mal étiquetées ont été supprimées.

Dans notre contexte, les échantillons à filtrer appartiennent à différentes classes. Or, cette information n'est pas utilisée dans le critère de décision basé sur le seuil n. Comme la dispersion des valeurs O_{RF} est spécifique à chaque classe, le seuillage sur les n plus fortes valeurs d'*outlier* peu avoir certaines limitations.

Considérons une classe c_i pour laquelle les échantillons ont des scores d'*outlier* en moyenne plus élevés que les échantillons appartenant aux autres classes. La suppression des n échantillons ayant les plus forts scores d'*outlier* peut alors conduire à l'élimination de l'ensemble des échantillons de la classe c_i. De plus, il parait plus judicieux de supprimer un grand nombre d'échantillons dans les premières itérations, puis de réduire au fil des itérations, car en théorie de moins en moins d'échantillons mal étiquetés sont présents. Pour prendre en compte cette observation, une solution possible consiste à utiliser un critère de décision défini par classe basé sur les valeurs des scores d'*outlier* des échantillons. Dans ce cas, les valeurs du seuil utilisées dans l'étape de décision sont différentes entre les classes et changent au cours des itérations.

Ainsi, le second filtrage itératif proposé applique un critère de décision par classe. Par hypothèse, nous considérons que la distribution des scores O_{RF} pour les échantillons appartenant à une même classe est proche d'une loi normale. Ainsi, les plus forts scores d'*outlier* d'une classe représentent théoriquement des valeurs extrêmes de la distribution. En prenant en compte cette hypothèse, l'utilisation de la règle des 3σ est proposée comme critère de décision. Pour chaque échantillon p ayant pour classe de référence c_r, la règle des 3σ considère que l'échantillon p est un *outlier* si son score d'*outlier* $O_{RF}(p)$ représente

une valeur extrême, *i.e.* $O_{RF}(p) > \overline{O_{RF}^{c_r(p)}} + 3\sigma_{O_{RF}^{c_r(p)}}$ où $\overline{O_{RF}^{c_r(p)}}$ et $\sigma_{O_{RF}^{c_r(p)}}$ représentent la moyenne et l'écart-type des scores d'*outlier* pour les échantillons appartenant à la classe $c_r(p)$ [67]. À chaque itération, la règle des 3σ est appliquée sur tous les échantillons. Les échantillons ne respectant pas cette règle sont alors supprimés de l'ensemble des échantillons d'apprentissage. Ce filtrage a pour avantage de ne pas supprimer le même nombre d'échantillons par classe. Cette stratégie peut particulièrement être bénéfique dans le cas ou les données mal étiquetées ne sont pas réparties équitablement entre classe.

Pour le filtrage utilisant la règle des 3σ comme critère de décision, la définition d'un critère d'arrêt n'est pas nécessaire. En effet, la détection des données mal étiquetées pour chaque classe s'arrête automatiquement lorsque tous les échantillons de cette classe respectent la règle des 3σ.

Afin de résumer les différentes informations, le Tableau 7.1 montre les caractéristiques des deux filtrages itératifs proposés. La principale différence entre les deux stratégies est le choix du critère de décision. Contrairement au seuil n, le critère basé sur la règle des 3σ utilise la classe fournie par la donnée de référence pour l'étape de décision. Pour les deux filtrages, la même méthode de détection de données mal étiquetées et le même traitement sont utilisés à chaque itération.

Tableau 7.1 – Filtrages itératifs proposés basés sur les scores d'*outlier* du *Random Forest* (RF).

	Global	**Par classe**
Méthode de détection : scores d'*outlier* du RF	O_{RF}^{Breiman}, $O_{RF}^{\text{DistanceLCA}}$ et $O_{RF}^{\text{PuretyLCA}}$	O_{RF}^{Breiman}, $O_{RF}^{\text{DistanceLCA}}$ et $O_{RF}^{\text{PuretyLCA}}$
Critère de décision	seuil n	3σ
Traitement	suppression	suppression
Critère d'arrêt	$O_{RF}^{t-1}(n) - O_{RF}^{t}(n) \simeq 0$	$\forall p, O_{RF}(p) \leq \overline{O_{RF}^{c_r(p)}} + 3\sigma_{O_{RF}^{c_r(p)}}$

Basé sur la combinaison des prédictions des arbres du *Random Forest*

Les filtrages itératifs proposés ici utilisent l'ensemble des prédictions des arbres construits par un modèle RF. Dans les filtrages présentés à la Section 7.1.2, la prédiction de chaque classifieur pour chaque classe est comptabilisée. Les échantillons identifiés comme étant mal étiquetés sont alors ceux pour lesquels la majorité des classifieurs prédisent une autre classe que celle fournie par la donnée de référence.

En pratique, le nombre de prédiction pour une classe donnée divisée par le nombre total de prédictions (*i.e.* de classifieurs) peut être vu comme la probabilité que l'échantillon appartienne à cette classe. Par conséquent, le décompte des prédictions peut alors être interprété comme un vecteur de probabilité d'appartenance aux classes.

Dans la suite, le vecteur de probabilité d'appartenance pour un échantillon x est noté $\boldsymbol{p}(x)$. Il est défini tel que $\boldsymbol{p}(x) = [p_{c_1}(x), ..., p_{c_N}(x)]$ avec N le nombre de classes. La probabilité $p_{c_i}(x)$ représente donc la probabilité que l'échantillon x appartienne à la classe c_i.

67. En pratique, les valeurs de moyenne et de l'écart-type sont très sensibles aux données mal étiquetées. Pour chaque classe, les échantillons ayant un score d'*outlier* supérieurs à cinq fois la moyenne ne sont pas pris en compte dans le calcul de l'écart-type.

Dans les filtrages présentés à la Section 7.1.2, le vecteur de probabilité est calculé en utilisant l'ensemble des algorithmes de classification étudiés. L'idée ici est d'utiliser le fait que le RF soit une méthode d'ensemble. Par conséquent, le vecteur de probabilité peut être directement calculé en utilisant l'ensembles des arbres du RF [Sluban, 2014]. En considérant ce vecteur, un échantillon est, dans la littérature, identifié comme étant mal étiqueté lorsque la probabilité pour la classe de la donnée de référence est inférieure à 50 %. Dans ces travaux, nous considérons qu'un échantillon est mal étiqueté si le maximum de probabilité est obtenue pour classe différente de celle fournie par la donnée de référence.

Comme proposé par Breiman (BP), le vecteur de probabilité est calculé en comptabilisant le nombre de prédictions par classe pour l'ensemble des arbres. La probabilité $p_{c_i}(x)$ est alors calculée sur l'ensemble des arbres de la manière suivante :

$$p_{c_i}^{\mathrm{BP}}(x) = \frac{1}{K}\sum_{k=1}^{K} p_{c_i}^{k}(x), \tag{7.1}$$

avec K le nombre d'arbres dans la forêt, et $p_{C_i}^{k}(x)$ la probabilité que le k-ième arbre prédise la classe c_i pour l'échantillon x. La probabilité $p_{c_i}^{k}(x)$ est calculée en étudiant la composition de la feuille $n_k(x)$ où tombe l'échantillon x dans le k-ième arbre. En particulier, le nombre d'échantillons par classe qui ont servi dans la construction de l'arbre k qui sont tombés dans la feuille $n_k(x)$ sont comptabilisés dans le vecteur $\boldsymbol{m}^{\boldsymbol{n_k(x)}}$. Ce dernier est défini tel que : $\boldsymbol{m}^{\boldsymbol{n_k(x)}} = \left[m_{c_1}^{n_k(x)}, ..., m_{c_N}^{n_k(x)}\right]$. La probabilité $p_{c_i}^{k}(x)$ est alors définie de la manière suivante par Breiman :

$$p_{c_i}^{k}(x) = \begin{cases} 1 \text{ si } \operatorname{argmax}(\boldsymbol{m}^{\boldsymbol{n_k(x)}}) = c_i \\ 0 \text{ sinon.} \end{cases} \tag{7.2}$$

Le vecteur de probabilité proposé par Breiman utilise que partiellement l'information fournie dans le vecteur $\boldsymbol{m}^{\boldsymbol{n_k(x)}}$. Considérons un problème de détection à deux classes pour lequel un modèle RF est composé de deux arbres ($K = 2$). Considérons également un échantillon x appartenant à la classe c_1 qui tombe dans les feuilles $n_1(x)$ et $n_2(x)$ pour lesquelles $\boldsymbol{m}^{\boldsymbol{n_1(x)}} = [100, 0]$ et $\boldsymbol{m}^{\boldsymbol{n_2(x)}} = [4, 6]$. Dans ce cas, la probabilité $p_{c_1}^{\mathrm{BP}}(x)$ que x appartienne à la classe c_1 est de 0,5. Pour cet exemple, les valeurs $p_{c_1}^{1}(x)$ et $p_{c_2}^{2}(x)$ sont toutes les deux égales à 1. Cependant, la pureté des deux nœuds est différente. En particulier, les échantillons d'apprentissage tombés dans $n_2(x)$ sont répartis quasiment équi-probablement entre les classes c_1 et c_2. Ainsi, la probabilité $p_{c_2}^{2}(x)$ ne devrait pas être la même que la probabilité $p_{c_1}^{1}(x)$. Afin de prendre en compte cette information, une nouvelle définition du vecteur de probabilité est proposée. Elle est notée $p_{C_i}^{RP1}(x)$ pour l'échantillon x, et s'exprime de la manière suivante :

$$p_{c_i}^{\mathrm{RP1}}(x) = \frac{1}{K}\sum_{k=1}^{K} p_{c_i}^{k}(x) = \frac{m_{c_i}^{n_k(x)}}{\sum_{j=1}^{N} m_{c_j}^{n_k(x)}}. \tag{7.3}$$

En reprenant l'exemple présenté précédemment, la probabilité $p_{c_1}^{\mathrm{RP1}}(x)$ est dorénavant de 0,7. Par conséquent, la $p_{c_1}^{\mathrm{RP1}}(x)$ est plus précise que précédemment. Cependant, une information n'est pas prise en compte dans le calcul des deux vecteurs de probabilités. Dans l'exemple précédent, l'échantillon x suit le même parcours que 100 échantillons appartenant à la classe c_1 dans le premier arbre alors qu'il suit le même parcours de seulement 10 échantillons lors de la construction du second arbre. Comme l'échantillon suit plus d'échantillons dans le premier arbre que le second, le premier arbre devrait avoir

une influence plus forte dans le calcul du vecteur de probabilité. Néanmoins, tous les arbres ont le même poids dans les deux vecteurs de probabilité décrits précédemment. Pour prendre en compte cette limitation, la probabilité $\boldsymbol{p}_{c_i}^{RP2}(x)$ suivante est proposée :

$$\boldsymbol{p}_{c_i}^{\mathrm{RP2}}(x) = \frac{\sum_{k=1}^{K} m_{c_i}^{n_k(x)}}{\sum_{k=1}^{K}\sum_{j=1}^{N} m_{c_j}^{n_k(x)}}. \tag{7.4}$$

Dans le cas de l'exemple précédent, la probabilité $\boldsymbol{p}_{c_1}^{RP2}(x)$ est alors de 94,5 %.

En utilisant ces trois vecteurs de probabilité, trois filtrages itératifs sont étudiés. Dans ces travaux, un échantillon x est mal étiqueté si la probabilité maximale est obtenue pour une classe différente de celle fournie par la donnée de référence. Ainsi, l'échantillon x est mal étiqueté si $c_r(x) \neq \underset{c_i}{\operatorname{argmax}}(\boldsymbol{p}(x))$.

Concernant le traitement appliqué aux données identifiées comme mal étiquetées à chaque itération, deux stratégies sont testées :

1. La suppression des échantillons dont la classe obtenant le maximum de probabilité est différente de la classe fournie par la donnée de référence.

2. Le ré-étiquetage des échantillons identifiés comme mal étiquetés. La nouvelle étiquette de ces échantillons est celle obtenant le maximum de probabilité. Dans cette approche, tous les échantillons sont donc gardés au cours des itérations.

Dans ces filtrages, la définition du critère d'arrêt n'est pas nécessaire. En effet, les processus s'arrêtent automatiquement lorsque les prédictions pour tous les échantillons sont en accord avec les étiquettes des échantillons.

Le Tableau 7.2 résume les principales caractéristiques de ces filtrages itératifs se basant sur la combinaison des prédictions des arbres du RF. Pour ces filtrages, deux catégories sont identifiées en fonction du type de filtrage appliqué aux échantillons identifiés comme étant mal étiquetés.

Tableau 7.2 – Filtrages itératifs proposés basés sur les prédictions des arbres du *Random Forest* (RF).

	Suppression	**Ré-étiquetage**
Méthode de détection : combinaison des prédictions du RF	$p_{c_i}^{\mathrm{BP}}(x)$, $p_{c_i}^{\mathrm{RP1}}(x)$ et $p_{c_i}^{\mathrm{RP2}}(x)$	$p_{c_i}^{\mathrm{BP}}(x)$, $p_{c_i}^{\mathrm{RP1}}(x)$ et $p_{c_i}^{\mathrm{RP2}}(x)$
Critère de décision	$c_r(x) \neq \underset{c_i}{\operatorname{argmax}}(\boldsymbol{p}(x))$	$c_r(x) \neq \underset{c_i}{\operatorname{argmax}}(\boldsymbol{p}(x))$
Traitement des données mal étiquetées	suppression	ré-étiquetage $c_r(x) \leftarrow \underset{c_i}{\operatorname{argmax}}(\boldsymbol{p}(x))$
Critère d'arrêt	$\forall x, c_r(x) = \underset{c_i}{\operatorname{argmax}}(\boldsymbol{p}(x))$	$\forall x, c_r(x) = \underset{c_i}{\operatorname{argmax}}(\boldsymbol{p}(x))$

7.3 Présentation des expérimentations

7.3.1 Évaluation

Les méthodes de filtrage décrites dans les parties précédentes visent à obtenir un ensemble d'échantillons d'apprentissage « propres ». Dans cette thèse, l'objectif est de nettoyer les données mal étiquetées d'un ensemble d'échantillons qui sera ensuite utilisé pour l'apprentissage d'un algorithme de classification. Comme vu au Chapitre 5, les données mal étiquetées font diminuer les performances de classification. Par conséquent, l'évaluation des méthodes de filtrage est réalisée en se basant sur deux critères.

Le premier critère d'évaluation utilisé vise à quantifier le gain du filtrage des données mal étiquetées sur les performances de la classification [Brodley and Friedl, 1999; Gamberger et al., 1999; Verbaeten and Van Assche, 2003; Zhu et al., 2003]. Pour ce faire, les données en sortie de l'étape de filtrage sont utilisées pour l'apprentissage d'un algorithme de classification. Dans cette étude, l'algorithme du RF présenté à Section 2.4.2 est utilisé après l'étape de filtrage. En effet, cet algorithme a montré ses bonnes performances – précision, stabilité, robustesse – aux Chapitres 4 et 5. Pour les expérimentations réalisées dans ce chapitre, les paramètres du RF utilisés sont les suivants : le nombre d'arbres K est de 100, le nombre de variables aléatoires sélectionnées à chaque nœud m est égal à $\sqrt{p}$ avec p la dimension du vecteur de variables, la profondeur maximale max_depth est de 25, et le nombre minimal d'échantillons par nœud $min\_samples$ est de 10.

Pour évaluer le gain du filtrage, les performances de l'algorithme de classification obtenues en utilisant les échantillons résultant de l'étape de filtrage sont comparées à trois résultats de référence. Le premier résultat de référence est celui obtenu en utilisant les données bruitées (cas sans filtrage), tandis que les deux autres résultats correspondent aux cas idéaux suivants :

— les résultats sont ceux obtenus en utilisant les données non-bruités (cas sans bruit),

— les résultats sont ceux obtenus après l'utilisation d'un filtrage idéal qui supprimerait toutes (et uniquement) les données mal étiquetées (cas filtrage idéal).

Si le filtrage est efficace, les performances de classification obtenues en utilisant les données filtrées doivent être meilleures que celles obtenues avec les données bruitées. Idéalement, ces résultats doivent être proches des résultats des deux cas idéaux (cas sans bruit et filtrage idéal).

Le second critère d'évaluation vise à mesurer la précision du filtrage, *i.e.* sa capacité à correctement identifier les données mal étiquetées [Brodley and Friedl, 1999; Verbaeten and Van Assche, 2003; Zhu et al., 2003]. Pour évaluer cette précision, la terminologie introduite dans Brodley and Friedl [1999] et reprise dans la récente *review* de Frénay and Verleysen [2014], est ici utilisée. La Figure 7.6 illustre les différentes notions. L'ensemble vert représente l'ensemble T des échantillons d'apprentissage, tandis que le cercle rouge représente l'ensemble des données mal étiquetées M. Le cercle bleu représente les données identifiées comme étant mal étiquetées I après l'étape de filtrage. Ainsi, l'intersection des cercles bleu et rouge correspond à l'ensemble des échantillons mal étiquetés correctement identifiés.

À partir de ces notations, trois métriques sont couramment utilisées : 1) la précision FP[68], 2) l'erreur de type 1 ER_1, et 3) l'erreur de type 2 ER_2.

La précision FP de l'étape de filtrage correspond au rapport entre le nombre d'échantillons mal étiquetés identifiés et le nombre total d'échantillons supprimés. Elle s'exprime

68. L'acronyme FP est pour *filter precision* en anglais.

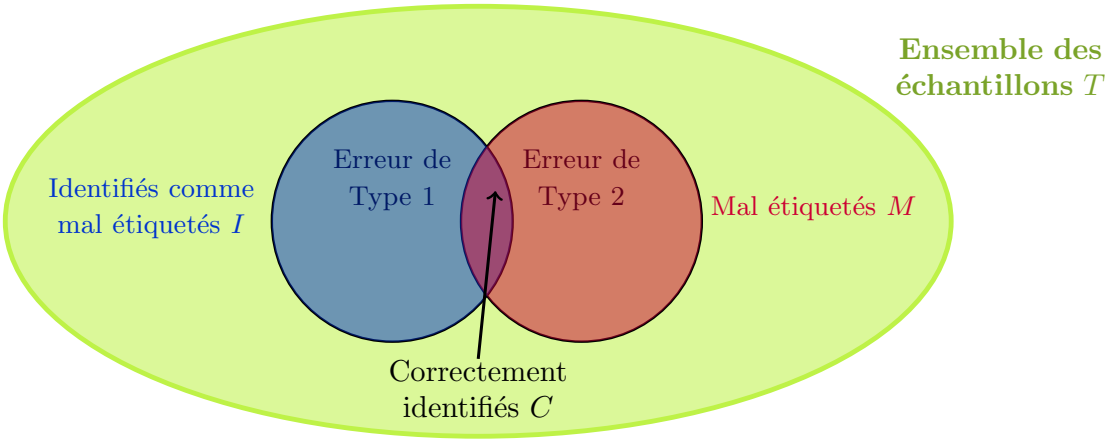


FIGURE 7.6 – Illustration des types d'erreurs lors du filtrage des données mal étiquetées

de la manière suivante :

$$FP = \frac{\blacksquare}{\blacksquare} = \frac{|C|}{|I|}. \tag{7.5}$$

L'erreur de type 1 ER_1 correspond au rapport du nombre d'échantillons correctement étiquetés qui sont identifiés comme mal étiquetés sur le nombre total d'échantillons correctement étiquetés. Elle s'exprime de la manière suivante :

$$ER_1 = \frac{\blacksquare \backslash \blacksquare}{\blacksquare \backslash \blacksquare} = \frac{|I| \backslash |C|}{|T| \backslash |M|}. \tag{7.6}$$

L'erreur de type 2 ER_2 correspond au rapport du nombre d'échantillons mal étiquetés qui ne sont pas identifiés comme tels sur le nombre total d'échantillons mal étiquetés. Elle s'exprime de la manière suivante :

$$ER_2 = \frac{\blacksquare \backslash \blacksquare}{\blacksquare} = \frac{|M| \backslash |C|}{|M|}. \tag{7.7}$$

Dans le cas de l'utilisation de processus itératifs, l'ensemble de ces métriques peuvent être calculées après chaque itération. Il est ainsi possible d'observer l'évolution des métriques au fur et à mesure des itérations.

7.3.2 Données satellitaires et données de référence

Les données utilisées sont quasiment identiques à celles du Chapitre 6. Ainsi, trois jeux de données sont utilisées dans ces expérimentations : simulées, SPOT-Landsat, et Sentinel-2.

Données simulées et SPOT-Landsat

Les données simulées présentées et évaluées au Chapitre 6 sont étudiées ici. Dans ces données, chaque échantillon est décrit par le profil de NDVI simulé pour quinze dates. Par ailleurs, ces données sont composées de cinq classes de végétation.

Concernant les données SPOT-Landsat, les variables utilisées ici sont les bandes spectrales de chaque image et le profil de NDVI. Ces données correspondent donc au jeu de

données BS-NDVI numéro 5 du Tableau 5.4 introduit au Chapitre 5. Elles sont composées de cinq classes, et chaque échantillon est décrit par un vecteur de variables de dimension 119.

Les deux jeux de données sont composés de 500 échantillons d'apprentissage par classe. Le bruit d'étiquetage présenté à la Section 5.2.2 est utilisé pour la génération des données mal étiquetées. Les expérimentations menées ici sont réalisées pour deux niveaux de bruit : 20 % et 40 %.

Afin d'évaluer les performances de classification obtenues après le filtrage, des échantillons test ne contenant pas de données mal étiquetées sont nécessaires. Pour les données simulées et SPOT-Landsat, ces échantillons test sont les mêmes que ceux utilisés au Chapitre 5. Ainsi, chaque jeu de données est composé de 500 échantillons test par classe.

Le Tableau 7.3 montre les résultats de référence décrits à la Section 7.3.1. La première ligne montre les résultats obtenus pour les données simulées, tandis que la seconde ligne montre les résultats obtenus pour les données SPOT-Landsat. La première colonne montre les valeurs d'OA obtenues par un RF lorsqu'aucune donnée mal étiquetée n'est présente dans les échantillons d'apprentissage. La deuxième colonne et la quatrième colonne montrent les résultats obtenus lorsque l'apprentissage du RF est réalisé en utilisant les échantillons d'apprentissage corrompus par 20 et 40 % de bruit respectivement. Enfin, la troisième et cinquième colonne montrent les valeurs d'OA obtenues après un filtrage idéal. Dans ce dernier cas, toutes les données mal étiquetées sont détectées, puis supprimées par l'étape de filtrage. Pour ces résultats de classification, l'algorithme utilisé est le RF dont la configuration est présentée à la Section 7.3.1.

Tableau 7.3 – Valeurs d'*Overall Accuracy* (OA) obtenues pour les données simulées et SPOT-Landsat dans différentes configurations.

	0 % sans bruit	**20 %** sans filtrage	**20 %** filtrage idéal	**40 %** sans filtrage	**40 %** filtrage idéal
Simulées	88,2	83,8	87,3	67,0	84,6
SPOT-Landsat	92,7	90,0	92,3	85,6	92,6

Le Tableau 7.3 montre que la présence de données mal étiquetées influence fortement les performances de la classification, particulièrement pour les données simulées. La perte d'OA est plus forte pour les données simulées car les échantillons sont décrits seulement par le profil de NDVI. Comme montré au Chapitre 5, l'ajout de variables permet d'être plus robuste à la présence de données mal étiquetées.

Par ailleurs, les résultats du Tableau montrent le gain sur les performances apporté par le filtrage idéal. Ce dernier permet par exemple d'augmenter l'OA de 17 % pour un niveau de bruit de 40 % dans le cas des données simulées. Cependant, une perte d'OA peut être observée entre le filtrage idéal et le cas sans bruit. Cette perte est normal puisque la suppression des données mal étiquetées a pour conséquence de réduire le nombre d'échantillons d'apprentissage.

Données Sentinel-2

Pour les données Sentinel-2 acquises en 2016, l'ensemble des échantillons disponibles est étiqueté avec deux données de référence.

La première donnée de référence correspond au RPG de l'année 2014, tandis que la second correspond aux données terrain 2016. Puisque les images Sentinel-2 sont acquises en 2016, les données RPG 2014 sont obsolètes. Ainsi, les échantillons extraits de ces données qui ont changé d'occupation des sols entre 2014 et 2016 sont considérés comme

des données mal étiquetées. Les données terrain 2016 sont quant à elles considérées comme idéales.

Dans ce chapitre, les échantillons décrits par les étiquettes 2014 sont utilisés en entrée de l'étape de filtrage. L'objectif est donc de nettoyer ces données afin de pouvoir réaliser un apprentissage robuste qui obtienne de bonnes performances. Plus exactement, les échantillons utilisés sont ceux présentés dans le Tableau 6.7 à la Section 6.4.3. Au total, six classes de végétation sont présentes, et chaque classe est composée de 500 échantillons. La dimension de chaque vecteur de variable est de 300.

Par ailleurs, les données terrain 2016 sont principalement utilisées pour extraire les échantillons test qui servent à évaluer les performances de classification des différents jeux de données utilisés à l'entrée du classifieur. De plus, elle est aussi utilisée pour extraire des échantillons d'apprentissage non-bruités qui permettront d'évaluer les résultats du cas idéal sans bruit.

Afin de résumer ces informations, le Tableau 7.4 détaille par classe le nombre d'échantillons test, le nombre d'échantillons d'apprentissage et le pourcentage de données mal étiquetées. La dernière colonne correspond au pourcentage de données mal étiquetées par classe présentes dans les données utilisées en apprentissage.

Tableau 7.4 – Nombre d'échantillons test, nombre d'échantillons d'apprentissage et pourcentage de données mal étiquetées pour les données Sentinel-2.

	nb. d'échantillons test (2016)	**nb. d'échantillons d'apprentissage (2014)**	**% ME**
CP	197 992		29,2
M	125 573		24,4
C	43 384	500	30,0
T	115 618		29,8
V	11 973		0,0
P	72 200		20,0
Total	566 740	3000	$\sim$ 22,2 %

CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.
% ME : pourcentage de données mal étiquetées

Comme pour les données simulées et SPOT-Landsat, trois résultats de référence sont calculés : 1) le cas sans bruit, 2) le cas filtrage idéal, et 3) le cas sans filtrage. Ces résultats de classification sont évalués en utilisant les échantillons test décrits dans la première colonne du Tableau 7.4.

Le Tableau 7.5 montre les valeurs d'OA et de F-Score par classe obtenues. La première colonne montre les performances obtenues dans le cas idéal sans bruit. La deuxième colonne montre les performances obtenues sans filtrage dont l'apprentissage est réalisé avec les données bruitées de 2014. Enfin, la troisième colonne montre les valeurs d'OA obtenues après l'utilisation d'un filtrage idéal. Pour ces résultats, l'algorithme de classification utilisé est le RF dont la configuration est présentée à la Section 7.3.1.

Le Tableau 7.5 montre que l'OA diminue de plus de 15 % en présence de données mal étiquetées. Toutes les classes sont impactées par la présence des données mal étiquetées, mais pas de la même manière. Par exemple, la classe colza est très sévèrement touchée par la présence des données mal étiquetées avec un F-Score de 40 %. Au contraire, la classe vigne est moins impactée par la présence des données mal étiquetées. Ce résultat est attendu puisque la classe vigne ne contient pas de données mal étiquetées.

Tableau 7.5 – Valeurs d'*Overall Accuracy* (OA) et de F-Scores obtenues pour les données Sentinel-2 lorsque les données sans bruit, sans filtrage et parfaitement filtrées sont utilisées pour l'apprentissage d'un *Random Forest*.

	2016 sans bruit	**2014** sans filtrage	**2014** filtrage idéal
Céréales à paille	95,5	79,5	92,9
Maïs	96,6	84,3	93,8
Colza	91,4	40,0	80,9
Tournesol	95,8	79,6	93,5
Vignes	84,2	78,8	77,3
Prairies	89,5	79,2	85,4
OA	94,4	78.2	90,9

7.4 Résultats des expérimentations

Dans ces travaux, les stratégies de filtrage étudiées sont divisées en deux catégories : les filtrages non-itératifs et les filtrages itératifs. En suivant ce découpage, cette partie s'intéresse à évaluer et comparer ces deux stratégies. L'évaluation de ces filtrages sera réalisée en utilisant les métriques présentées à la Section 7.3.1.

Pour les deux catégories de filtrage, trois méthodes de détection de données mal étiquetées sont considérées : 1) les méthodes d'édition couramment utilisées dans la littérature, 2) les méthodes basées sur les scores d'*outlier* du RF qui ont montré des bonnes performances au Chapitre 6, et 3) les méthodes basées sur la combinaison des prédictions des arbres construits par le RF. Pour chacune de ces méthodes, le critère de décision permettant d'identifier les données mal étiquetées est analysé.

Une première étude consiste à appliquer une seule fois les méthodes de détection. Outre le critère de décision, ces filtrages non-itératifs nécessitent de définir le traitement à appliquer aux données identifiées comme étant mal étiquetées.

Une deuxième étude consiste à étudier les méthodes de détection de manière itérative. Pour ces filtrages itératifs, les échantillons mal étiquetés sont détectés et traités à chaque itération. Le filtrage s'arrête alors lorsque le critère d'arrêt est respecté. Pour chacune des approches étudiées les critères de décision et d'arrêt seront évalués.

Dans un premier temps, les filtrages non-itératifs sont étudiés. Dans un deuxième temps, les processus itératifs sont analysés. Les meilleurs filtrages non-itératifs et itératifs sont ensuite étudiés sur les données Sentinel-2.

7.4.1 Étude de filtrages non-itératifs

Cette première étude vise à analyser des filtrages utilisant des méthodes de détection de données mal étiquetées de manière non-itérative. Pour ce faire, trois méthodes de détection de données mal étiquetées sont utilisées. La sensibilité des filtrages proposés vis-à-vis de leurs paramètres est analysée pour deux jeux de données. Plus spécifiquement, les données simulées et SPOT-Landsat décrites dans la Section 7.3.2 sont utilisées.

Les différents filtrages sont évalués en mesurant le gain de l'étape de filtrage sur les performances de classification. Pour ce faire, les échantillons restants après l'étape de filtrage sont utilisés pour l'apprentissage d'un algorithme du RF dont la configuration est présentée à la Section 7.3.1. De plus, la précision des filtrages est étudiée en calculant les mesures FP, ER_1 et ER_2 définies à la Section 7.3.1.

Concernant les méthodes de détection de données mal étiquetées, les trois méthodes suivantes sont étudiées :

1. La méthode ENN. Cette méthode classique calcule un score d'*outlier* binaire. Ainsi, les échantillons sont directement identifiés comme étant soit correctement étiquetés soit mal étiquetés. La définition d'un critère de décision est donc automatique. Pour ce filtrage, les échantillons identifiés comme étant mal étiquetés sont supprimés. Par ailleurs, la méthode ENN nécessite la configuration du paramètre k. Une étude sur la sensibilité vis-à-vis de ce paramètre est présentée.
2. Les méthodes basées sur les scores d'*outlier* du RF. Dans ces approches, le score d'*outlier* est non-binaire. Un critère de décision est donc nécessaire afin d'identifier les données mal étiquetées. Le critère de décision utilisé ici considère les n échantillons ayant les plus forts scores d'*outlier* comme des données mal étiquetées. Ces échantillons sont ensuite supprimés de l'ensemble des échantillons d'apprentissage. Pour le calcul des scores O_{RF}, les trois mesures de similarité – Breiman, DistanceLCA et PuretyLCA – étudiées au Chapitre 6 sont utilisées.
3. Les méthodes basées sur la combinaison des prédictions du RF. Dans cette approche, les trois vecteurs de probabilité d'appartenance aux classes (BP, RP1, et RP2) proposés à la Section 7.2.2 sont utilisés. Pour chaque échantillon, le critère de décision identifie comme mal étiquetés les échantillons pour lesquels la classe obtenant le maximum de probabilité est différente de celle fournie par la donnée de référence. De plus, deux types de traitement des données mal étiquetées sont étudiés : 1) la suppression, 2) le ré-étiquetage.

Dans la suite, les résultats des filtrages basés sur ces trois méthodes de détection sont détaillés. L'analyse des paramètres des méthodes, de leur critère de décision et des types de traitement appliqué aux données filtrées est réalisée.

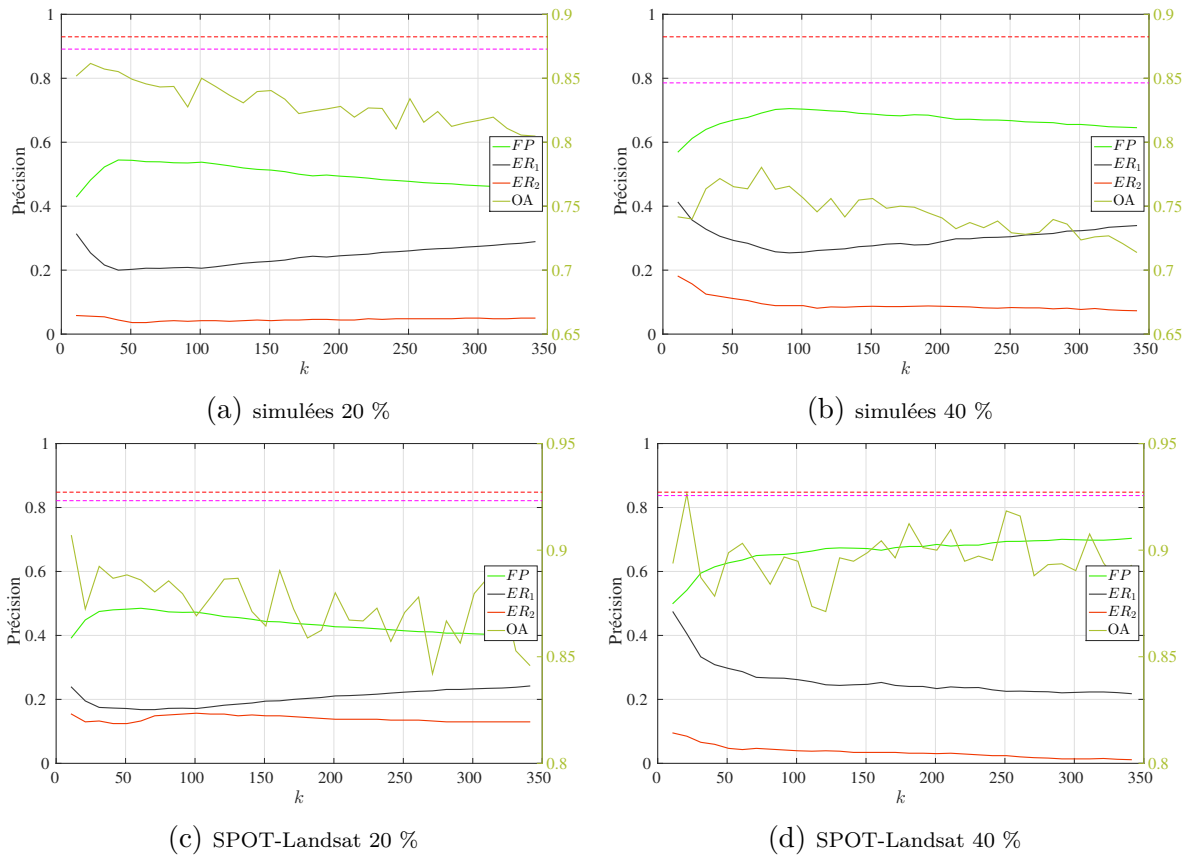
Basé sur les méthodes d'édition

Comme indiqué précédemment, le premier filtrage non-itératif étudié est basé sur l'utilisation de la méthode ENN présentée dans la Section 6.1.2. Cette méthode de détection nécessite la configuration du paramètre de voisinage k.

L'objectif de cette étude est d'évaluer l'apport du filtrage sur les deux critères d'évaluation décrits à la Section 7.3.1. Ces deux critères sont évalués pour différentes valeurs de k afin d'étudier la sensibilité de ce paramètre. Pour ce faire, des valeurs de k allant de 1 à 371 par pas de 10 sont testées.

La Figure 7.7 montre les résultats obtenus avec la méthode ENN pour différentes valeurs de k. Les résultats sont affichés pour les deux critères d'évaluation. L'axe des ordonnées de gauche montre les valeurs des métriques associées à la précision du filtrage, tandis que l'axe des ordonnées de droite montre les valeurs d'OA obtenues après l'étape de filtrage. Concernant les mesures de précision, la courbe en vert représente la mesure FP, en noir l'erreur ER_1 et en rouge l'erreur ER_2. La première ligne montre les résultats obtenus pour les données simulées, tandis que la seconde ligne montre les résultats obtenus pour les données SPOT-Landsat. La première colonne correspond à un niveau de bruit de 20 %, et la seconde colonne à un niveau de bruit de 40 %.

Sur cette Figure, deux lignes horizontales permettent de visualiser les valeurs d'OA obtenues dans les deux cas optimaux montrés dans le Tableau 7.3. En particulier, la ligne horizontale en pointillé rouge indique les performances de classification obtenues par le RF appris sur les échantillons d'apprentissage non-bruités (cas sans bruit), tandis que la ligne horizontale en pointillé magenta indique les performances de classification obtenues par le RF appris sur les données parfaitement filtrées.



(a) simulées 20 % (b) simulées 40 %

(c) SPOT-Landsat 20 % (d) SPOT-Landsat 40 %

FIGURE 7.7 – Valeurs d'*Overall Accuracy* (OA) et de précision de la méthode *Edited Nearest Neighbor* (ENN) en fonction du paramètre de voisinage k. Les données simulées et SPOT-Landsat avec 20 et 40 % de données mal étiquetées sont utilisées.

Concernant l'évaluation des performances de classification, la Figure 7.7 montre que le filtrage basé sur l'ENN permet d'améliorer les valeurs d'OA. Cette affirmation peut être corroborée en regardant les performances de classification obtenues en utilisant les échantillons d'apprentissage sans filtrage. Ces résultats, montrés dans le Tableau 7.3, sont rappelés ici :

— simulées 20 % : OA = 83,8 %,

— simulées 40 % : OA = 67,0 %,

— SPOT-Landsat 20 % : OA = 90,0 %,

— SPOT-Landsat 40 % : OA = 85,6 %.

Concernant la configuration du paramètre k, les résultats obtenus montrent que les valeurs d'OA sont stables pour différentes valeurs de k. Pour l'ensemble des jeux de données étudiés, les valeurs d'OA ont tendance a diminué lorsque la valeur de k augmente.

Concernant la précision du filtrage, les courbes en rouge montrent que l'erreur ER_2 reste très faible pour tous les jeux de données. Ainsi, la méthode ENN est capable d'identifier toutes les données mal étiquetées présentes dans les échantillons d'apprentissage. Néanmoins, les courbes de précision et d'erreur ER_1 (en vert et noir respectivement) montrent que la méthode ENN fait de la sur-détection avec de nombreux faux positifs. Ces résultats sont donc en accord avec les résultats observés au Chapitre 6.

Afin de mieux comprendre ces premiers résultats, les nombres d'échantillons par classe restants après l'étape de filtrage et utilisés pour l'apprentissage de l'algorithme de classification sont montrés dans le Tableau 7.6. Les résultats sont affichés pour différentes

valeurs du paramètre k sur les données simulées. Les quatre premières colonnes montrent le nombre d'échantillons lorsque le niveau de bruit est de 20 %, et les quatre dernières lorsque le niveau de bruit est de 40 %. Pour rappel, le nombre d'échantillons par classe avant le filtrage est de 500.

Tableau 7.6 – Nombre d'échantillons d'apprentissage par classe restant pour différentes valeurs de k après l'utilisation de la méthode *Edited Nearest Neighbor* (ENN) pour les données simulées.

	20 %				40 %			
k	1	31	61	91	1	31	61	91
Maïs	304	373	380	381	243	233	256	264
Maïs ensilage	244	304	303	301	199	238	208	198
Sorgho	233	265	266	259	131	153	165	179
Tournesol	250	298	292	285	199	226	239	252
Soja	233	355	365	377	232	283	311	315
Total	1264	1595	1606	1603	1004	1133	1179	1208

Le Tableau 7.6 montre que le filtrage basé sur la méthode ENN est agressif. En effet, quelque soit la valeur de k, environ 1400 échantillons sont supprimés pour un niveau de bruit de 20 %. Cependant, le nombre de données mal étiquetées est en réalité de 500 $(0,2 \times 2500)$. De manière similaire, environ 1200 échantillons sont supprimés pour un niveau de bruit de 40 %, alors que le nombre de données mal étiquetées est en réalité de 800 $(0,4 \times 2500)$. Par ailleurs, le nombre de données supprimées par classe n'est pas identique alors que le niveau de bruit est identique par classe.

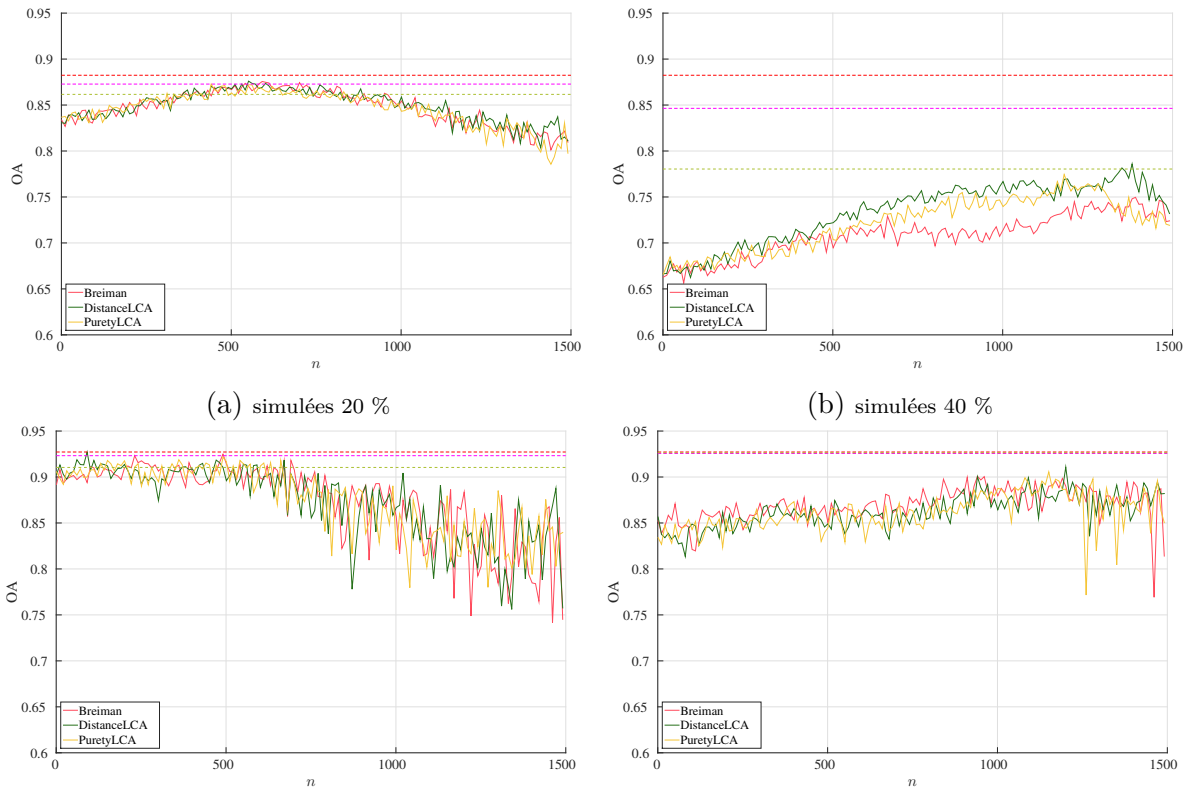
Basé sur le score d'*outlier* du *Random Forest*

Dans cette partie, le filtrage non-itératif étudié est basé sur l'utilisation des scores d'*outlier* du RF. Comme ces scores d'*outlier* sont non-binaires, un critère pour décider quelles sont les données mal étiquetées est nécessaire. Dans ce travail, le critère consiste à supprimer les n échantillons ayant les plus forts scores d'*outlier*.

Afin d'évaluer la sensibilité du filtrage vis-à-vis du paramètre n, les évaluations sont réalisées pour des valeurs de n allant de 1 à 1491 par pas de 10. Par ailleurs, l'utilisation des scores d'*outlier* du RF nécessite également l'apprentissage d'un modèle du RF. Pour ces expérimentations, la configuration est identique à celle du RF utilisé pour l'évaluation des performances de classification.

Les valeurs d'OA en fonction du paramètre n sont montrées dans la Figure 7.8. La première ligne montre les résultats obtenus pour les données simulées, tandis que la seconde ligne montre les résultats obtenus pour les données SPOT-Landsat. La première colonne correspond à un niveau de bruit de 20 %, tandis que la seconde colonne correspond à un niveau de bruit de 40 %. Chaque courbe représente une méthode : en rouge Breiman, en vert foncé DistanceLCA et en jaune PuretyLCA. Les lignes horizontales en pointillé rouge et magenta indiquent les valeurs d'OA du RF appris pour des données sans bruit et parfaitement filtrées respectivement. Une troisième ligne horizontale en pointillé vert clair indique le maximum d'OA obtenu par la méthode ENN.

Pour les différentes mesures de similarité, la Figure 7.8 montre qu'il existe une valeur optimale de n qui permette d'obtenir une valeur maximale d'OA. Plus spécifiquement, le maximum d'OA est obtenu lorsque le nombre d'échantillons supprimés n est proche du nombre réel de données mal étiquetées. Pour un niveau de bruit de 20 %, le nombre total d'échantillons mal étiquetés est de 500. Cette valeur est donc proche des maximums d'OA observés sur les Figures 7.8a 7.8c. Pour ces valeurs de n, l'étude sur la précision du



(a) simulées 20 % (b) simulées 40 %

(c) SPOT-Landsat 20 % (d) SPOT-Landsat 40 %

FIGURE 7.8 – Valeurs d'*Overall Accuracy* (OA) pour les méthodes Breiman, DistanceLCA et PuretyLCA en fonction du paramètre de seuil n. Les données simulées et SPOT-Landsat avec 20 et 40 % de données mal étiquetées sont utilisées.

filtrage permettra de déterminer si les données supprimées correspondent uniquement à des données mal étiquetées ou pas.

Concernant l'utilisation de différentes mesures de similarité, aucune différence significative peut être observée, exceptée à la Figure 7.8b. Dans ce cas là, les mesures de similarité proposées DistanceLCA et PuretyLCA ont un maximum d'OA plus élevé que pour la mesure de similarité de Breiman.

Les résultats entre les données simulées et SPOT-Landsat sont similaires. Pour un bruit de 20 %, le filtrage permet d'atteindre les performances optimales (ligne en pointillé magenta). Pour un bruit de 40 %, le filtrage améliore les performances de classification mais les performances optimales ne sont pas atteintes. Dans le cas des données simulées contaminées par 40 % de bruit, le gain du filtrage est plus important. Cependant, l'écart entre les maxima d'OA atteint et les valeurs d'OA optimales est aussi plus important. Les petites fluctuations observées sont dues à l'aléatoire introduit dans l'apprentissage de l'algorithme de classification. Les résultats montrés ici ne sont pas moyennés pour l'apprentissage de différents classifieurs.

La comparaison des valeurs maximales d'OA avec le filtrage basé sur la méthode ENN (ligne horizontale en pointillé vert) montre que les filtrages basés sur les scores d'*outlier* obtiennent globalement de meilleurs résultats lorsque le niveau de bruit est faible (Figures 7.8a 7.8c). En revanche, le filtrage basé sur la méthode ENN permet d'obtenir un OA plus élevé lorsque le niveau de bruit est de 40 % (Figures 7.8b et 7.8d [69]).

69. Dans la Figure 7.8d, la ligne horizontale en pointillé vert représentant le maximum d'OA obtenu par la méthode ENN est superposée à ligne horizontale en pointillé rouge.

En complément de la Figure 7.8, le Tableau 7.7 indique le nombre d'échantillons par classe disponibles après l'étape de filtrage pour différentes valeurs de n sur les données simulées. Ces échantillons sont utilisés pour l'apprentissage du modèle RF dont les performances sont montrées à la Figure 7.8. Les quatre premières colonnes montrent le nombre d'échantillons lorsque le niveau de bruit est de 20 %, et les quatre dernières lorsque le niveau de bruit est de 40 %.

TABLEAU 7.7 – Nombre d'échantillons d'apprentissage restant pour différentes valeurs de n après un filtrage basé sur les scores classiques (Breiman) d'*outliers* du *Random Forest* pour les données simulées.

	20 %				40 %			
n	501	801	1101	1401	501	801	1101	1401
Maïs	378	309	274	225	332	282	258	227
Maïs ensilage	404	333	273	208	390	335	262	197
Sorgho	389	348	282	212	430	352	290	226
Tournesol	435	362	287	231	405	353	287	225
Soja	393	347	283	223	442	377	302	224
Total	1999	1699	1399	1099	1999	1699	1399	1099

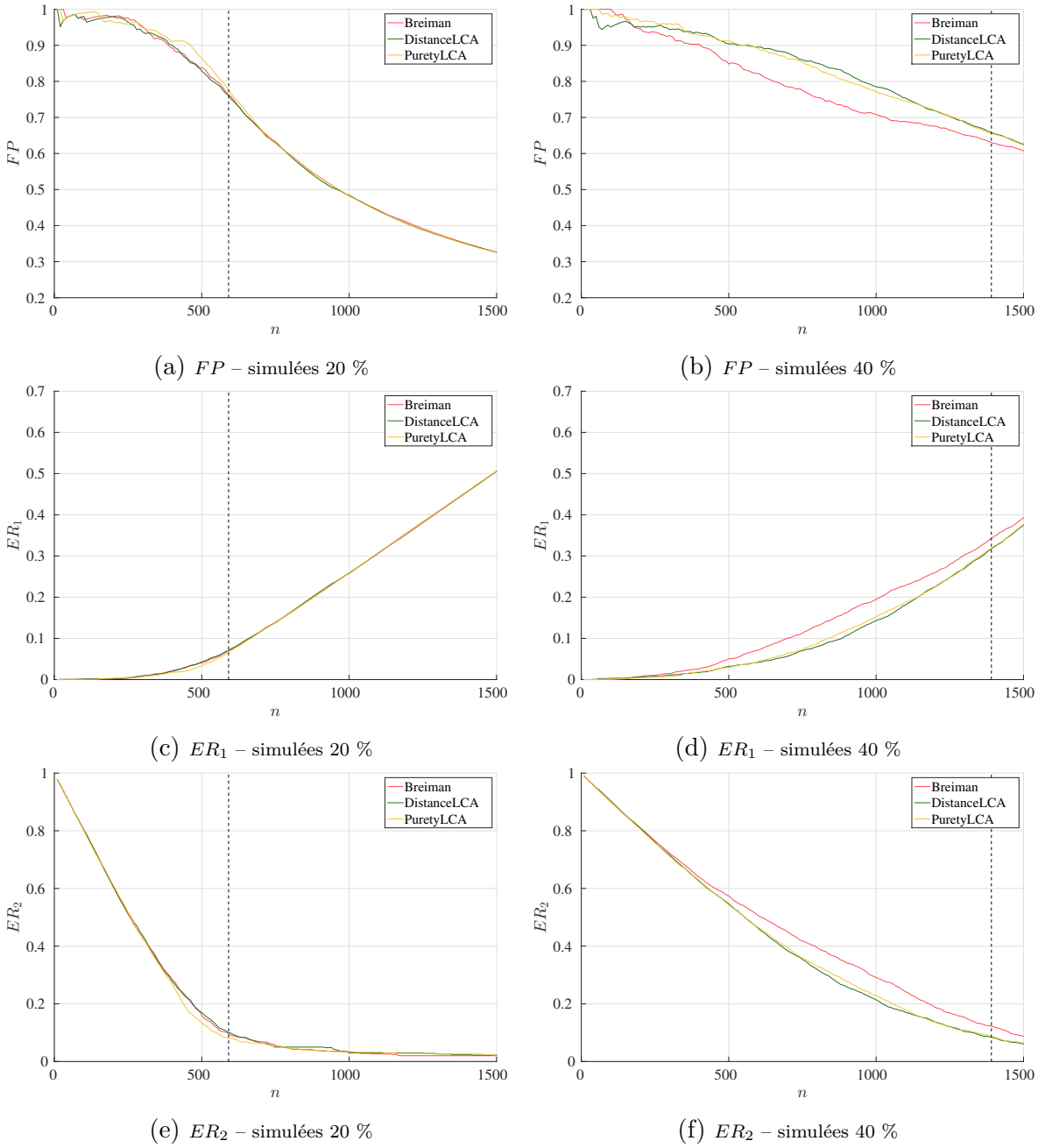
Le Tableau 7.7 montre que le nombre d'échantillons restant après l'étape de filtrage est équilibré entre les différentes classes. Pour un niveau de bruit de 20 %, le nombre d'échantillons est équitablement réparti pour les quatre valeurs de n étudiées. Pour un niveau de bruit de 40 %, le nombre d'échantillons est équilibré seulement pour les deux plus grandes valeurs de n (1101 et 1401). Pour des valeurs de n inférieures à 1101, le filtrage élimine d'abord les échantillons appartenant à la classe maïs. Au contraire, peu d'échantillons de soja sont supprimés. Ainsi, un léger déséquilibre est visible entre ces deux classes.

Les résultats du Tableau 7.7 ne permettent pas de savoir si les échantillons supprimés correspondent vraiment aux données mal étiquetées. Ainsi, la précision du filtrage est aussi étudiée. Les Figures 7.9 et 7.10 montrent ces résultats en fonction de la valeur du seuil n [70] pour les données simulées et SPOT-Landsat respectivement. Sur ces deux figures, la première ligne montre les valeurs de FP, la deuxième ligne de ER_1 et la troisième ligne de ER_2. La première colonne donne les résultats pour un niveau de bruit de 20 %, la seconde pour un niveau de bruit de 40 %. Chaque courbe représente une méthode : en rouge Breiman, en vert foncé DistanceLCA et en jaune PuretyLCA. Par ailleurs, la ligne verticale en pointillé noir indique la valeur n pour laquelle le maximum d'OA est obtenue pour la mesure de similarité Breiman (Figure 7.8). Comme cette valeur est similaire pour les mesures de similarité DistanceLCA et PuretyLCA, elle n'est pas affichée pour ces deux méthodes.

Les Figures 7.9 et 7.10 montrent des résultats similaires pour les différents jeux de données étudiées. Concernant la mesure de similarité utilisée, peu de différences sont visibles. Une seule différence est visible sur la Figure 7.9 lorsque le niveau de bruit est de 40 %. Dans ce cas, les mesures de similarité DistanceLCA et PuretyLCA sont plus précises et commettent moins d'erreur que la mesure de similarité de Breiman. Ce résultat peut expliquer les meilleures performances en classification observées à la Figure 7.8b pour ces deux mesures.

Les Figures 7.9 et 7.10 montrent que la valeur de n est cruciale et qu'elle dépend du

70. Dans ce cas là, la valeur $FP(n)$ correspond donc à $P@n$ la précision à n étudiée au Chapitre 6. De manière similaire, la valeur ER_2 est égale à $1 - \text{PA}$ avec PA le rappel.



(a) FP – simulées 20 %

(b) FP – simulées 40 %

(c) ER_1 – simulées 20 %

(d) ER_1 – simulées 40 %

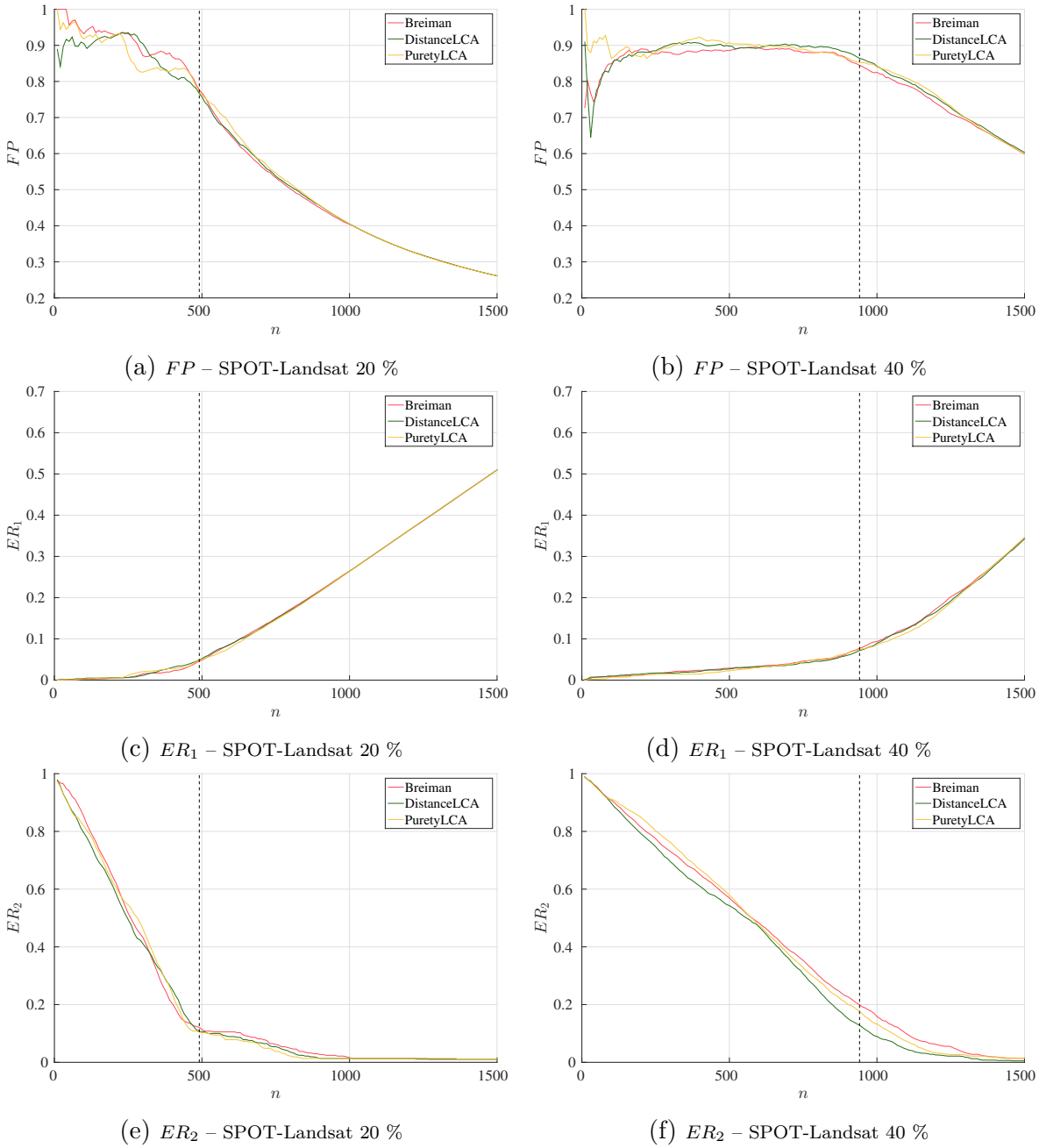
(e) ER_2 – simulées 20 %

(f) ER_2 – simulées 40 %

FIGURE 7.9 – Évolution de la précision du filtre (FP, ER_1 et ER_2) en fonction du paramètre de seuil n pour les méthodes Breiman, DistanceLCA, PuretyLCA. Les données simulées avec 20 et 40 % de données mal étiquetées sont utilisées.

niveau de bruit et des données étudiées. De manière générale, la précision FP (première ligne) est très élevée pour des petites valeurs de n. Une exception est visible sur la Figure 7.10b dans le cas des données SPOT-Landsat à un niveau de bruit de 40 %. Dans ce cas, des échantillons bien étiquetés sont supprimés pour des valeurs de n inférieures à 100 .

Par ailleurs, l'erreur ER_2 (troisième ligne) est très élevée pour des petites valeurs de n. Ces résultats sont attendus puisque les nombres de données mal étiquetées sont de 500 et 800 pour 20 et 40 % de bruit respectivement. Ainsi, la valeur de n est inférieure au nombre réel de données mal étiquetées. De manière complémentaire, l'erreur ER_1 (deuxième ligne) est quasiment nulle pour des petites valeurs de n. Dans ce cas, quasiment aucune donnée



(a) FP – SPOT-Landsat 20 %

(b) FP – SPOT-Landsat 40 %

(c) ER_1 – SPOT-Landsat 20 %

(d) ER_1 – SPOT-Landsat 40 %

(e) ER_2 – SPOT-Landsat 20 %

(f) ER_2 – SPOT-Landsat 40 %

FIGURE 7.10 – Évolution de la précision du filtre (FP, ER_1 et ER_2) en fonction du paramètre de seuil n pour les méthodes Breiman, DistanceLCA, PuretyLCA. Les données SPOT-Landsat avec 20 et 40 % de données mal étiquetées sont utilisées.

correctement étiquetée n'est identifiée comme mal étiquetée. Pour ces petites valeurs de n, le filtrage supprime donc uniquement des données mal étiquetées.

Les Figures 7.9 et 7.10 montrent également que le maximum d'OA (ligne horizontale en pointillé noir) est obtenu pour un compromis entre les trois métriques FP, ER_1 et ER_2. Plus précisément, le maximum est obtenu au moment où l'erreur ER_2 est quasiment nulle. Dans ce cas, tous les échantillons mal étiquetés sont quasiment supprimés. L'erreur ER_1 est aussi généralement très faible lorsque le maximum d'OA est atteint. Cependant, une exception existe pour les données simulées à 40 % de bruit (seconde colonne de la Figure 7.9). Dans ce cas, le maximum d'OA est atteint lorsque l'erreur ER_1 est supérieure à 0,3. Cela signifie que plus de 30 % de données correctement étiquetées ont été supprimées

lorsque le maximum d'OA a été atteint. La suppression de ces échantillons correctement étiquetés peut expliquer alors la différence observée entre le maximum d'OA atteint et les OA obtenus dans des cas optimaux (lignes horizontales magenta et rouge pour la Figure 7.8b).

Ces premiers résultats montrent qu'un filtrage basé sur le score d'*outlier* du RF permet d'améliorer les performances de la classification. Néanmoins, la configuration du paramètre n joue un rôle essentiel sur l'efficacité de ce filtrage. Par conséquent, les performances obtenues en classification après le filtrage dépendent directement de n.

Dans la réalité, les échantillons test non-bruités utilisés pour calculer l'OA ne sont pas disponibles. Ainsi, il n'est pas possible de fixer la valeur de n en s'appuyant sur les valeurs d'OA montrées à la Figure 7.8.

Théoriquement, les valeurs des scores d'*outlier* des données mal étiquetées sont plus fortes que celles des données correctement étiquetées. De plus, les scores d'*outlier* des échantillons correctement étiquetés doivent être idéalement similaires et quasiment nuls. En notant $O_{RF}(n)$ la valeur du score d'*outlier* du n-ième échantillon supprimé, une solution est de fixer la valeur de n lorsque $O_{RF}(n-1) - O_{RF}(n) \simeq 0$. Idéalement, le maximum d'OA doit être atteint pour cette valeur de n.

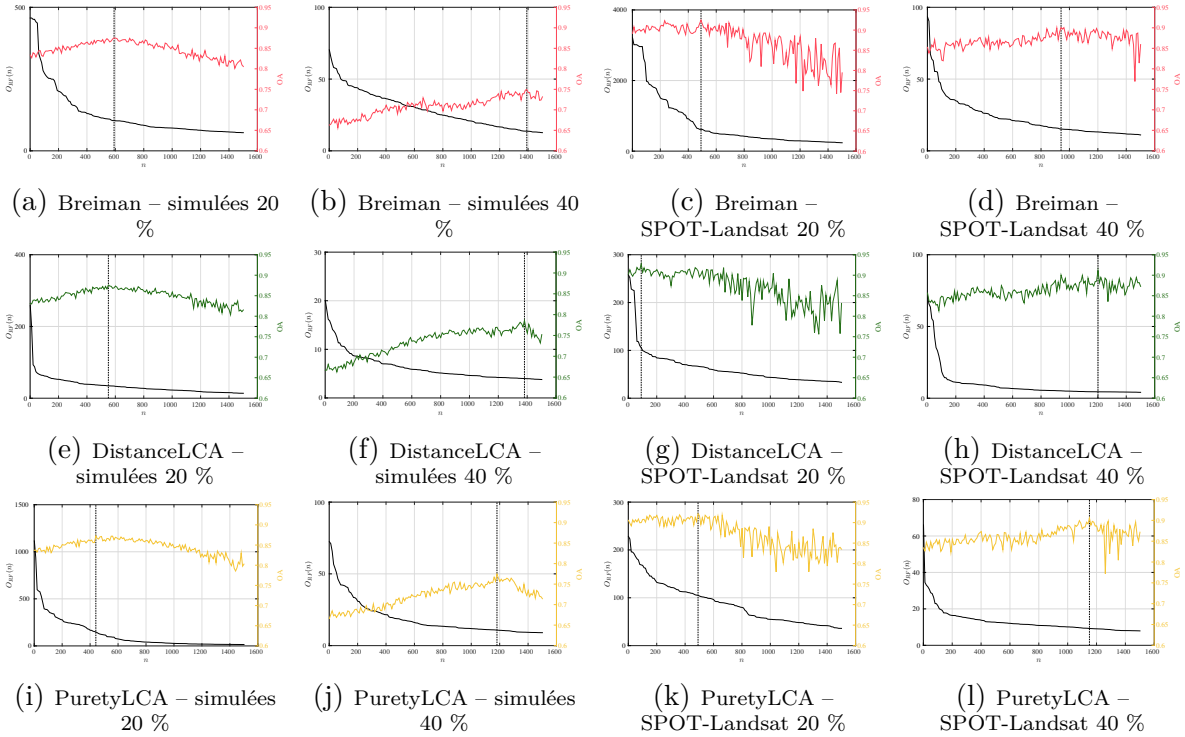
Pour étudier cette hypothèse, la Figure 7.11 montre le score d'*outlier* du n-ième échantillon supprimé et les valeurs d'OA pour différentes valeurs de n. La valeur $O_{RF}(n)$ est montrée sur l'axe des ordonnées de gauche. À noter que la plage des valeurs observées pour $O_{RF}(n)$ est spécifique pour chaque figure. Les valeurs de l'OA en fonction de n sont rappelées sur l'axe des ordonnées de droite. Chaque ligne représente une mesure de similarité : Breiman, DistanceLCA et PuretyLCA. Chaque colonne représente un jeu de données : données simulées avec 20 % de données mal étiquetées, données simulées 40 %, données SPOT-Landsat 20 % et données SPOT-Landsat 40 %. Par ailleurs, la ligne verticale en pointillé noir indique la valeur du seuil n pour laquelle le maximum d'OA est obtenue.

Les Figures 7.11a, 7.11b, 7.11c, 7.11g et 7.11i représentent des cas plus ou moins idéaux. Sur ces figures, le maximum d'OA est atteint au moment où les valeurs $O_{RF}(n)$ se stabilisent. Dans un cas applicatif, la valeur de n pourrait donc être fixée en observant les valeurs de score $O_{RF}(n)$ pour différentes valeurs de n. Dans ce cas, la valeur de n choisie correspondrait au moment où les valeurs $O_{RF}(n)$ deviendraient stables et quasiment nulles. Cependant, pour d'autres cas de la Figure 7.11, notamment la dernière colonne, ce critère semble moins évident à utiliser.

Basé sur la combinaison des prédictions des arbres du *Random Forest*

Le dernier filtrage étudié ici utilise de manière non-itérative les stratégies décrites dans la Section 7.2.2. Pour ce faire, l'algorithme du RF est appris en utilisant tous les échantillons contenant les données mal étiquetées à détecter. Les paramètres du RF sont configurés de la même manière que pour le calcul des scores d'*outlier* : $K = 100$, $m = \sqrt{p}$, $max\_depth = 25$ et $min\_samples = 10$.

L'ensemble des arbres construit est ensuite utilisé pour calculer les différents vecteurs de probabilité d'appartenance aux classes présentés à la Section 7.2.2. Si le maximum de probabilité est obtenu pour une classe autre que celle de la classe fournie par la donnée de référence, alors l'échantillon est identifié comme étant mal étiqueté. Grâce au calcul du vecteur de probabilité, deux types de traitements sont étudiés : la suppression et le ré-étiquetage. Pour l'évaluation de ces méthodes, les données et les critères d'évaluation sont toujours identiques à ceux utilisés précédemment. Ainsi, l'algorithme de classification



(a) Breiman – simulées 20 %
(b) Breiman – simulées 40 %
(c) Breiman – SPOT-Landsat 20 %
(d) Breiman – SPOT-Landsat 40 %
(e) DistanceLCA – simulées 20 %
(f) DistanceLCA – simulées 40 %
(g) DistanceLCA – SPOT-Landsat 20 %
(h) DistanceLCA – SPOT-Landsat 40 %
(i) PuretyLCA – simulées 20 %
(j) PuretyLCA – simulées 40 %
(k) PuretyLCA – SPOT-Landsat 20 %
(l) PuretyLCA – SPOT-Landsat 40 %

FIGURE 7.11 – Évolution de l'*Overall Accuracy* (OA) et des scores d'*outlier* en fonction du paramètre de seuil n pour les méthodes Breiman, DistanceLCA, PuretyLCA. Les données simulées et SPOT-Landsat avec 20 et 40 % de données mal étiquetées sont utilisées.

utilisé après l'étape de filtrage est identique pour les différentes stratégies de filtrage évaluées dans cette partie.

Les valeurs d'OA obtenues pour les méthodes basées sur la combinaison des prédictions sont montrées dans le Tableau 7.8. De plus, les résultats des autres filtrages non-itératifs étudiés précédemment sont aussi indiqués dans le Tableau 7.8. Pour le filtrage ENN et les filtrages basés sur les scores d'*outlier* du RF, les valeurs d'OA montrées sont obtenues pour une configuration optimale des paramètres k et n. De plus, les valeurs d'OA, des cas sans filtrage et filtrage idéal présentées dans le Tableau 7.3, obtenues par le RF sont aussi affichées sur la troisième et quatrième ligne respectivement. Les valeurs en gras indiquent les meilleures valeurs d'OA pour chaque jeu de données.

Pour les filtrages basés sur la combinaison des prédictions du RF, certaines méthodes n'identifient aucun échantillon comme étant mal étiqueté. Dans ces cas là, un tiret indique qu'aucun résultat n'est disponible.

Le Tableau 7.8 montre que les filtrages basés sur la méthode ENN et les scores d'*outlier* calculés par le RF obtiennent les meilleurs résultats. Plus précisément, l'utilisation des méthodes ENN et DistanceLCA conduit à de très bons résultats, parfois même au-dessus des valeurs du filtrage idéal. Pour le cas des données simulées à 40 %, cas où les données mal étiquetées sont les plus nuisibles, un gain d'OA supérieur à 11 % peut être observé. Néanmoins, ces performances sont encore loin de celles obtenues avec un classifieur appris sur les données parfaitement filtrées.

Concernant les filtrages basés sur une méthode de détection combinant les prédictions des arbres du RF, les performances obtenues sont bien en-deçà des autres méthodes analysées. En particulier, les vecteurs de probabilité BP et RP1 échouent à améliorer les performances de classification. De plus, l'utilisation de ces deux vecteurs ne permet pas la détection de données mal étiquetées dans le cas des données SPOT-Landsat. Le troisième

Tableau 7.8 – Valeurs de l'*Overall Accuracy* (OA) obtenues pour les filtrages non-itératifs basés sur les méthodes d'édition, les scores d'*outlier* du *Random Forest* (RF) et les prédictions des arbres du RF. Les données utilisées sont les données simulées et SPOT-Landsat avec 20 et 40 % de bruit. Les valeurs en gras indiquent les meilleures valeurs d'OA.

		Simulées		**SPOT-Landsat**	
		20 %	40 %	20 %	40 %
OA sans filtrage		83,8	67,0	90,0	85,6
OA filtrage idéal		87,3	84,6	92,3	92,6
Méthodes d'édition	ENN	86,2	78,0	91,0	**92,6**
Score d'*outlier* O_{RF}	O_{RF}^{Breiman}	**87,6**	75,0	92,5	90,1
	$O_{RF}^{\text{DistanceLCA}}$	**87,6**	**78,6**	**92,7**	91,1
	$O_{RF}^{\text{PuretyLCA}}$	87,3	77,4	92,2	90,6
Prédictions O_{RF} : suppression	$p_{C_i}^{\text{BP}}(x)$	84,3	66,8	-	-
	$p_{C_i}^{\text{RP1}}(x)$	83,8	66,8	-	-
	$p_{C_i}^{\text{RP2}}(x)$	84,7	68,4	90,1	85,6
Prédictions du RF : ré-étiquetage	$p_{C_i}^{\text{BP}}(x)$	83,7	67,3	-	-
	$p_{C_i}^{\text{RP1}}(x)$	83,3	66,4	-	-
	$p_{C_i}^{\text{RP2}}(x)$	83,0	68,6	91,0	86,0

calcul de vecteur de probabilité RP2 permet une légère amélioration de l'OA dans le cas du ré-étiquetage des échantillons mal étiquetés. Avec ces résultats, la comparaison entre les deux types de traitement – suppression ou ré-étiquetage – est difficile.

En conclusion, la suppression des échantillons identifiés comme mal étiquetés par la méthode ENN ou les méthodes basées sur les scores d'*outlier* calculés par le RF sont prometteuses. Néanmoins, les méthodes de détection de données mal étiquetées Breiman, DistanceLCA et PuretyLCA nécessitent la configuration non-triviale du seuil n utilisé comme critère de décision. Ainsi, l'utilisation de filtrages itératifs décrits à la Section 7.2 peut répondre à cette problématique.

7.4.2 Étude de filtrages itératifs

Le principe du filtrage itératif est d'appliquer une méthode de détection de données mal étiquetées de manière itérative. Dans cette partie, trois filtrages itératifs sont étudiés en utilisant les trois méthodes de détection de la Section 7.4.1. Pour ces méthodes, la configuration d'un critère pour arrêter le filtrage est nécessaire.

Comme dans la Section 7.4.1, l'évaluation des filtrages itératifs est réalisée en calculant la précision du filtrage et les performances de classification obtenues en utilisant les données filtrées. Les différentes évaluations sont ici réalisées à chaque itération. Afin d'évaluer le gain en classification, le RF décrit à la Section 7.3.1 est utilisé comme algorithme de classification.

Les résultats de classification seront comparés à trois résultats de référence : 1) ceux obtenus en utilisant des données non-bruitées, et 2) ceux obtenus en utilisant des données parfaitement filtrées, et 3) ceux obtenus en utilisant les données bruitées. Ces résultats de référence sont présentés dans le Tableau 7.3 à la Section 7.3.1. Par ailleurs, les expéri-

mentations sont réalisées sur les données simulées et SPOT-Landsat avec des niveaux de bruit de 20 et 40 %.

Dans la suite, les résultats des trois stratégies de filtrage itératif sont détaillés. Pour chaque filtrage, l'analyse des paramètres des méthodes, des critères de décision et d'arrêt, et des types de traitement appliqué aux données filtrées est réalisée.

Basé sur les méthodes d'édition

Le premier filtrage itératif étudié repose sur l'utilisation de la méthode d'édition ENN. Dans la littérature, plusieurs variantes itératives de la méthode ENN existent dont la méthode RENN. Dans cette dernière, la méthode ENN est utilisée à chaque itération pour détecter les données à supprimer. Comme dans la version ENN non-itérative, la configuration du paramètre k est donc aussi nécessaire pour la méthode RENN. Le filtrage itératif RENN s'arrête automatiquement lorsque plus aucune donnée mal étiquetée n'est identifiée.

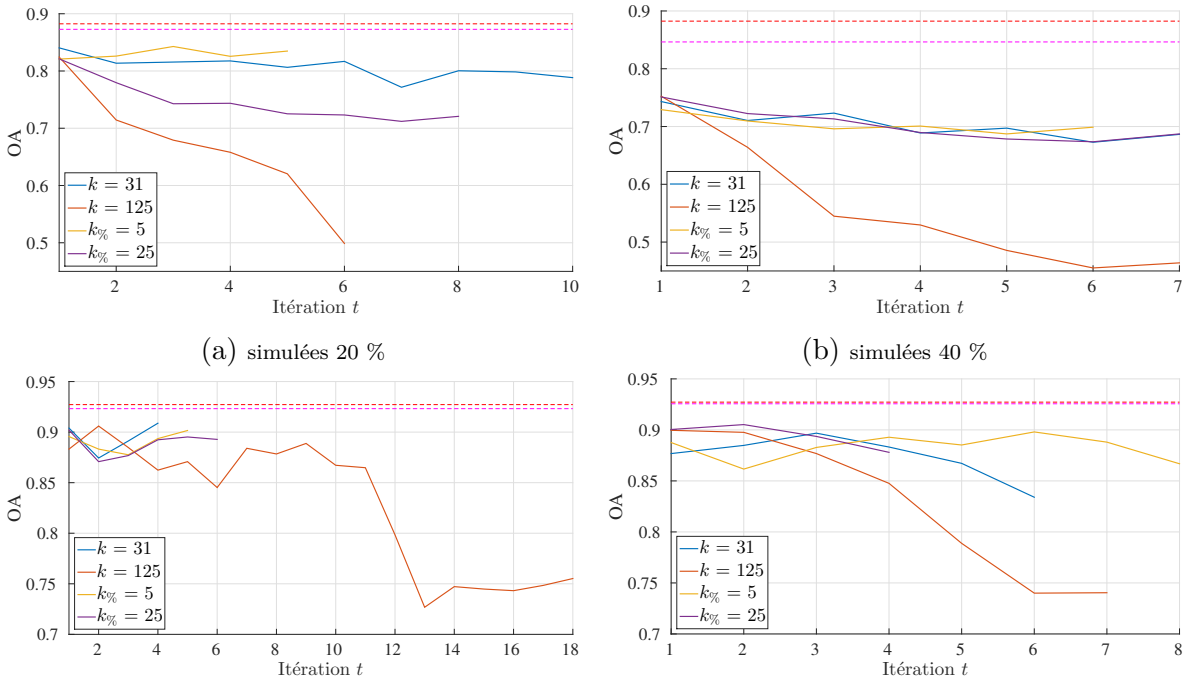
Dans la Section 7.4.1, il a été montré que la configuration optimale de k dépendait du niveau de bruit présent dans les données. Ainsi, une limitation importante de la méthode RENN est d'utiliser une valeur de k identique pour chaque classe alors que dans un cas réel le niveau de bruit n'est pas nécessairement identique pour toutes les classes. Par ailleurs, le Tableau 7.6 montre que le nombre d'échantillons supprimé pour différentes valeurs de k par la méthode ENN varie entre les classes. La taille du voisinage considérée peut alors être trop importante pour certaine classe par rapport au nombre d'échantillons présent dans cette classe. La valeur du paramètre k est alors souvent inadaptée après plusieurs itérations.

Pour pallier à cet inconvénient, un nouveau filtrage itératif basé sur l'utilisation de la méthode ENN est ici proposé. La principale nouveauté repose sur la définition d'un nouveau paramètre k nommé $k_{c_i}^t$ spécifique à la classe c_i et mis à jour à chaque itération t. Plus spécifiquement, la valeur de $k_{c_i}^t$ correspond au pourcentage $k_{\%}$ des échantillons appartenant à la classe c_i pour l'itération t. Par exemple, considérons la classe c_i composée de 500 échantillons. Si la méthode ENN est appliquée avec une valeur k fixée à 125, le score d'*outlier* d'un échantillon de la classe c_i est évalué en considérant ses 125 plus proches voisins pour toutes les itérations. En revanche, en considérant le filtrage proposé avec un pourcentage $k_{\%}$ fixé à 25 %, la valeur $k_{c_i}^1$ pour la première itération est égal à 125. Si 100 échantillons pour la classe c_i sont éliminés à la première itération, alors la valeur $k_{c_i}^2$ pour la deuxième itération est égal à 100. Pour ce nouveau filtrage, le seul paramètre nécessaire est donc le pourcentage $k_{\%}$.

La Figure 7.7 de la Section 7.4.1 a mis en évidence que la précision du filtrage ENN et les performances de classification obtenues étaient peu influencées par la configuration de k. Ainsi, les études sont ici menées pour seulement quelques valeurs de k et $k_{\%}$.

La Figure 7.12 montre les performances de classification pour chaque itération t. La première ligne montre les résultats obtenus pour les données simulées, tandis que la seconde ligne montre les résultats obtenus pour les données SPOT-Landsat. La première colonne correspond à un niveau de bruit de 20 %, et la seconde colonne à un niveau de bruit de 40 %. Les courbes bleue et rouge montrent les résultats obtenus pour une valeur k fixée à 31 et 125 respectivement. Les courbes jaune et violette montrent les résultats obtenus pour une valeur $k_{\%}$ égale à 5 et 25 % respectivement. Par ailleurs, les lignes horizontales en pointillé rouge et magenta indiquent les valeurs d'OA obtenues par le RF appris sur les données sans bruit et parfaitement filtrées respectivement (Tableau 7.3).

Sur la Figure 7.12, les courbes sont de longueurs différentes puisque chaque filtrage



(a) simulées 20 % (b) simulées 40 %

(c) SPOT-Landsat 20 % (d) SPOT-Landsat 40 %

FIGURE 7.12 – Évolution de l'*Overall Accuracy* (OA) au cours des itérations pour différents processus itératifs basés sur les méthodes d'édition.

s'arrête automatiquement. Pour quasiment toutes les configurations, le nombre d'itérations nécessaires est inférieur à 10.

Les résultats observés pour les deux filtrages itératifs étudiés – RENN ou la variante proposée – sont similaires sur les premières itérations. Cependant, l'utilisation de la méthode RENN avec une valeur élevée de k (*i.e.* $k = 25$) montre qu'une valeur fixée de k fait chuter les valeurs d'OA après quelques itérations. Dans ce cas là, la valeur élevée de k conduit à la suppression de certaines classes. Ce problème n'est pas rencontré avec l'utilisation d'une valeur adaptative $k_{\%}$. Par contre, une différence est observée entre les deux valeurs de $k_{\%}$ utilisées. Pour ces jeux de données, une valeur élevée $k_{\%} = 25$ conduit à de moins bonnes performances qu'une valeur plus petite de 5 %.

Les résultats des deux filtrages itératifs peuvent être comparés aux résultats du filtrage non-itératifs ENN qui correspondent aux résultats obtenus à la première itération. Comme les maximums d'OA sont généralement obtenus dès la première itération dans la Figure 7.12, l'utilisation de la méthode ENN de manière itérative n'est pas bénéfique. Ces résultats étaient attendus puisque la Section 7.4.1 a montré que la méthode ENN identifie quasiment la totalité des données mal étiquetées (erreur ER_2 faible sur la Figure 7.7) en privilégiant la sur-détection (erreur ER_1 élevée). Si les données mal étiquetées sont presque toutes supprimées dès la première itération, alors les itérations suivantes sont peu utiles.

Basé sur le score d'*outlier* du *Random Forest*

Cette partie s'intéresse aux deux processus itératifs basés sur le score d'*outlier* du RF présentés dans la Section 7.2.2. Les caractéristiques de ces deux filtrages sont rappelées dans le Tableau 7.1.

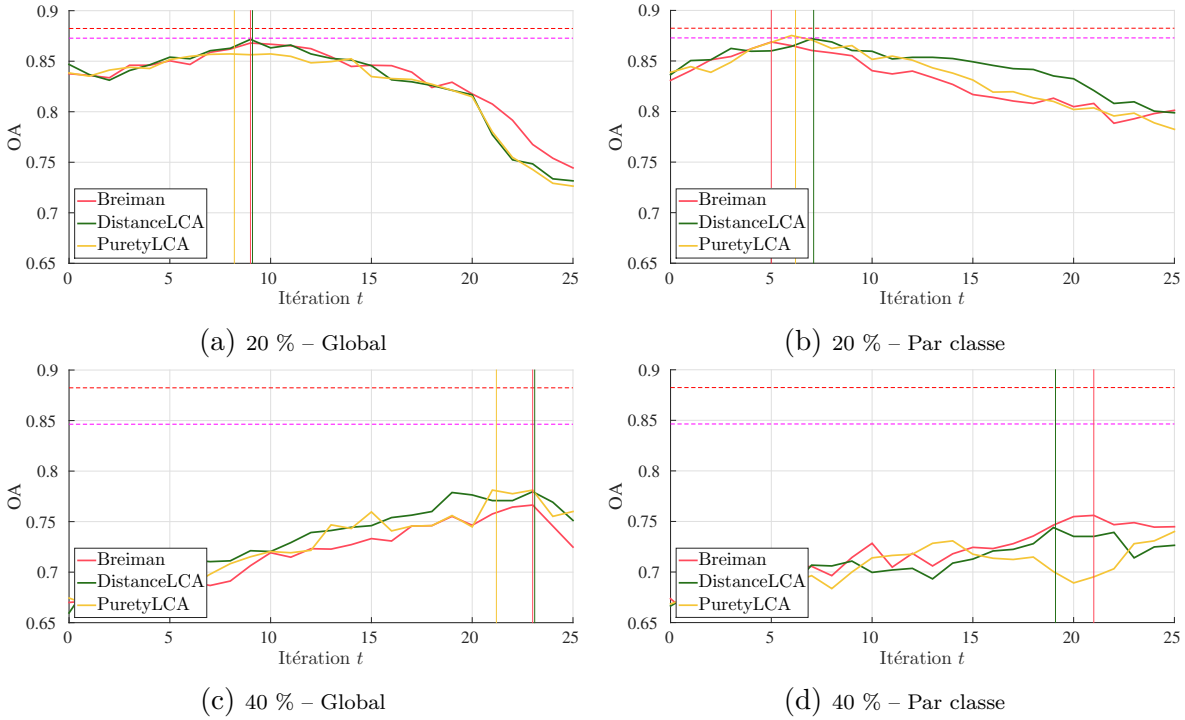
Dans le filtrage itératif global, les n échantillons ayant les plus forts scores d'*outlier* sont supprimés à chaque itération. La valeur du seuil n est ici fixée à 50 grâce à une étude

de sensibilité qui n'est pas présentée. Cette valeur permet d'obtenir un bon compromis entre la précision du filtrage et le nombre d'itérations nécessaires pour éliminer les données mal étiquetées. Le Chapitre 6 a notamment montré les faibles différences entre la précision à 10 et 50 (Tableaux 6.4 et 6.5). Comme le processus ne s'arrête pas automatiquement, un critère d'arrêt basé sur la valeur des scores d'*outlier* a été défini.

Dans le filtrage itératif par classe, les échantillons supprimés sont ceux qui ne respectent pas la règle des 3σ. Comme expliqué à la Section 7.2.2, ce filtrage itératif s'arrête automatiquement lorsque l'ensemble des échantillons respectent la règle des 3σ.

Les filtrages itératifs par classe et global nécessitent le calcul des scores d'*outlier* O_{RF} à chaque itération. La configuration du RF utilisée est identique pour chaque itération à celle de la Section 7.4.1.

La première évaluation mesure l'apport des filtrages global et par classe sur les performances de classification. L'algorithme de classification utilisé après l'étape de filtrage est toujours le RF présenté à la Section 7.3.1. La Figure 7.13 montre l'évolution des valeurs d'OA au cours des itérations pour les deux processus itératifs sur les données simulées. Les résultats sont affichés pour les 25 premières itérations car dans cette première étude le critère d'arrêt n'est pas évalué. La première ligne correspond à un niveau de bruit de 20 % , tandis que la seconde ligne correspond à un niveau de bruit de 40 %. La première colonne correspond au filtrage itératif global, et la seconde colonne au filtrage itératif par classe. Chaque courbe représente une mesure de similarité : en rouge pour Breiman, en vert pour DistanceLCA, et en jaune pour PuretyLCA. Par ailleurs, les lignes horizontales en pointillé rouge et magenta indiquent les valeurs d'OA obtenues par le RF appris sur les données sans bruit et parfaitement filtrées respectivement. Les lignes verticales en pointillé correspondent à la position des maximums d'OA observés pour chaque mesure de similarité. La couleur des lignes verticales correspond à la mesure de similarité évaluée.



(a) 20 % – Global (b) 20 % – Par classe

(c) 40 % – Global (d) 40 % – Par classe

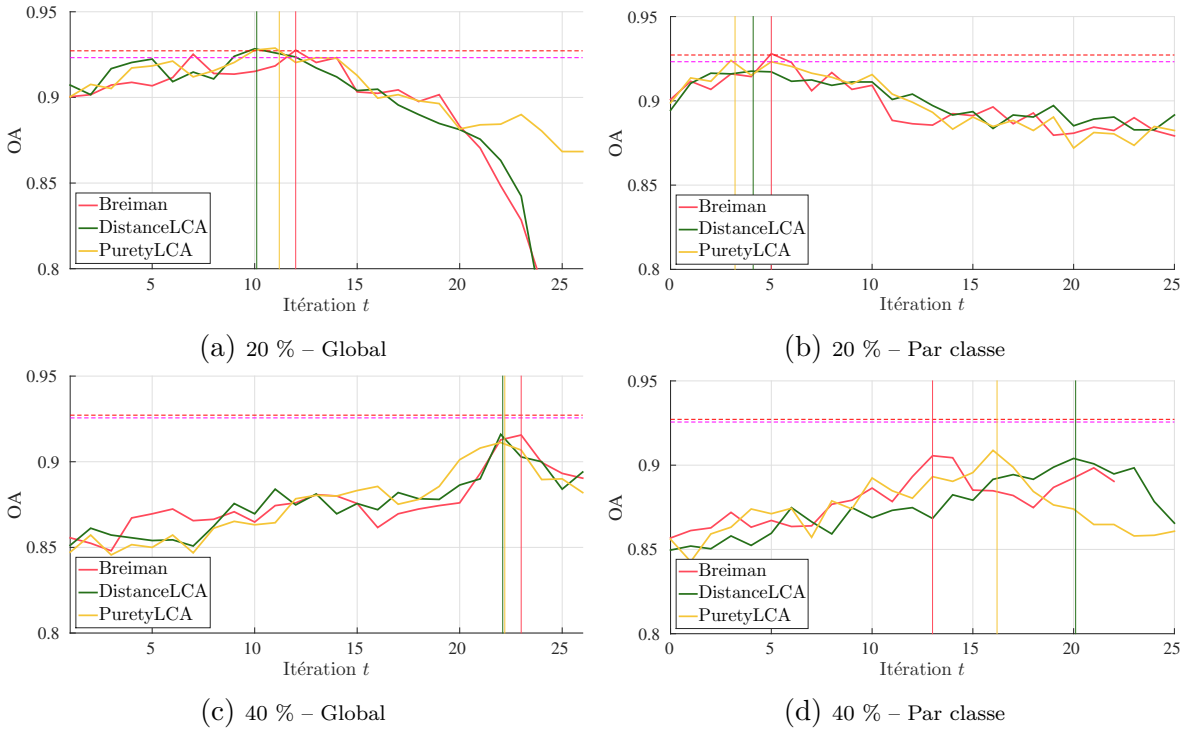
FIGURE 7.13 – Évolution de l'*Overall Accuracy* (OA) au cours des itérations pour les processus itératifs basés sur les scores O_{RF} pour les données simulées.

Pour toutes les configurations, la Figure 7.13 montre que l'étape de filtrage a une incidence positive sur les performances de la classification. Cette affirmation peut être

vérifiée en regardant les valeurs d'OA à l'itération $t = 0$, qui correspondent aux valeurs d'OA obtenues par le RF appris sur les données bruitées [71]. Lorsque le niveau de bruit est de 20 %, l'utilisation des filtrages itératifs permet même d'atteindre les performances qui seraient obtenues avec un filtrage idéal (droite horizontale en pointillé magenta). Lorsque le niveau de bruit est de 40 %, les valeurs d'OA augmentent de plus 10 %. Néanmoins, les résultats des Figures 7.13c et 7.13d montrent que les filtrages proposés peuvent être améliorés puisque les valeurs d'OA obtenues avec les filtrages idéaux (droites horizontales en pointillé rouge et magenta) ne sont pas atteintes.

Concernant le choix de la mesure de similarité, les maximums d'OA obtenus sont très similaires entre les mesures Breiman, DistanceLCA et PuretyLCA. Par ailleurs, la comparaison des résultats entre les deux colonnes de la Figure 7.13 montre que peu de différences sont visibles entre le critère de décision global et par classe : les résultats sont similaires. Néanmoins, le filtrage global permet d'obtenir des valeurs d'OA légèrement plus élevées que le filtrage par classe pour un niveaux de bruit de 40 %.

La même évaluation est réalisée pour les données SPOT-Landsat. En suivant le même style que précédemment, la Figure 7.14 montre l'évolution des valeurs d'OA au cours des itérations. Les résultats sont similaires à ceux obtenus à la Figure 7.13. Pour ce jeux de données, les performances optimales sont quasiment atteintes. De plus, le critère de décision global obtient des maximums d'OA légèrement plus élevés pour un niveau de bruit à 40 %.



(a) 20 % – Global (b) 20 % – Par classe

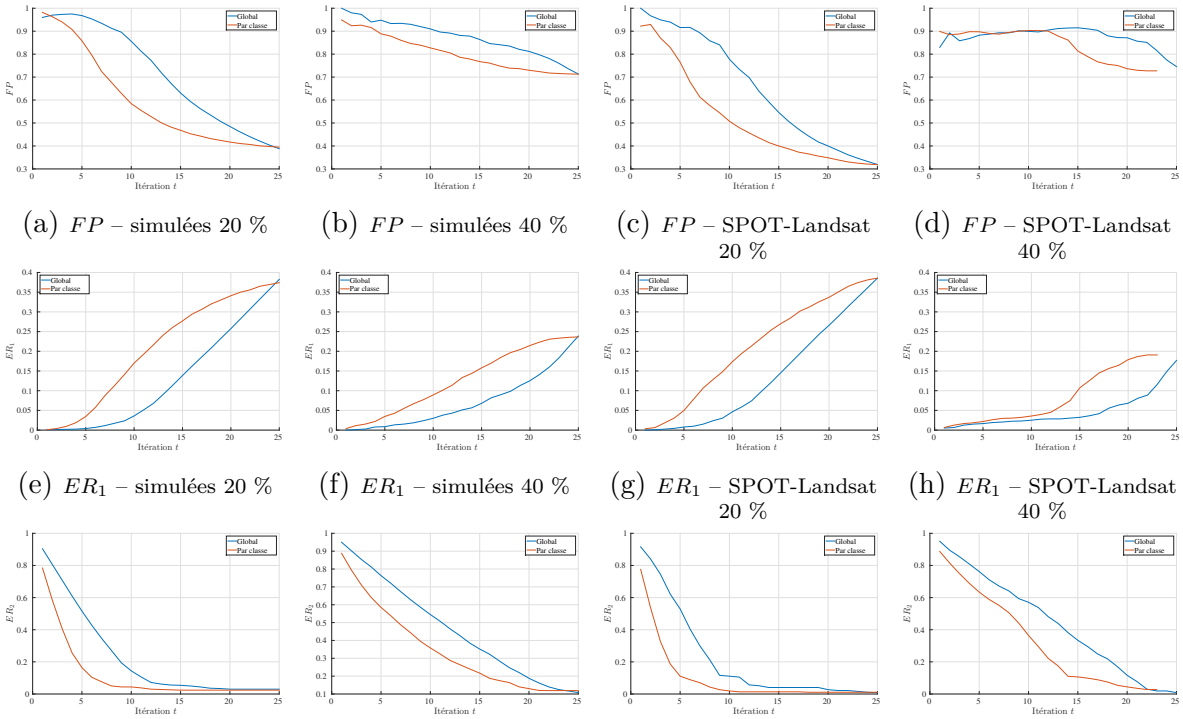
(c) 40 % – Global (d) 40 % – Par classe

FIGURE 7.14 – Évolution de l'*Overall Accuracy* (OA) au cours des itérations pour les processus itératifs basés sur les scores O_{RF} pour les données SPOT-Landsat.

Afin de comparer plus précisément les filtrages global et par classe, leur précision est maintenant évaluée. Pour ce faire, les trois mesures FP, ER_1, et ER_2 sont utilisées. Comme les trois mesures de similarité étudiées ont des performances comparables, les

71. Les différences d'OA visibles entre les trois méthodes (Breiman, DistanceLCA, et PuretyLCA) à l'itération $t = 0$ sont dues à l'utilisation de l'aléatoire (échantillons *bootstrap* et principe du *random feature selection*) lors de la construction des arbres du RF.

résultats sont montrés uniquement pour Breiman. Pour les deux filtrages, la Figure 7.15 montre les valeurs des métriques à chaque itération. Chaque courbe correspond à un filtrage : global en bleu, par classe en rouge. Plus précisément, la première ligne montre les valeurs de FP, la deuxième ligne de ER_1 et la troisième ligne de ER_2. Chaque colonne représente un jeu de données. Les deux premières colonnes montrent les résultats obtenus pour les données simulées pour un niveau de bruit de 20 et 40 %. Les deux dernières colonnes montrent les résultats obtenus pour les données SPOT-Landsat pour un niveau de bruit de 20 et 40 %.



(a) FP – simulées 20 % (b) FP – simulées 40 % (c) FP – SPOT-Landsat 20 % (d) FP – SPOT-Landsat 40 %

(e) ER_1 – simulées 20 % (f) ER_1 – simulées 40 % (g) ER_1 – SPOT-Landsat 20 % (h) ER_1 – SPOT-Landsat 40 %

(i) ER_2 – simulées 20 % (j) ER_2 – simulées 40 % (k) ER_2 – SPOT-Landsat 20 % (l) ER_2 – SPOT-Landsat 40 %

FIGURE 7.15 – Évolution de la précision du filtre (FP, ER_1 et ER_2) en fonction des itérations t pour les deux processus itératifs global et par classe. Les données simulées et SPOT-Landsat avec 20 et 40 % de données mal étiquetées sont utilisées.

Malgré des valeurs d'OA similaires, la Figure 7.15 montre que les deux filtrages proposés ont des comportements différents. Le filtrage par classe privilégie la sur-détection des données mal étiquetés : l'erreur ER_2 est faible, tandis que l'erreur ER_1 est élevée. Au contraire, le filtrage global a une erreur ER_2 plus élevée que celle du filtrage par classe. Il détecte donc moins de données mal étiquetées. Néanmoins, sa valeur de FP est plus élevée car il identifie moins de faux positifs. Ainsi, le choix entre le filtrage global ou par classe dépend de l'application. Si la suppression d'un grand nombre de données n'est pas importante, le filtrage par classe est intéressant car il nécessite moins de d'itérations. En revanche, le filtrage global peut être privilégié lorsque le nombre d'échantillons d'apprentissage est limité. Il permettra d'éviter la suppression d'un trop grand nombre d'échantillons.

Afin de continuer la comparaison des filtrages global et par classe, la possibilité de configurer facilement leur critère d'arrêt doit être étudiée. Dans les Figures 7.13 et 7.14, les filtrages devraient idéalement s'arrêter lorsque le maximum d'OA est atteint. Comme mentionné précédemment, un critère d'arrêt basé sur les valeurs d'OA n'est pas possible dans une application réelle. En effet, les échantillons test non-bruités utilisés pour calculer

les valeurs d'OA ne sont pas disponibles. Le critère d'arrêt doit donc être indépendant des performances de classification.

Pour le filtrage global, le critère d'arrêt étudié repose sur la valeur $O^t_{RF}(n)$, présentée à la Section 7.2.2. Cette valeur correspond au score d'*outlier* du n-ième échantillon supprimé à chaque itération t. Pour le filtrage par classe, le processus s'arrête automatiquement lorsque tous les échantillons respectent la règle des 3σ. Les Figures 7.13b et 7.14b montrent que les maximum d'OA sont atteints vers la cinquième itération. Pourtant, le filtrage par classe ne s'arrête pas pendant les vingt-cinq premières itérations. Afin de s'assurer des bonnes performances du filtrage par classe, un critère d'arrêt doit alors être défini.

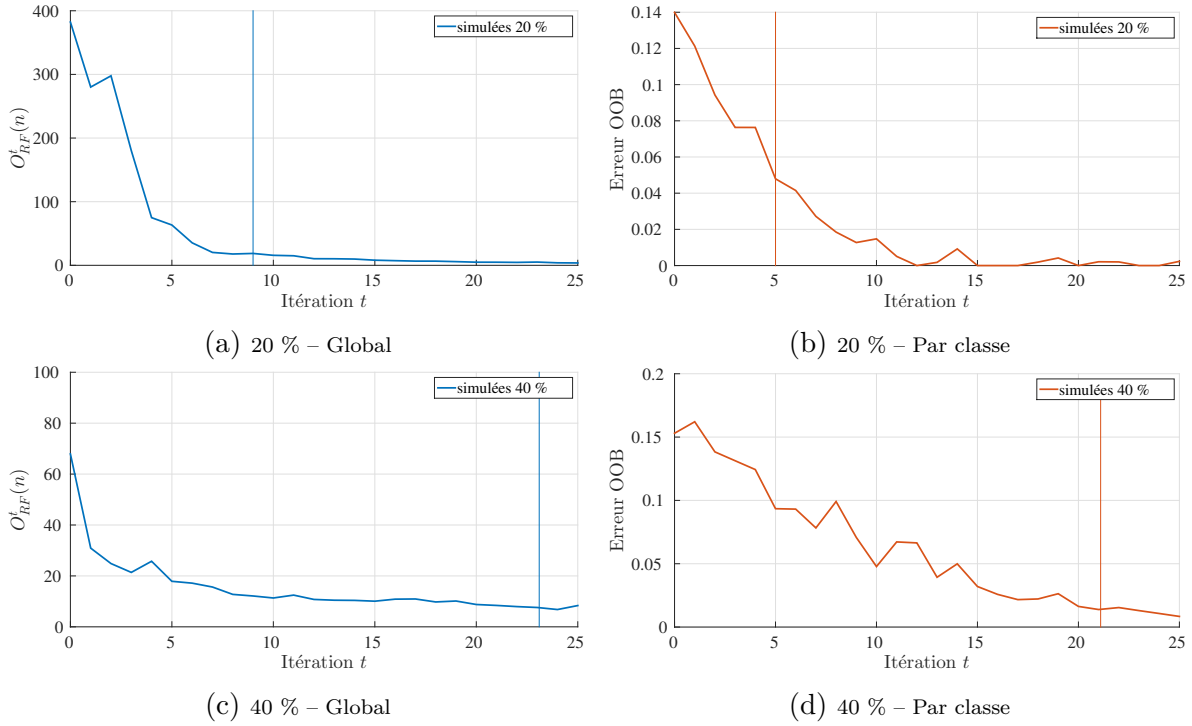
Dans ces travaux, nous nous intéressons à l'erreur OOB présentée à la Section 2.4.2 afin de définir le critère d'arrêt pour le filtrage par classe. Pour rappel, l'erreur OOB est calculée en utilisant les échantillons OOB de chaque arbre. L'idée ici est d'étudier l'erreur OOB commise par le modèle du RF appris à chaque itération. Idéalement cette erreur ErrOOB correspond à celle obtenue sur des échantillons test indépendants, ainsi ErrOOB $= 1 - OA$ [Lawrence et al., 2006; Rodríguez-Galiano et al., 2012; Waske and Benediktsson, 2007]. Dans notre contexte, l'erreur OOB doit être élevée tant que le nombre de données mal étiquetées est important. Puis, elle doit théoriquement se stabiliser lorsque la majorité des données mal étiquetées est supprimée.

Pour résumer, la stabilité de la valeur $O_{RF}(n)$ du n-ième échantillon supprimé et de l'erreur OOB est étudiée comme critère d'arrêt pour le filtrage global et par classe respectivement. Ces deux critères d'arrêt sont observés pour les vingt-cinq premières itérations. Comme précédemment, les résultats sont présentés seulement pour la mesure de similarité Breiman. Ainsi, la Figure 7.16 montre ces résultats pour les données simulées. Des résultats similaires, non-présentés ici, sont obtenus pour les données SPOT-Landsat. La première ligne correspond à un niveau de bruit de 20 %, tandis que la seconde ligne correspond à un niveau de bruit de 40 %. La première colonne montre les valeurs du score $O^t_{RF}(n)$ pour le filtrage itératif global, tandis que la seconde colonne montre les valeurs de l'erreur OOB pour le filtrage itératif par classe. Les lignes verticales indiquent l'itération pour lequel le maximum d'OA est atteint.

Concernant le filtrage itératif global, la Figure 7.16 montre que les valeurs $O^t_{RF}(n)$ décroissent rapidement pendant les premières itérations. Pour le niveau de bruit à 20 %, les valeurs $O^t_{RF}(n)$ se stabilisent proche de 0 au moment où le maximum d'OA est atteint. En revanche, les valeurs $O^t_{RF}(n)$ se stabilisent aux alentours de la dixième itération pour un niveau de bruit à 40 %. Pourtant, le maximum d'OA n'est pas encore atteint. Il l'est seulement après la vingtième itération. Pour ces données, les valeurs $O^t_{RF}(n)$ diminuent très rapidement alors qu'un nombre important de données mal étiquetées est encore présent. La diminution des scores $O^t_{RF}(n)$ n'est donc peut être pas due uniquement à la suppression des données mal étiquetées. Il est probable que la dispersion des scores d'*outlier* O_{RF} change au cours des itérations. Comme le nombre d'échantillons diminue, les arbres du RF sont moins profonds et contiennent moins de nœuds. Cette modification dans la structure des arbres peut alors avoir une conséquence sur la dispersion des scores d'*outlier*.

Concernant le processus itératif par classe, l'erreur OOB semble intéressante comme critère d'arrêt. Pour un niveau de bruit de 40 %, les valeurs de l'erreur OOB se stabilisent au moment où le maximum d'OA est atteint. Pour un niveau de bruit de 20 %, les valeurs de l'erreur OOB se stabilisent aux alentours de la dixième itération, soit cinq itérations après l'obtention du maximum d'OA. Or, la Figure 7.13b montre que les valeurs d'OA sont stables entre la cinquième et la dixième itération. Ainsi, ce n'est pas un problème si le filtrage est arrêté à la dixième itération.

En conclusion, les deux filtrages itératifs proposés semblent complémentaires. D'une



(a) 20 % – Global (b) 20 % – Par classe

(c) 40 % – Global (d) 40 % – Par classe

FIGURE 7.16 – Évaluation des critères d'arrêt pour les filtrages itératifs global et par classe pour les données simulées.

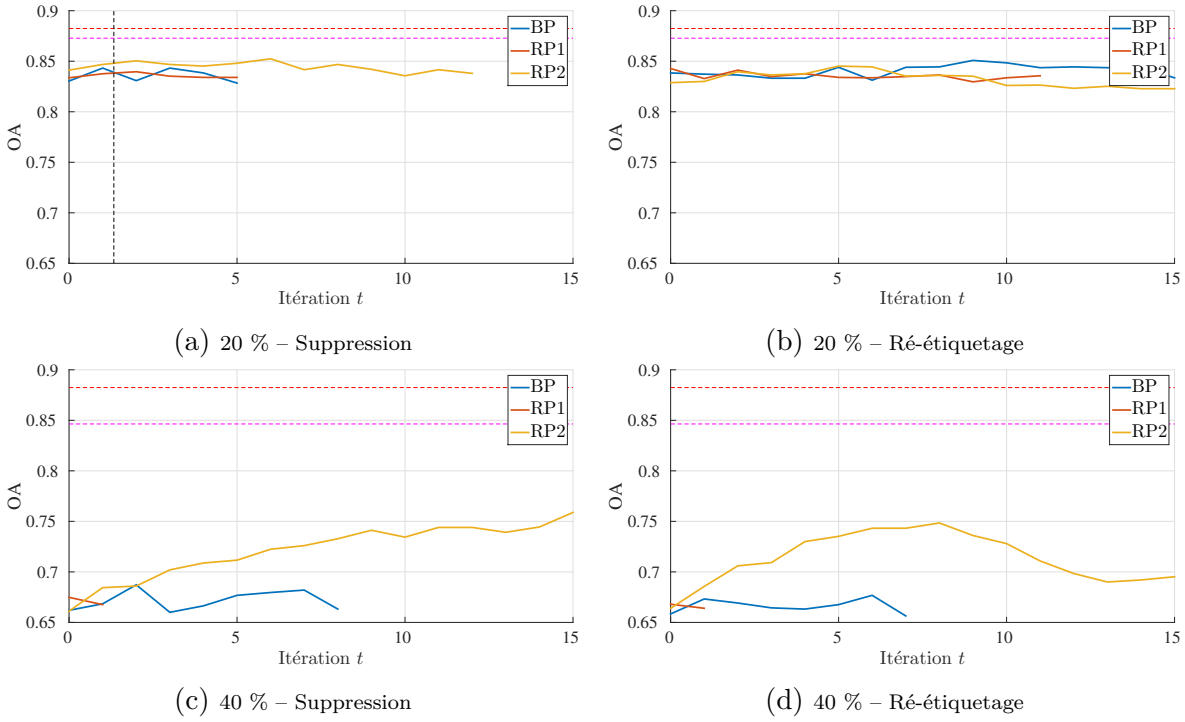
part, le processus itératif global n'élimine pas de données correctement étiquetées mais il manque certaines données mal étiquetées. Par ailleurs, le critère d'arrêt basé sur l'évolution des valeurs $O_{RF}(n)$ au cours des itérations n'est pas satisfaisant pour toutes les situations. D'autre part, le processus itératif par classe, plus agressif, permet la suppression de la quasi-totalité des données mal étiquetées, mais un nombre conséquent de données correctement étiquetées est aussi supprimé. De plus, les valeurs de l'erreur OOB peuvent être utilisées comme critère d'arrêt pour ce processus. Pour les deux filtrages, les trois mesures de similarité évaluées montrent des résultats quasiment identiques.

Basé sur la combinaison des prédictions des arbres du *Random Forest*

Les filtrages itératifs étudiés ici sont décrits à la Section 7.2.2. Ces filtrages sont divisés en deux catégories en fonction du traitement appliqué aux données mal étiquetées. La première catégorie supprime ces données à chaque itération, tandis que la seconde catégorie utilise le vecteur de probabilité pour ré-étiqueter ces données à chaque itération.

L'évaluation est réalisée avec les mêmes critères que les précédentes études présentées dans ce chapitre. Ainsi, les mêmes jeux de données et la même configuration pour l'apprentissage de tous les algorithmes du RF sont utilisés. Ces informations sont notamment disponibles pour les versions non-itératives de ce type de filtrage.

La Figure 7.17 montre l'évolution de l'OA au cours des itérations pour les données simulées. Les différentes lignes montrent les résultats pour des niveaux de bruit de 20 et 40 %. La première colonne correspond au traitement qui supprime les échantillons identifiés comme étant mal, tandis que la seconde colonne correspond au ré-étiquetage. Pour chaque figure, les résultats obtenus en utilisant les trois calculs de vecteur de probabilité sont présentés : en bleu pour BP, en rouge pour RP1, et en jaune pour RP2. Les lignes horizontales en pointillé rouge et magenta indiquent les valeurs d'OA obtenues par le RF appris sur les données sans bruit et parfaitement filtrées respectivement.



(a) 20 % – Suppression

(b) 20 % – Ré-étiquetage

(c) 40 % – Suppression

(d) 40 % – Ré-étiquetage

FIGURE 7.17 – Évolution de l'*Overall Accuracy* (OA) au cours des itérations pour les processus itératifs basés sur les prédictions des arbres du *Random Forest* appliqués sur les données simulées pour 20 et 40 % de données mal étiquetées.

Pour rappel, ces filtrages s'arrêtent automatiquement lorsque la classe qui obtient le maximum de probabilité est identique à celle fournie par la donnée de référence. Sur la Figure 7.12, les filtrages étudiés s'arrêtent alors à différentes itérations.

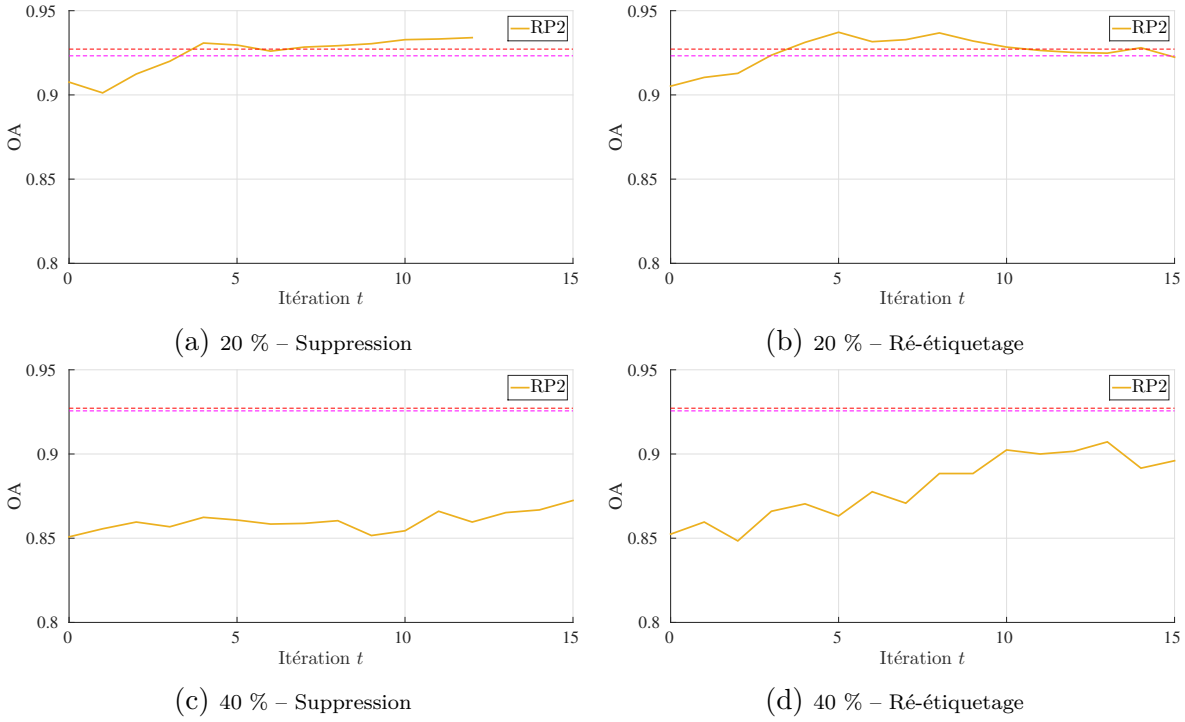
La Figure 7.17 montre que le gain apporté par l'étape de filtrage est différent entre les deux niveaux de bruit. Ainsi, les performances de classification ne sont pas améliorées pour un niveau de bruit de 20 %. En revanche, elles augmentent pour un niveau de bruit de 40 % lorsque le vecteur de probabilité RP2 est utilisé. Cependant, ces performances sont loin des cas idéaux (droites horizontales en pointillé).

Par ailleurs, les vecteurs de probabilité BP et RP1 échouent à détecter les données mal étiquetées. L'utilisation de ces deux vecteurs conduit à des filtrages restrictifs qui identifient très peu de données mal étiquetées. Ces pauvres résultats sont peut être la conséquence du choix du critère de décision. Dans les approches classiques de la littérature décrites à la Section 7.1.2, les échantillons sont identifiés comme mal étiquetés si leur probabilité qu'ils appartiennent à la classe fournie par la donnée de référence est inférieure à 50 %. Dans ces travaux, la probabilité de la classe de référence doit être différente de la probabilité maximale, qui peut être bien inférieure à 50 %.

De plus, la Figure 7.17 montre que les maximums d'OA obtenus pour la suppression et le ré-étiquetage sont similaires. Cependant, ce maximum est atteint plus rapidement en utilisant le ré-étiquetage.

La même évaluation est réalisée sur les données SPOT-Landsat. En suivant le même style que précédemment, la Figure 7.18 montre ces nouveaux résultats pour le calcul du vecteur de probabilité RP2. Comme pour la version non-itérative de ces filtrages, l'utilisation des vecteurs de probabilité BP et RP1 ne permet pas la détection de données mal étiquetées.

Dans cet exemple, le ré-étiquetage des données identifiées comme étant mal étiquetées permet d'obtenir des valeurs d'OA plus élevées que la suppression. Par ailleurs, les



(a) 20 % – Suppression

(b) 20 % – Ré-étiquetage

(c) 40 % – Suppression

(d) 40 % – Ré-étiquetage

FIGURE 7.18 – Évolution de l'*Overall Accuracy* (OA) au cours des itérations pour les processus itératifs basés sur les prédictions des arbres du *Random Forest* appliqués sur les données SPOT-Landsat contaminées par 20 et 40 % de données mal étiquetées.

performances optimales sont obtenues lorsque le niveau de bruit est de 20 %.

Comparaison des performances entre les filtrages itératifs

Dans cette partie, les filtrages itératifs étudiés sont comparés. Pour ce faire, les maximums d'OA obtenus pour les différents filtrages sont montrés dans le Tableau 7.9. De plus, les valeurs d'OA, présentées dans le Tableau 7.3, sont aussi affichées sur la troisième et quatrième ligne respectivement. Les valeurs en gras indiquent les meilleures valeurs d'OA pour chaque jeu de données.

Le Tableau 7.9 montre que les filtrages itératifs basés sur les scores d'*outlier* du RF obtiennent globalement les meilleures performances. Avec ces filtrages, les valeurs d'OA atteintes sont très proches du cas idéal (filtrage idéal) excepté pour les données simulées à 20 %. Ce dernier jeu de données représente un challenge. Si on compare les résultats entre les cas sans filtrage et filtrage idéal, une baisse d'OA de 17 % est observée.

Concernant les autres filtrages, la stratégie de ré-étiquetage des échantillons basée sur le vecteur de probabilité RP2 obtient les meilleures performances sur les données SPOT-Landsat. Cependant, ce type de filtrage est moins performant pour les données simulées. Les filtrages itératifs basés sur la méthode ENN et sur les prédictions du RF (suppression) obtiennent de moins bons résultats.

La comparaison de ces résultats avec ceux du Tableau 7.8 obtenus pour les filtrages non-itératifs montre que les résultats sont similaires. L'utilisation de filtrages itératifs ne permet donc pas d'améliorer les performances de l'algorithme de classification final par rapport à un filtrage non-itératif. Cependant, les résultats des filtrages non-itératifs présentés dans le Tableau 7.8 sont obtenus avec les valeurs optimales des paramètres k et n. Or, les valeurs de ces paramètres ne sont pas triviales à configurer. De plus, le filtrage itératif par classe ne nécessite pas la configuration du paramètre n.

Tableau 7.9 – Valeurs de l'*Overall Accuracy* (OA) obtenues pour les filtrages itératifs basés sur les méthodes d'édition, les scores d'*outlier* du *Random Forest* (RF) et les prédictions des arbres du RF. Les données utilisées sont les données simulées et SPOT-Landsat avec 20 et 40 % de bruit. Les valeurs en gras indiquent les meilleures valeurs d'OA.

		Simulées		**SPOT-Landsat**	
		20 %	40 %	20 %	40 %
OA sans filtrage		83,8	67,0	90,0	85,6
OA filtrage idéal		87,3	84,6	92,3	92,6
ENN	k	84,0	75,3	91,0	90,0
	$k_{\%}$	84,3	75,1	90,2	90,5
Score d'*outlier* O_{RF} : global	O_{RF}^{Breiman}	86,8	76,6	92,0	90,0
	$O_{RF}^{\text{DistanceLCA}}$	**87,2**	78,0	92,6	90,3
	$O_{RF}^{\text{PuretyLCA}}$	85,7	**78,1**	92,0	**90,7**
Score d'*outlier* O_{RF} : par classe	O_{RF}^{Breiman}	86,9	75,6	92,8	90,6
	$O_{RF}^{\text{DistanceLCA}}$	**87,2**	74,4	91,8	90,4
	$O_{RF}^{\text{PuretyLCA}}$	87,5	77,1	91,5	90,1
Prédictions du RF : suppression	$p_{C_i}^{\text{BP}}(x)$	84,3	68,7	-	-
	$p_{C_i}^{\text{RP1}}(x)$	84,0	66,8	-	-
	$p_{C_i}^{\text{RP2}}(x)$	82,2	75,9	93,4	88,2
Prédictions du RF : ré-étiquetage	$p_{C_i}^{\text{BP}}(x)$	85,1	67,7	-	-
	$p_{C_i}^{\text{RP1}}(x)$	84,1	66,4	-	-
	$p_{C_i}^{\text{RP2}}(x)$	84,5	73,6	**93,7**	**90,7**

Afin de confirmer les résultats obtenus sur les différents types de filtrage, la meilleure version – non-itérative ou itérative – de chaque filtrage est gardée pour l'évaluation des données Sentinel-2.

7.4.3 Étude des données Sentinel-2

Les filtrages ayant obtenus les meilleurs résultats dans les Sections 7.4.1 et 7.4.2 sont évalués ici avec les données Sentinel-2 (Section 7.3.2). Les évaluations sont réalisées principalement sur les performances de classification obtenues après l'étape de filtrage. L'algorithme de classification final est une nouvelle fois l'algorithme du RF présenté à la Section 7.3.1. Ces résultats seront comparés avec les deux résultats de référence présentés dans le Tableau 7.4 (sans bruit et filtrage idéal).

Plus précisément, les trois catégories de filtrage étudiées ici sont :

— La méthode ENN. Cette méthode nécessite uniquement la configuration du paramètre de voisinage k. Ainsi, une étude de sensibilité vis-à-vis du paramètre k est réalisée en faisant varier sa valeur de 1 à 341 par pas de 10.

— Les filtrages itératifs global et par classe basés sur les scores d'*outlier* O_{RF}. Concernant le filtrage global, la valeur du critère de décision n est fixée à 50 comme pour les études de la Section 7.4.2. De plus, le critère d'arrêt basé sur l'évolution de $O^t_{RF}(n)$ est étudié. Concernant le filtrage par classe, la règle des 3σ est appliquée par classe comme critère de décision. De plus, l'évolution des valeurs de l'erreur OOB est utilisée pour définir un critère d'arrêt. Comme les études sur les différentes mesures de similarité ne sont pas concluantes, les résultats sont montrés seulement pour la mesure de similarité proposée par Breiman.

— Les filtrages itératifs basés sur la combinaison des prédictions des arbres du RF. Les deux types de traitements appliqués aux données identifiées comme étant mal étiquetées sont étudiés. Par ailleurs, le vecteur de probabilité est calculé uniquement en utilisant l'équation 7.4.

Dans un premier temps, le filtrage non-itératif basé sur la méthode ENN est évalué. Dans un second temps, le meilleur résultat de la méthode ENN est comparé avec les performances obtenues par les filtrages itératifs retenus.

L'apport du filtrage ENN est mesuré pour les deux critères d'évaluation décrits à la Section 7.3.1. De plus, la sensibilité de la méthode ENN vis-à-vis du paramètre k est évaluée en testant des valeurs de k allant de 1 à 371 par pas de 10. Ainsi, la Figure 7.19 montre les résultats des deux critères d'évaluation pour les différentes valeurs de k. L'axe des ordonnées de gauche montre les valeurs des métriques associées à la précision du filtrage, tandis que l'axe des ordonnées de droite montre les valeurs d'OA obtenues après l'étape de filtrage. Concernant les mesures de précision, la courbe en vert représente la mesure FP, en noir l'erreur ER_1 et en rouge l'erreur ER_2. Par ailleurs, les lignes horizontales en pointillé rouge et magenta indiquent les valeurs d'OA obtenues par le RF appris sur les données sans bruit et parfaitement filtrées respectivement (Tableau 7.4).

La Figure 7.19 montre que le maximum d'OA est obtenu pour une valeur de k égale à 20. À cette itération, la précision de l'étape de filtrage est maximale, et les erreurs ER_1 et ER_2 se stabilisent aux alentours de 0,1. Cela signifie que seulement 10 % des données mal étiquetées ne sont pas détectées par la méthode ENN. Pour ces données, la sensibilité de la méthode vis-à-vis du paramètre k est identique à celle observée pour les données simulées et SPOT-Landsat.

Le meilleur résultat obtenu par le filtrage ENN est maintenant comparé avec les filtrages itératifs. Pour ces derniers, la Figure 7.20 montre l'évolution de l'OA au cours

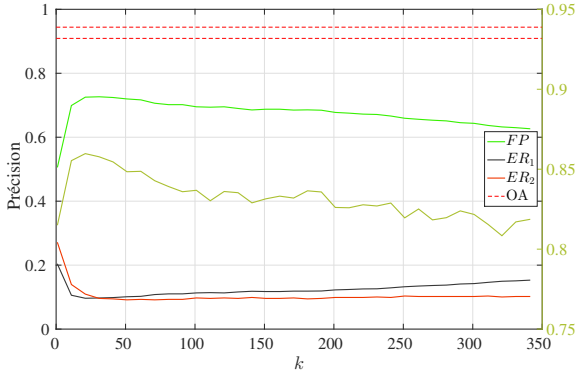


FIGURE 7.19 – Valeurs d'*Overall Accuracy* (OA) et de précision de la méthode *Edited Nearest Neighbor* (ENN) en fonction du paramètre de voisinage k pour les données Sentinel-2.

des itérations. Chaque courbe représente un filtrage itératif. Les courbes bleue et rouge montrent les résultats pour les filtrages itératifs global et par classe respectivement. Les deux autres courbes montrent les résultats pour les filtrages itératifs basés sur les prédictions du RF en utilisant le vecteur RP2 : en jaune pour la suppression, et en violet pour le ré-étiquetage. Par ailleurs, les lignes horizontales en pointillé rouge et magenta représentent les valeurs d'OA obtenues par le RF appris sur les données sans bruit et parfaitement filtrées respectivement. De plus, la ligne horizontale en pointillé vert montre le maximum d'OA obtenu pour la méthode ENN sur la Figure 7.19.

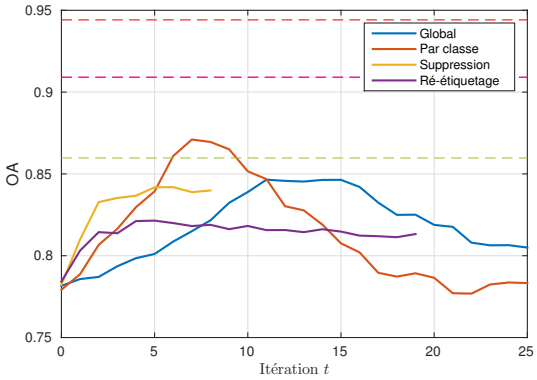


FIGURE 7.20 – Valeurs d'*Overall Accuracy* (OA) pour différents filtrages itératifs sur les données Sentinel-2.

Les courbes des filtrages itératifs basés sur la combinaison des prédictions du RF (en jaune et violet) sont de longueurs différentes puisque ces deux filtrages itératifs s'arrêtent automatiquement.

La Figure 7.20 montre que l'ajout d'une étape de filtrage pour « nettoyer » les échantillons d'apprentissage permet d'améliorer les performances de classification. La valeur d'OA obtenue par le RF appris sur les données bruitées peut être observée à l'itération $t = 0$ [72]. Par ailleurs, cette figure montre que les filtrages basés sur les scores d'*outlier* permettent d'obtenir les meilleurs performances en classification. De plus, la suppression pour le filtrage basé sur les prédictions des arbres du RF permet ici d'obtenir une valeur d'OA quasiment identique à celle du filtrage global. Ce qui n'était pas le cas lors des études

72. Les petites différences d'OA visibles entre les quatre filtrages (global, par classe, suppression et ré-étiquetage) à l'itération $t = 0$ sont dues à l'utilisation de l'aléatoire (échantillons *bootstrap* et principe du *random feature selection*) lors de la construction des arbres du RF.

sur les données simulées et SPOT-Landsat. En revanche, les performances obtenues avec le ré-étiquetage des données se stabilisent à une valeur d'OA peu élevée. Malgré une forte amélioration des performances, les maximums d'OA atteints sont encore loin des valeurs d'OA des deux résultats de référence idéaux (sans bruit et filtrage parfait).

Dans les données Sentinel-2, le niveau de bruit est différent par classe. Or, l'étude de l'OA ne permet pas de visualiser les éventuelles différences de performances entre les classes. Ainsi, les valeurs de F-Score pour les six classes, obtenues pour chaque filtrage, sont comparées. Pour la méthode ENN, les valeurs sont montrées pour la configuration optimale du paramètre k. Pour les filtrages itératifs, les valeurs sont montrées pour l'itération t où le maximum d'OA est obtenu. Le Tableau 7.10 montre ces résultats. Afin de faciliter les comparaisons, les trois premières colonnes rappellent les performances obtenues par un RF appris en utilisant 1) les donnée sans bruit, 2) les données bruitées, et 3) les données parfaitement filtrées. Les valeurs en gras indiquent les valeurs de F-Score par classe et d'OA les plus élevées.

Tableau 7.10 – Valeurs d'*Overall Accuracy* (OA) et de F-Scores obtenues pour les données Sentinel-2 pour différentes stratégies de filtrage. Les valeurs en gras indiquent les meilleurs résultats obtenus parmi les filtrages étudiés.

	2016 sans bruit	**2014** sans filtrage	**2014** filtrage idéal	**ENN** $k = 21$	**Global** $t = 11$	**Par Classe** $t = 7$	**Suppression** $t = 6$	**Ré-étiquetage** $t = 4$
CP	95,5	79,5	92,9	88,5	86,0	**89,9**	85,5	83,0
M	96,6	84,3	93,8	**90,7**	89,3	90,2	89,7	88,5
C	91,4	40,0	80,9	**79,6**	75,2	75,4	79,3	79,3
T	95,8	79,6	93,5	**90,7**	89,2	**90,7**	90,3	90,4
V	84,2	78,8	77,3	58,3	**69,2**	63,1	58,0	49,0
P	89,5	79,2	85,4	77,1	77,0	**80,5**	73,9	71,1
OA	94,4	78.2	90,9	86,0	84,6	**87,1**	84,2	82,1

CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.

Le Tableau 7.10 montre des différences dans les valeurs de F-Score obtenues entre les différents filtrages étudiés. Par exemple, le filtrage par classe obtient la valeur d'OA la plus élevée, et aussi les valeurs de F-Score les plus élevées pour les classes céréales à paille, tournesol et vigne. En revanche, les valeurs de F-Score pour la classe de colza sont en-deçà des autres filtrages étudiés. Les filtrages ENN et global obtiennent aussi de bons résultats pour une majorité des classes. Néanmoins, leurs F-Scores sur la classe prairie sont plus faibles que celui obtenu pour le filtrage par classe.

Par ailleurs, il est intéressant de noter le comportement de la valeur du F-Score pour la classe vigne, qui ne contient pas de données mal étiquetées. Pour cette classe, la valeur de F-Score obtenue sans filtrage est plus élevée que les valeurs de F-Score obtenues après l'utilisation d'une étape de filtrage. Une analyse (non-montrée ici) des valeurs de précision et de rappel indique que le rappel de la classe vigne reste identique mais que sa précision diminue. Autrement dit, de nombreux échantillons appartenant à d'autres classes sont prédits comme de la vigne lorsque le classifieur est appris sur les échantillons résultant de l'étape de filtrage. Ainsi, nous soupçonnons que les filtrages utilisés suppriment des échantillons correctement étiquetés informatifs (*e.g.* en bordure de classe). Les frontières de décision sont alors moins bien définies entre classes, et des échantillons n'appartenant pas à la classe vigne sont prédits comme de la vigne.

Pour mieux comprendre les différentes valeurs d'OA obtenues, la précision de l'étape de filtrage est étudiée pour les trois filtrages obtenant les valeurs d'OA les plus élevées. Ainsi, le Tableau 7.11 montre les valeurs de FP, ER_1 et ER_2 pour le filtrage ENN et les filtrages itératifs global et par classe. Les valeurs de k et t sont celles pour lesquelles le

maximum d'OA est atteint. Les valeurs en gras indiquent le filtrage le plus performant. De plus, les valeurs d'OA sont rappelées.

Tableau 7.11 – Valeurs des précisions FP, ER_1 et ER_2 (en pourcentage) pour le filtrage ENN et les filtrages itératifs global et par classe. Les valeurs en gras indiquent le filtrage le plus performant.

	ENN $k = 21$	**Global** $t = 11$	**Par Classe** $t = 7$
OA	86,0	84,6	**87,1**
FP	72,1	**78,5**	63,4
ER_1	9,8	**6,9**	15,5
ER_2	11,0	11,3	**5,7**

Le Tableau 7.11 montre que le filtrage itératif par classe obtient la plus petite valeur d'erreur ER_2, mais aussi la plus faible précision FP et la plus forte erreur ER_1. De plus, ce filtrage permet d'obtenir les valeurs d'OA les plus élevées. Ainsi, le meilleur résultat est ici obtenu par un filtrage agressif qui supprime un grand nombre d'échantillons même correctement étiquetés.

Malgré une différence d'OA de 1,4 %, les précisions du filtrage ENN et du filtrage itératif global sont similaires. Ce résultat est donc contre-intuitif, le filtrage itératif global obtient une valeur d'OA moins élevée alors que sa précision est meilleure.

Afin de mieux comprendre ce denier résultat, le nombre d'échantillons par classe restant après l'étape de filtrage est analysé pour les trois filtrages. Chaque échantillon est donné en fonction de sa classe de référence (c_r) et de sa classe vérité terrain (c_{vt}). Les Tableaux 7.13, 7.14 et 7.15 montrent ces résultats pour les filtrages ENN, global et par classe respectivement. Les valeurs de k et t sont celles pour lesquelles le maximum d'OA est atteint. Les lignes indiquent l'occupation des sols en 2014, tandis que les colonnes donnent l'occupation des sols en 2016. Ainsi, la colonne total indique le nombre d'échantillons utilisés par classe pour l'apprentissage du RF après l'étape de filtrage. De plus, la dernière colonne « % ME » indique le pourcentage de données mal étiquetées présentes dans chaque classe. Les valeurs en gras sur la diagonale indiquent donc le nombre d'échantillons correctement étiquetés utilisés pour l'apprentissage. Afin de comparer plus facilement ces résultats, le Tableau 6.7, présenté à la Section 6.4.3, donnant le nombre d'échantillons présents avant l'étape de filtrage est ici dupliqué (Tableau 7.12).

L'analyse de ces quatre tableaux permet de mieux comprendre les différences de F-Score et de précision observées entre les trois filtrages. Par exemple, les trois filtrages n'ont pas les mêmes difficultés à détecter les échantillons mal étiquetés appartenant à la classe prairie. En effet, le pourcentage de données mal étiquetées présentes dans la classe prairie est de 6,7 et 11,6 % après un filtrage ENN ou global. Ce pourcentage est moins élevé pour le filtrage itératif par classe, mais 84 données correctement étiquetées de la classe vigne ont été supprimées pour ce filtrage.

Dans ce problème de détection, certains échantillons mal étiquetés sont plus faciles à identifier que d'autres. Par exemple, les échantillons appartenant à une classe de culture d'été mais étiquetés par une classe de culture d'hiver sont faciles à identifier. En effet, le vecteur de variables de ces données mal étiquetées sera différent des autres échantillons correctement étiquetés dans la classe. Ainsi, les 8 échantillons référencés comme maïs alors qu'ils appartiennent à la classe de colza sont supprimés par tous les filtrages. En revanche, les trois filtrages échouent à supprimer une grande partie des échantillons étiquetés comme tournesol alors qu'ils appartiennent à la classe maïs. Ces données sont plus

Tableau 7.12 – Nombre d'échantillons d'apprentissage avant l'étape de filtrage en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt})

2016 (c_{vt}) / **2014** (c_r)	**CP**	**M**	**C**	**T**	**V**	**P**	**Total**	**% ME**
CP	**354**	20	45	66	1	14	500	29,2
M	89	**378**	8	16	0	9	500	24,4
C	56	12	**350**	72	0	10	500	30,0
T	59	32	47	**351**	0	11	500	29,8
V	0	0	0	0	**500**	0	500	0,0
P	76	16	8	0	0	**400**	500	20,0

CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.
% ME : pourcentage de données mal étiquetées

Tableau 7.13 – Nombre d'échantillons utilisés pour l'apprentissage, pour les données Sentinel-2, en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) après un filtrage *Edited Nearest Neighbor* ($k = 21$).

2016 (c_{vt}) / **2014** (c_r)	**CP**	**M**	**C**	**T**	**V**	**P**	**Total**	**% ME**
CP	**288**	0	4	3	0	1	296	1,5
M	2	**336**	0	2	0	1	341	1,5
C	2	0	**327**	2	0	1	332	0,0
T	4	14	0	**300**	0	1	319	0,1
V	0	0	0	0	**491**	0	491	0,0
P	27	9	0	0	0	**366**	402	6,7

CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.
% ME : pourcentage de données mal étiquetées

Tableau 7.14 – Nombre d'échantillons utilisés pour l'apprentissage, pour les données Sentinel-2, en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) après un filtrage itératif global ($t = 11$).

2016 (c_{vt}) / **2014** (c_r)	**CP**	**M**	**C**	**T**	**V**	**P**	**Total**	**% ME**
CP	**292**	0	4	2	0	0	298	2,0
M	0	**351**	0	1	0	1	353	0,6
C	2	0	**345**	0	0	0	347	0,6
T	0	15	0	**332**	0	3	350	5,1
V	0	0	0	0	**496**	0	496	0,0
P	40	7	0	0	0	**359**	406	11,6

CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.
% ME : pourcentage de données mal étiquetées

Tableau 7.15 – Nombre d'échantillons utilisés pour l'apprentissage, pour les données Sentinel-2, en fonction de leur classe de référence (c_r) et de leur classe vérité terrain (c_{vt}) après un filtrage itératif par classe ($t = 7$).

2016 (c_{vt}) / **2014** (c_r)	**CP**	**M**	**C**	**T**	**V**	**P**	**Total**	**% ME**
CP	**328**	0	5	3	0	5	341	3,8
M	0	**312**	0	0	0	1	313	0,3
C	0	0	**329**	0	0	0	329	0,0
T	0	13	0	**307**	0	2	322	4,7
V	0	0	0	0	**382**	0	382	0,0
P	9	0	0	0	0	**316**	325	2,8

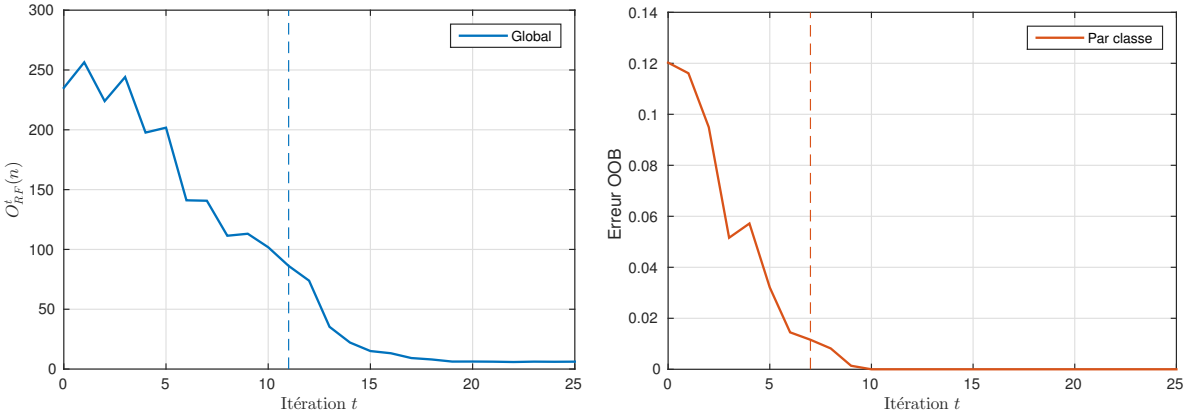
CP : Céréales à paille. M : Maïs. C : Colza. T : Tournesol. V : Vignes. P : Prairies.
% ME : pourcentage de données mal étiquetées

difficiles puisque le tournesol et le maïs sont deux cultures d'été : les vecteurs de variables sont donc similaires.

De plus, les Tableaux 7.13, 7.14 et 7.15 confirment les résultats précédents, notamment sur le fait que le filtrage itératif par classe est très agressif. En effet, la comparaison des nombres d'échantillons encore présents par classe (avant-dernière colonne) montre que ce dernier est bien inférieur pour le filtrage itératif par classe.

Par ailleurs, le Tableau 7.15 montre que la quasi-totalité des échantillons mal étiquetés est supprimée pour le filtrage itératif par classe quand le maximum d'OA est atteint. Pourtant, la valeur d'OA obtenue est encore loin de celle obtenue lorsque l'apprentissage est réalisé sur les données parfaitement filtrées. Deux raisons sont possibles : 1) soit les dernières données mal étiquetées encore présentes perturbent l'apprentissage de l'algorithme de classification, 2) soit les échantillons correctement étiquetés qui ont été supprimés sont très informatifs pour l'algorithme de classification. Une étude dont les résultats ne sont pas présentés ici a montré que la suppression des dernières données mal étiquetées (*i.e.* la configuration du Tableau 7.15 avec des 0 hors-diagonale) augmente l'OA de 2 %, tandis que la ré-injection des données supprimées mais correctement étiquetées (*i.e.* la configuration du Tableau 7.15 avec la diagonale identique au Tableau 7.12) permet d'augmenter l'OA de 4 %.

Concernant les deux filtrages itératifs global et par classe, le critère d'arrêt doit être analysé. Comme dans la Section 7.4.2, la stabilité des valeurs $O_{RF}(n)$ et de l'erreur OOB sont analysées. La Figure 7.21 montre ces deux résultats : à gauche pour le filtrage itératif global et à droite pour le filtrage itératif par classe. Les lignes verticales en pointillé indiquent l'itération où le maximum d'OA est atteint.



(a) Global (b) Par classe

FIGURE 7.21 – Étude du critère d'arrêt sur les données Sentinel-2 pour les filtrages itératifs global et par classe.

Concernant le filtrage itératif global, la Figure 7.21a montre que les valeurs $O_{RF}(n)$ se stabilisent autour de la quinzième itération alors que le maximum d'OA est atteint à la onzième itération. Cependant, l'analyse de l'évolution des valeurs d'OA, affichées à la Figure 7.20, montre que les valeurs d'OA sont stables entre la onzième et la quinzième itération. Ce n'est donc pas un problème d'arrêter le filtrage à la onzième itération.

Concernant le filtrage par classe, la Figure 7.21b montre que les valeurs de l'erreur OOB se stabilisent juste après l'obtention du maximum d'OA. Comme pour les données simulées et SPOT-Landsat, ce critère d'arrêt est donc parfaitement adapté pour le filtrage itératif par classe.

7.5 Conclusion

Dans ce chapitre, un cadre méthodologique permettant de prendre en compte la présence de données mal étiquetées dans le processus de classification a été étudié. La solution proposée consiste à ajouter une étape de filtrage afin de « nettoyer » les échantillons d'apprentissage. L'objectif principal de cette étape de filtrage est alors d'améliorer les performances de classification par rapport à celles qui seraient obtenues en utilisant les échantillons bruités.

Les filtrages étudiés sont composés de deux étapes principales : 1) la détection des données mal étiquetées, et 2) le traitement appliqué à ces données détectées. Après avoir réalisé un état de l'art sur la détection des données mal étiquetées, deux catégories de filtrage ont été analysées. La première s'appuie sur des méthodes de détection d'*outlier*, tandis que la seconde approche repose sur l'utilisation d'un ensemble d'algorithmes de classification. Cette catégorisation a donné lieu à l'étude de deux stratégies.

La première stratégie repose sur l'utilisation des scores d'*outlier* calculés par le RF. Dans cette méthode, un critère de décision est nécessaire pour identifier les données mal étiquetées. Le premier critère étudié repose sur la suppression des n échantillons ayant les plus forts scores d'*outlier*. Cependant, il ne permet pas de prendre en compte les différences de niveaux de bruit entre les classes. Ainsi, un second critère pour lequel le nombre d'échantillons identifiés est différent par classe a été utilisé. Ce critère est basé sur la règle des 3σ. Ces deux critères sont notamment utilisés dans des filtrages itératifs où les données identifiées comme mal étiquetées sont supprimées à chaque itération. Pour ces filtrages itératifs, deux critères d'arrêt ont aussi été étudiés.

La seconde stratégie se base sur la combinaison des prédictions des arbres du RF. Les filtrages proposés utilisent les prédictions des arbres du RF pour calculer un vecteur de probabilité d'appartenance aux classes pour chaque échantillon. L'analyse de ce vecteur est ensuite réalisée pour identifier les données mal étiquetées. Dans un premier temps, le vecteur de probabilité d'appartenance aux classes proposé par Breiman a été étudié. L'existence de plusieurs limitations a conduit à proposer deux nouveaux calculs du vecteur de probabilité. Les différences entre les trois calculs de probabilité ont pu être observées lors des expérimentations. Dans ces travaux, deux types de traitement appliqué aux données identifiées comme étant mal étiquetées sont étudiés : la suppression et le ré-étiquetage.

Ces deux stratégies ont été évaluées sur trois jeux de données au cours de trois expérimentations différentes. En outre, elles ont été utilisées de manière non-itérative et itérative.

Dans un premier temps, l'ensemble des filtrages proposé est évalué de manière non-itérative. Le gain sur les performances de la classification des différents filtrages est notamment quantifié. De plus, ces résultats sont comparés avec ceux de la méthode d'édition ENN, qui avait montré sa précision pour identifier les données mal étiquetées au Chapitre 6. Pour cette méthode, une étude de sensibilité vis-à-vis du paramètre k est aussi réalisée. les premières évaluations montrent les bonnes performances des méthodes ENN pour lesquelles la sensibilité du paramètre k est faible. Elle met aussi en avant les bonnes performances des méthodes basées sur l'utilisation des scores d'*outlier* O_{RF} pour lesquelles le paramètre de seuil n est toutefois difficile à configurer. Les autres filtrages non-itératifs obtiennent des résultats peu satisfaisants.

Dans un deuxième temps, l'ensemble des filtrages proposé est utilisé de manière itérative. Ainsi, deux critères d'arrêt sont étudiés pour les filtrages utilisant le score d'*outlier* du RF. Dans ces expérimentations, la méthode RENN, version itérative de la méthode

ENN, a aussi été évaluée. Dans cette méthode, l'utilisation d'une taille de voisinage k identique par classe et par itération présente des désavantages. Ainsi, une variante de la méthode RENN est proposée afin d'adapter la valeur de k au nombre d'échantillons disponibles par classe. Les différentes évaluations réalisées mettent en avant les bonnes performances des filtrages basés sur les valeurs O_{RF}. En revanche, les filtrages itératifs basés sur des méthodes d'édition n'apportent pas un gain supplémentaire par rapport à l'utilisation du filtrage non-itératif ENN. Les performances des filtrages itératifs basés sur les prédictions du RF sont globalement moins bonnes. Néanmoins, le filtrage utilisant le calcul de vecteur de probabilité RP2 obtient des résultats satisfaisants.

Dans un troisième temps, les meilleurs filtrages sont évalués sur les données Sentinel-2 plus complexes. Pour ces données, le bruit correspond à une situation réelle, et le niveau de bruit est différent entre les classes. Cette étude confirme les résultats obtenus pour les deux premières évaluations. Parmi tous les filtrages, ceux basés sur les scores d'*outlier* O_{RF} obtiennent les meilleurs résultats. En particulier, le filtrage itératif par classe permet d'atteindre la valeur d'OA la plus élevée sur les données Sentinel-2. Pour les filtrages basés sur les prédictions des arbres du RF, les stratégies de suppression et de ré-étiquetage obtiennent des performances en-deçà des autres filtrages. Enfin, le filtrage basé sur la méthode ENN montre de bonnes performances avec une configuration simple de son paramètre k.

Parmi l'ensemble des filtrages, deux comportements différents sont observés. D'une part, le filtrage itératif global est précis, mais n'identifie pas certaines données mal étiquetées. D'autre part, la méthode ENN et le filtrage itératif par classe sont plus agressifs et éliminent un nombre important d'échantillons. Pour ces deux filtrages, les données mal étiquetées sont souvent toutes supprimées, mais des données correctement étiquetées sont également supprimées. Le choix du filtrage dépend donc de l'application.

Dans le contexte où un grand nombre d'échantillons est disponible, les approches plus agressives sont intéressantes. Dans un tout autre contexte où la précision du filtrage est importante, l'utilisation du filtrage global peut être intéressante. Dans ce dernier cas, les échantillons supprimés représentent quasi-uniquement des données mal étiquetées. Ainsi, les échantillons supprimés par le filtrage peuvent être utiles pour d'autres applications, comme la détection de changement.

Les évaluations sur les filtrages itératifs ont également mis en lumière l'importance de la configuration du critère d'arrêt. En effet, les meilleurs filtrages itératifs ne s'arrêtent pas automatiquement. Pour le filtrage global, l'analyse des scores d'*outlier* des échantillons supprimés n'a pas permis systématiquement d'arrêter le filtrage itératif de manière idéale. En revanche, l'étude des valeurs de l'erreur OOB pour le filtrage par classe a montré son potentiel.

L'ensemble de ces études sont menées pour un nombre restreint d'échantillons d'apprentissage (500 par classe) et pour cinq ou six classes de végétation uniquement. Des études complémentaires sont donc nécessaires afin de confirmer ces premiers résultats encourageants. Des études avec un plus grand nombre d'échantillons et une nomenclature plus complexe sont nécessaires. De plus, les filtrages étudiés ici doivent être comparés avec les filtrages basés sur les prédictions d'un ensemble d'algorithmes de classification comme celui proposé par Brodley and Friedl [1999].

Cinquième partie

Conclusion

Chapitre 8

Conclusion générale

Sommaire

8.1 Conclusions **207**
8.2 Perspectives **210**
8.2.1 Perspectives méthodologiques 210
8.2.2 Perspectives applicatives 212

8.1 Conclusions

L'objectif général de la thèse vise à améliorer la production des cartes d'occupation des sols à partir de nouvelles séries temporelles d'images satellitaires comme celles fournies par les capteurs Sentinel-2. En particulier, cette thèse s'intéresse à la classification supervisée de ces nouvelles données. Deux défis ont été identifiés dans ces travaux de thèse. Le premier concerne le choix de l'algorithme de classification supervisée, tandis que le second s'intéresse à la prise en compte des erreurs d'étiquetage souvent présentes dans la phase d'apprentissage de l'algorithme de classification.

Dans le contexte de la classification de séries temporelles sur de grandes étendues, le choix du classifieur est critique. Pour notre problématique, l'algorithme de classification doit permettre de trouver un bon compromis entre les critères suivants : la précision, le temps de calcul pas trop élevé, le paramétrage facile, la stabilité lors de traitement sur de grandes étendues et la robustesse à la présence de données mal étiquetées. En outre, l'état-de-l'art réalisé au Chapitre 2 sur les méthodes de classification a mis en avant les bonnes performances et propriétés de deux algorithmes supervisés : le RF et le SVM. Ainsi, ces deux algorithmes sont étudiés afin de répondre au premier défi de cette thèse.

En outre, les performances de l'algorithme de classification sont directement liées à la qualité des données d'apprentissage. Ces dernières sont décrites par un vecteur de variables extrait de données satellitaires et une étiquette extraite des données de référence. Un vecteur de variables idéal doit contenir une information suffisante et pertinente afin de bien caractériser les différentes classes à identifier. Dans le domaine de la cartographie de l'occupation des sols, les nouvelles séries temporelles d'images satellitaires ouvrent de nouvelles opportunités. En effet, la haute résolution temporelle de ces données est un atout pour la caractérisation des occupations des sols qui évoluent au cours du temps. Pourtant, le choix des variables à extraire de ces données est un challenge à étudier.

Par ailleurs, la qualité des étiquettes extraites de la donnée de référence est directement liée à la qualité de l'algorithme de classification, et donc de la carte produite. En télédétection, des données de qualité sont difficiles à obtenir. Elles sont de plus, parfois trop anciennes par rapport aux dates d'acquisitions des données satellitaires. Pour ces raisons, les données de référence utilisées pour les problèmes de classification contiennent bien souvent des erreurs, connues sous le nom de données mal étiquetées. Ainsi, la première partie de la thèse s'interroge sur le choix des algorithmes de classification, des données satellitaires à fournir en entrée et sur l'impact du bruit présent dans les données de référence.

Le Chapitre 4 a été consacré à l'étude des deux algorithmes de classification sélectionnés. De plus, la question sur les données satellitaires à fournir en entrée du système de classification a aussi été abordée. Dans un premier temps, le choix de l'algorithme de classification a été discuté notamment en comparant les performances de classification du RF et du SVM. Tout d'abord, le RF a montré ses bonnes performances de classification, son paramétrage facile, et un temps d'apprentissage moins important que l'algorithme du SVM. Puis, le choix des données à fournir en entrée a aussi été étudié, en proposant notamment l'utilisation de primitives temporelles calculées sur un indice de végétation. Différents vecteurs de variables extraits des données satellitaires ont alors été comparés. Il a été vu que l'ajout des primitives temporelles proposées ne permettait pas d'améliorer les performances de classification. Finalement, la stabilité du RF lors de traitements sur de grandes étendues a été abordée. En particulier, les performances de classification ont été évaluées en s'éloignant de la zone d'apprentissage, sur des paysages complexes. Cette étude a montré que si les échantillons d'apprentissage ne caractérisent pas bien les variabilités des paysages, alors les performances de classification ne sont pas bonnes. Une solution possible pour améliorer les performances de classification lors d'un apprentissage sur de grande étendues consisterait alors à stratifier les échantillons d'apprentissage en fonction des paysages.

Par ailleurs, l'influence des échantillons d'apprentissage sur les performances de classification n'a jamais été étudiée en télédétection. Le Chapitre 5 a alors proposé de quantifier cette influence pour les algorithmes du RF et du SVM. Pour ce faire, un jeu de données simulées et une procédure de génération de bruit ont spécifiquement été développés. Finalement, plusieurs expérimentations ont été menées pour différentes configurations en modifiant le nombre de classes, d'échantillons et de variables. Les principaux résultats montrent que les deux algorithmes de classification sont robustes pour des petits niveaux de bruit. Toutefois, la présence de données mal étiquetées influence moins les performances de classification du RF sur l'ensemble des configurations testées. Par ailleurs, les deux algorithmes de classification sont impactés négativement par la présence d'un grand nombre de données mal étiquetées.

Ces premières études ont donc montré l'intérêt du RF pour la classification sur de grandes étendues. Cependant, les résultats du Chapitre 5 ont mis en évidence que les performances de classification sont affectées par la présence de données mal étiquetées. Ainsi, le deuxième défi, abordé dans la deuxième partie de la thèse, vise à prendre en compte la présence des échantillons mal étiquetés dans le processus de classification. Pour ce faire, un cadre méthodologique ayant pour objectif d'améliorer la qualité des échantillons d'apprentissage a été proposé. Dans ce cas là, deux problématiques ont été soulevées. D'une part, les échantillons mal étiquetés doivent être détectés. D'autre part, une stratégie doit être définie afin de nettoyer ces données mal étiquetées.

Afin de répondre à la première problématique, un ensemble de méthodes de détection de données mal étiquetées et de détection d'*outliers* a été étudié au cours du Chapitre 6.

L'étude bibliographique montre que les méthodes d'*outlier* peuvent être utilisées pour la détection des données mal étiquetées. De plus, cette étude a mis en avant les limitations des approches existantes pour notre problématique. En particulier, une majorité de ces méthodes nécessite la définition d'une distance pour étudier la similarité entre les échantillons. Or, cette distance est complexe à définir dans le cas où les échantillons sont décrits par des séries temporelles. Ainsi, nous avons contourné cette difficulté en utilisant une méthode de détection d'*outlier* qui s'appuie sur la structure de l'ensemble des arbres construits par l'algorithme du RF. Dans cette méthode initialement proposée par Breiman, le score d'*outlier* calculé repose sur la mesure de similarité entre échantillons. Afin de mieux tirer partie de la structure des arbres, nous avons également proposé deux nouvelles mesures de similarité.

Au total, onze méthodes de détection ont été comparées sur trois jeux de données contenant différents niveaux de bruit. La sensibilité des différentes méthodes vis-à-vis de leur paramètres a aussi étudiée. Les principaux résultats ont montré la précision des scores d'*outlier* calculés par le RF permettant d'identifier les données mal étiquetées. Par, ailleurs les méthodes d'édition, couramment utilisées dans la littérature, ont montré également leur capacité à identifier la quasi-totalité des échantillons mal étiquetés. Cependant, ces méthodes ont généralement un taux de faux positifs élevé puisqu'elles font de la surdétection.

Le Chapitre 6 a donc montré que les méthodes de détection d'*outlier* étudiées peuvent être utilisées pour la détection des données mal étiquetées. Ainsi, ces méthodes peuvent être utilisées dans un cadre méthodologique permettant de prendre en compte la présence de données mal étiquetées parmi les échantillons d'apprentissage. La solution choisie consiste à appliquer une étape de filtrage pour nettoyer l'ensemble des échantillons disponibles dans la donnée de référence. Ensuite, les échantillons nettoyés sont utilisés pour l'apprentissage de l'algorithme de classification produisant les cartes d'occupation des sols.

Dans ce contexte, le Chapitre 7 a présenté plusieurs stratégies de filtrage, basées notamment sur les méthodes les plus prometteuses du Chapitre 6. Certaines stratégies de filtrage déjà existantes dans la littérature ont été étudiées et adaptées à notre problématique. Les stratégies de filtrage proposées utilisent, de manière non-itérative et itérative, différentes méthodes de détection, différents critères de décision et d'arrêt ainsi que différents traitements appliqués aux données identifiées comme mal étiquetées. Une des contributions importante de ce chapitre est notamment la proposition de deux nouveaux filtrages itératifs basés 1) sur les scores d'*outlier* et 2) sur les prédictions du RF.

Les principaux résultats ont montré que les différents filtrages testés permettent d'améliorer les performances de classification par rapport à l'utilisation de données bruitées. Plusieurs évaluations ont été menées en testant notamment les filtrages sur trois jeux de données. En particulier, une étude a été réalisée sur des données Sentinel-2 pour lesquelles la présence de données mal étiquetées est due à l'utilisation d'une donnée de référence obsolète. Pour ces données, les résultats ont alors montré que les meilleures stratégies de filtrage permettent un gain d'OA supérieur à 8 %, par rapport à l'utilisation des données bruitées. Une analyse approfondie sur les différents filtrages a montré des comportements différents. Certaines stratégies agressives, qui privilégient la suppression d'un grand nombre de données, donnent de bons résultats. Pour résumer, le Chapitre 7 a donc mis en évidence que le filtrage des données mal étiquetées est une solution possible et pertinente pour améliorer les performances des algorithmes de classification.

8.2 Perspectives

Les perspectives de ces travaux sont multiples et concernent à la fois les aspects méthodologiques et applicatifs. Les perspectives méthodologiques sont d'abord présentées, puis les perspectives applicatives sont décrites.

8.2.1 Perspectives méthodologiques

Plusieurs perspectives méthodologiques peuvent être envisagées en lien avec les deux défis abordés dans la thèse. Ainsi, ces perspectives peuvent concerner le choix des données satellitaires, le choix de l'algorithme de classification, ou encore les stratégies de filtrage des données mal étiquetées.

Dans ces travaux, les données fournies en entrée du système de classification se résument aux bandes spectrales ainsi qu'aux primitives spectrales et temporelles. Il est toutefois possible d'ajouter un plus grand nombre de primitives, notamment en utilisant le contexte spatial. Par exemple, les primitives spatiales, comme les textures d'Haralick, peuvent permettre de mieux caractériser la texture des différentes surfaces. De même l'utilisation d'approches objets ou super-pixels peut permettre d'une part de consommer moins de mémoire, et de réduire les confusions observées au Chapitre 4 sur des objets macroscopiques comme les zones arborées. Afin d'enrichir les données en entrée, une autre solution consiste à ajouter des données complémentaires. Par exemple, les données radar peuvent permettre de s'affranchir des (mauvaises) conditions lors des acquisitions avec des capteurs passifs [Balzter et al., 2015; Inglada et al., 2016; Kussul et al., 2017]. En effet, la mise à disposition gratuite des données Sentinel-1 et Sentinel-2 permet d'exploiter la synergie des deux types d'images sur de grandes étendues [Inglada et al., 2016; Waske and Benediktsson, 2007; Waske and van der Linden, 2008].

Cependant, l'ajout de ces variables conduit à de nouvelles problématiques liées notamment à la très grande dimension du problème de classification et à la gestion des gros volumes de données. Ainsi, étudier des méthodes de sélection ou de réduction du volume des données en entrée peut être intéressant dans des futurs travaux. Cela permettrait d'une part d'assurer le bon fonctionnement des algorithmes de classification, et d'autre part de réduire le temps d'apprentissage des algorithmes.

Concernant le choix de l'algorithme de classification, le RF a montré son intérêt dans le contexte de la classification de séries temporelles d'images satellitaires. Cependant, certaines limitations ont aussi pu être observées au cours des différentes études. Par exemple, des performances très faibles sont obtenues dans le Chapitre 4 pour les classes sous-représentées (*e.g.* le sorgho) ou les classes difficiles à classer présentant de forte variabilité (*e.g.* les prairies). Une solution pour ces classes plus difficiles peut consister à combiner les performances de différents algorithmes de classification. Par exemple, les résultats du Chapitre 4 montrent une certaine complémentarité entre les algorithmes du RF et du SVM. De plus, les algorithmes de classification utilisés ici ne prennent pas en compte la temporalité des données. Si les vecteurs de variables étaient mélangés, les résultats resteraient identiques. Les primitives temporelles calculées dans ces travaux ont essayé d'exploiter cette information. Cependant, le bruit sur l'estimation des paramètres de la double sigmoïde utilisée n'a pas permis une bonne exploitation de l'information temporelle.

Par ailleurs, le Chapitre 4 a mis en évidence les difficultés de la classification sur la zone des Pyrénées, et de manière générale sur des paysages présentant des caractéristiques différentes de celle de la zone d'apprentissage. La mise en place d'une stratification des échantillons d'apprentissage en fonction par exemple des caractéristiques des paysages

semble appropriée lors de traitements sur de grandes étendues.

En classification supervisée, il est courant de diviser les données de référence en deux sous-ensembles : le premier est utilisé pour l'apprentissage, tandis que le second est utilisé pour l'évaluation. Ainsi, la présence de données mal étiquetées impacte à la fois l'étape d'apprentissage et l'étape d'évaluation. Les travaux de cette thèse s'intéressent uniquement aux conséquences sur l'étape d'apprentissage, tandis que d'autres travaux considèrent uniquement la présence de données mal étiquetées dans l'étape d'évaluation [Congalton and Green, 2008; Foody, 2002, 2013]. Il serait intéressant d'enrichir ces études en analysant la présence de données mal étiquetées à la fois dans les échantillons d'apprentissage et dans les échantillons test. Cette étude permettrait entre autre de déterminer la sur-estimation faite lors de l'étape d'évaluation.

Concernant le deuxième défi abordé dans ces travaux, le Chapitre 7 montre que les filtrages itératifs améliorent les performances de la classification. Néanmoins, les performances maximales ne sont pas atteintes pour plusieurs jeux de données. Cette limitation peut être expliquée par le fait que les stratégies de filtrage proposées suppriment quelques échantillons correctement étiquetés, informatifs pour le problème de classification. Ainsi, l'ajout d'une étape supplémentaire analysant l'ensemble des échantillons supprimés peut être envisagé dans de futurs travaux.

D'une manière similaire aux approches d'apprentissage actif [Tuia et al., 2011], une stratégie consiste à ré-étiqueter les échantillons supprimés en impliquant par exemple un opérateur externe [Bouguelia, 2015; Rebbapragada, 2010]. Cette stratégie comporte des limitations car parfois l'opérateur peut échouer à ré-étiqueter correctement les échantillons supprimés, surtout s'ils sont trop difficiles (*e.g.* à la frontière entre plusieurs classes). Une stratégie plus avantageuse peut être l'utilisation d'algorithmes d'apprentissage semi-supervisé après l'étape de filtrage [Chapelle et al., 2009]. Dans ces approches des échantillons étiquetés et non-étiquetés sont utilisés. L'apprentissage semi-supervisé est alors vu comme un problème supervisé pour lequel les échantillons non-étiquetés sont utilisés pour ajuster la frontière de décision de l'algorithme de classification. Dans notre problématique, l'idée serait de désétiqueter les échantillons détectés comme mal étiquetés par l'étape de filtrage, et de les réutiliser dans un apprentissage semi-supervisé [Hughes et al., 2004]. Ainsi, les échantillons identifiés comme mal étiquetés influenceront la forme de la frontière de décision sans impacter le choix de la classe pour de nouveaux échantillons.

De plus, les méthodes de détection des données mal étiquetées étudiées ne prennent pas en compte le contexte spatial des échantillons d'apprentissage. Dans le contexte de ces travaux, les données mal étiquetées correspondent généralement à des polygones mal étiquetés. Ainsi, les échantillons mal étiquetés sont rarement isolés. Il serait donc possible de prendre en compte le voisinage spatial des échantillons dans les calculs des scores d'*outlier*.

Enfin, un aspect important à considérer dans ces travaux est l'adaptation des méthodes proposées pour pouvoir être appliquées sur un grand nombre d'échantillons. En effet, les configurations étudiées dans ces travaux se limitent à un faible nombre d'échantillons alors que la classification sur de grandes étendues nécessite un grand nombre d'échantillons. Or, les scores d'*outlier* calculés par les méthodes proposées utilisent la similarité d'un échantillon avec tous les échantillons de sa classe. Par conséquent, ce calcul de similarité réalisé entre chaque paire d'échantillons peut s'avérer coûteux. Il est donc important d'adapter les méthodes étudiées afin qu'elles puissent prendre en compte ce nombre important d'échantillons. Pour les scores d'*outlier* basés sur le RF, une solution possible est d'approximer le score $O_{RF}(p)$. Par exemple, la similarité entre p peut être calculée pour seulement un sous-ensemble des échantillons d'apprentissage qui appartienne à la même

classe de p.

8.2.2 Perspectives applicatives

Dans ces travaux, des filtrages itératifs des échantillons d'apprentissage sont proposés pour améliorer les performances des algorithmes de classification. La première application directe de ces travaux est l'obtention de cartes d'occupation des sols sur de grandes étendues. Par exemple, les données du RPG ou de l'OCS-GE pourrait être utilisée pour la classification d'une série temporelle d'images Sentinel-2. De manière similaire, les approches proposées pourraient être utilisées lorsque les étiquettes des échantillons d'apprentissage sont extraits d'une carte d'occupation des sols comme celle du CES OSO produite sur de grandes étendues.

Une deuxième application est la production de cartes d'occupation des sols de l'année en cours en utilisant des données de référence collectées les années précédentes. Ces cartes peuvent alors être utiles dans des applications réelles nécessitant un suivi régulier des territoires et une mise à jour cartographique régulière. Par exemple, le Ministère de l'Agriculture souhaite connaître les cultures semées le plus tôt possible dans l'année.

Initialement proposée pour améliorer le système de classification, la détection de données mal étiquetées ouvre des perspectives pour la détection de changements dans les bases de données d'occupation des sols. En effet, la détection des données mal étiquetées dans une base de données à partir d'images satellitaires peut permettre de détecter les changements d'occupation des sols. Dans le cadre de cette application, le nombre de faux négatifs (*i.e.* données mal étiquetées non détectées par le système) doit être quasiment nul afin de ne pas oublier des zones de changement.

Bibliographie

F. Achard, H. D. Eva, H.-J. Stibig, P. Mayaux, J. Gallego, T. Richards, and J.-P. Malingreau. Determination of deforestation rates of the world's humid tropical forests. *Science*, 297(5583):999–1002, 2002. Cité page 10.

C. C. Aggarwal. *Outlier Analysis*. 2013. Cité page 120.

R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. *European Conference on Machine Learning (ECML)*, pages 39–50, 2004. Cité page 77.

M. Akbari, A. R. Mamanpoush, A. Gieske, M. Miranzadeh, M. Torabi, and H. R. Salemi. Crop and land cover classification in Iran using Landsat 7 imagery. *International Journal of Remote Sensing*, 27(19):4117–4135, 2006. Cité page 68.

K. M. Ali and M. J. Pazzani. Error reduction through learning multiple descriptions. *Machine Learning*, 24(3):173–202, 1996. Cité page 162.

F. Angiulli and C. Pizzuti. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):203–215, 2005. Cité page 126.

O. Arino, D. Gross, F. Ranera, M. Leroy, P. Bicheron, C. Brockman, P. Defourny, C. Vancutsem, F. Achard, L. Durieux, L. Bourg, J. Latham, A. Di Gregorio, R. Witt, M. Herold, J. Sambale, S. Plummer, and J. L. Weber. GlobCover: ESA service for global land cover from MERIS. In *IEEE International Geoscience and Remote Sensing Symposium 2007 (IGARSS)*, pages 2412–2415. IEEE, 2007. Cité page 20.

I. Arnau Rodes. *Exploitation of high spatial, spectral and temporal resolution Earth observation imagery for large area land cover estimation*. PhD thesis, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), 2016. Cité pages 58, 69 , et 92.

J. J. Arsanjani, P. Mooney, A. Zipf, and A. Schauss. Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets. In *OpenStreetMap in GIScience*, pages 37–58. Springer, 2015. Cité page 18.

D. Arvor, M. Jonathan, M. S. P. Meirelles, V. Dubreuil, and L. Durieux. Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. *International Journal of Remote Sensing*, 32(22):7847–7871, 2011. Cité page 72.

M. H. A. Baig, L. Zang, T. Shuai, and Q. Tong. Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. *Remote Sensing Letters*, 5(5):423–431, 2014. Cité pages 70 et 71.

A. M. Baldridge, S. J. Hook, C. I. Grove, and G. Rivera. The ASTER spectral library version 2.0. *Remote Sensing of Environment*, 113(4):711–715, 2009. Cité page 12.

H. Balzter, B. Cole, C. Thiel, and C. Schmullius. Mapping CORINE land cover from Sentinel-1A SAR and SRTM digital elevation model data using Random Forests. *Remote Sensing*, 7(11):14876–14898, 2015. Cité page 210.

T. R. Bandaragoda. *Isolation based anomaly detection: A re-examination*. PhD thesis, Faculty of Information Technology, Monash University, 2015. Cité page 132.

R. Barandela, R. M. Valdovinos, and J. S. Sánchez. New applications of ensembles of classifiers. *Pattern Analysis & Applications*, 6(3):245–256, 2003. Cité page 163.

V. Barnett and T. Lewis. *Outliers in statistical data*. Wiley, 1974. Cité page 121.

E. Bartholomé and A. S. Belward. GLC2000: a new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing*, 26(9):1959–1977, 2005. Cité page 20.

E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1):105–139, 1999. Cité page 31.

P. S. A. Beck, C. Atzberger, K. A. Høgda, B. Johansen, and A. K. Skidmore. Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. *Remote Sensing of Environment*, 100(3):321–334, 2006. Cité pages 72 et 98.

M. Belgiu and L. Drăguţ. Random Forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016. Cité page 31.

A. S. Belward, J. E. Estes, and K. D. Kline. The igbp-dis global 1-km land-cover data set discover: A project overview. *Photogrammetric Engineering % Remote Sensing*, 65 (9):1013–1020, 1999. Cité page 20.

S. Bernard. *Forêts Aléatoires: De l'Analyse des Mécanismes de Fonctionnement à la Construction Dynamique*. PhD thesis, Université de Rouen, 2009. Cité page 46.

S. Bernard, L. Heutte, and S. Adam. Influence of hyperparameters on random forest accuracy. In *International Workshop on Multiple Classifier Systems*, pages 171–180. Springer, 2009. Cité page 46.

E. Berthier, E. Schiefer, G. K. C. Clarke, B. Menounos, and F. Rémy. Contribution of Alaskan glaciers to sea-level rise derived from satellite imagery. *Nature Geoscience*, 3 (2):92–95, 2010. Cité page 10.

S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland. Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613, 2011. Cité page 113.

B. Biggio, B. Nelson, and P. Laskov. Support Vector Machines under adversarial label noise. *Asian Conference on Machine Learning*, 20:97–112, 2011. Cité page 104.

T. Blaschke. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1):2–16, 2010. Cité pages 68 et 69.

S. Bojinski, M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, and M. Zemp. The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, 95(9):1431–1443, 2014. Cité page 4.

S. Bontemps, M. Boettcher, C. Brockmann, G. Kirches, C. Lamarche, J. Radoux, M. Santoro, E. Vanbogaert, U. Wegmüller, M. Herold, F. Achard, F. Ramoino, O. Arino, and P. Defourny. Multi-year global land cover mapping at 300 m and characterization for climate modelling: Achievements of the Land Cover component of the ESA Climate Change Initiative. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(7):323, 2015. Cité page 20.

M.-R. Bouguelia. *Classification et apprentissage actif à partir d'un flux de données évolutif en présence d'étiquetage incertain*. PhD thesis, Université de Lorraine, 2015. Cité page 211.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König. Overview of Random Forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Data Mining and Knowledge Discovery*, 2(6):493–507, 2012. Cité page 87.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. Cité page 41.

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. Cité pages 31, 45 , et 46.

L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and regression trees*. Taylor & Francis, 1984. Cité pages 43 et 46.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *ACM International Conference on Management of Data (SIGMOD)*, volume 29, pages 93–104. ACM, 2000. Cité pages 129 et 146.

G. J. Briem, J. A. Benediktsson, and J. R. Sveinsson. Multiple classifiers applied to multisource remote sensing data. *IEEE Transactions on Geoscience and Rremote Sensing*, 40(10):2291–2299, 2002. Cité page 31.

C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999. Cité pages 96, 102, 104, 160, 161, 162, 163, 171 , et 203.

C. E. Brodley, M. A. Friedl, et al. Identifying and eliminating mislabeled training instances. In *American Association for Artificial Intelligence (AAAI) / Innovative Applications of Artificial Intelligence (IAAI)*, pages 799–805. American Association for Artificial Intelligence, Menlo Park, CA (United States), 1996. Cité page 104.

G. Brown, J. Wyatt, R. Harris, and X. Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005. Cité page 40.

P. D. Broxton, X. Zeng, D. Sulla-Menashe, and P. A. Troch. A global land cover climatology using MODIS data. *Journal of Applied Meteorology and Climatology*, 53(6): 1593–1605, 2014. Cité page 21.

R. Bryll, R. Gutierrez-Osuna, and F. Quek. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition*, 36(6):1291–1302, 2003. Cité page 41.

T. Büschenfeld and J. Ostermann. Automatic refinement of training data for classification of satellite imagery. *ISPRS Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences*, pages 1–7, 2012. Cité page 123.

G. Büttner. CORINE land cover and land cover change products. In *Land Use and Land Cover Mapping in Europe*, pages 55–74. Springer, 2014. Cité page 15.

G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, pages 1–37, 2015. Cité page 138.

G. Camps-Valls, D. Tuia, L. Gomez-Chova, S. Jimenez, and J. Malo. *Remote Sensing Image Processing. Synthesis Lectures on Image, Video, and Multimedia Processing.* Morgan and Claypool, 2011. Cité page 68.

J. Cao, S. Kwong, and R. Wang. A noise-detection based AdaBoost algorithm for mislabeled data. *Pattern Recognition*, 45(12):4451–4465, 2012. Cité page 96.

M. J. Carlotto. Effect of errors in ground truth on classification accuracy. *International Journal of Remote Sensing*, 30(18):4831–4849, 2009. Cité page 97.

H. Carrão, P. Gonçalves, and M. Caetano. Contribution of multispectral and multitemporal information from MODIS images to land cover classification. *Remote Sensing of Environment*, 112(3):986–997, 2008. Cité page 31.

R. Caruana and A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *International Conference on Machine Learning (ICML)*, pages 161–168. ACM, 2006. Cité page 32.

G. C. Cawley and N. L. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010. Cité page 114.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15, 2009. Cité page 121.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. Cité page 39.

O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (Chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. Cité page 211.

N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*, pages 853–867. Springer, 2005. Cité page 77.

C. Chen, A. Liaw, and L. Breiman. Using Random Forest to learn imbalanced data. *University of California, Berkeley*, 110, 2004. Cité pages 77 et 164.

J. Chen, J. Chen, A. Liao, X. Cao, L. Chen, X. Chen, C. He, G. Han, S. Peng, M. Lu, W. Zhang, W. Tong, and J. Mills. Global land cover mapping at 30m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:7–27, 2015. Cité page 20.

P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989. Cité page 96.

M. Claverie, V. Demarez, B. Duchemin, O. Hagolle, D. Ducrot, C. Marais Sicre, J.-F. Dejoux, M. Huc, P. Keravec, P. Béziat, R. Fieuzal, E. Ceschia, and G. Dedieu. Maize and sunflower biomass estimation in southwest France using high spatial and temporal resolution remote sensing data. *Remote Sensing of Environment*, 124:844–857, 2012. Cité page 4.

R. R. Colditz. An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sensing*, 7(8):9655–9681, 2015. Cité page 77.

A. Comber, P. Fisher, C. Brunsdon, and A. Khmag. Spatial analysis of remote sensing image classification accuracy. *Remote Sensing of Environment*, 127:237–246, 2012. Cité page 49.

R. G. Congalton and K. Green. *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press, 2008. Cité pages 47, 77 , et 211.

C. Cortes and V. N. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995. Cité page 36.

E. P. Crist and R. C. Cicone. A physically-based transformation of Thematic Mapper data—The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote Sensing*, GE-22(3):256–263, 1984. Cité pages 70 et 71.

D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. Random Forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007. Cité page 87.

M. Dalla Mura, J. Benediktsson, B. Waske, and L. Bruzzone. Morphological attribute profiles for the analysis of very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(10):3747–3762, 2010. Cité page 68.

M. Dalponte, H. O. Orka, T. Gobakken, D. Gianelle, and E. Næsset. Tree species classification in boreal forests with hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 51(5):2632–2645, 2013. Cité page 82.

J. Dash, A. Mathur, G. M. Foody, P. J. Curran, J. W. Chipman, and T. M. Lillesand. Land cover classification using multi-temporal MERIS vegetation indices. *International Journal of Remote Sensing*, 28(6):1137–1159, 2007. Cité page 31.

P. Defourny, C. Vancutsem, P. Bicheron, C. Brockmann, F. Nino, L. Schouten, and M. Leroy. GlobCover: a 300 m global land cover product for 2005 using ENVISAT MERIS time series. In *Proceedings of the ISPRS Commission VII mid-term symposium, Remote sensing: from pixels to processes*, pages 8–11. Citeseer, 2006. Cité page 20.

R. S. DeFries and J. R. G. Townshend. NDVI-derived land cover classifications at a global scale. *International Journal of Remote Sensing*, 15(17):3567–3586, 1994. Cité pages 30 et 70.

P. A. Devijver and J. Kittler. On the edited nearest neighbor rule. In *IEEE International Conference on Pattern Recognition*, pages 72–80, 1980. Cité page 165.

A. Di Gregorio and L. J. Jansen. *Land cover classification system (LCCS): classification concepts and user manual for software version 1.0*. Food and Agriculture Organization of the United Nations, 2000. Cité page 252.

T. G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000a. Cité pages 31 et 40.

T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000b. Cité page 96.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: ESA's Optical High-Resolution Mission for {GMES} Operational Services. *Remote Sensing of Environment*, 120:25–36, 2012. Cité page 10.

P. Du, J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu. Multiple classifier system for remote sensing image classification: A review. *Sensors*, 12(4):4764–4792, 2012. Cité page 31.

D. C. Duro, S. E. Franklin, and M. G. Dubé. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*, 118:259–272, 2012. Cité page 82.

H. Eerens, D. Haesen, F. Rembold, F. Urbano, C. Tote, and L. Bydekerke. Image time series processing for agriculture monitoring. *Environmental Modelling & Software*, 53: 154–162, 2014. Cité page 72.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994. Cité page 41.

C. Englund and A. Verikas. A novel approach to estimate proximity in a Random Forest: An exploratory study. *Expert Systems with Applications*, 39(17):13046–13050, 2012. Cité pages 133 et 155.

A. Estabrooks, T. Jo, and N. Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, 2004. Cité page 77.

J. Estima and M. Painho. Exploratory analysis of OpenStreetMap for land use classification. In *ACM International Workshop on Crowdsourced and Volunteered Geographic Information (SIGSPATIAL)*, pages 39–46. ACM, 2013. Cité page 18.

H. D. Eva, A. S. Belward, E. E. De Miranda, C. M. Di Bella, V. Gond, O. Huber, S. Jones, M. Sgrenzaroli, and S. Fritz. A land cover map of South America. *Global Change Biology*, 10(5):731–744, 2004. Cité page 27.

M. Fauvel. Kernel matrix approximation for learning the kernel hyperparameters. In *IEEE International Geoscience and Remote Sensing Symposium 2012 (IGARSS)*, pages 5418–5421. IEEE, 2012. Cité page 38.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. Cité page 138.

J. J. Feddema, K. W. Oleson, G. B. Bonan, L. O. Mearns, L. E. Buja, G. A. Meehl, and W. M. Washington. The importance of land-cover change in simulating future climates. *Science*, 310(5754):1674–1678, 2005. Cité page 4.

W. Feng, S. Boukir, and L. Guo. Identification and correction of mislabeled training data for land cover classification based on ensemble margin. In *IEEE International Geoscience and Remote Sensing Symposium 2015 (IGARSS)*, pages 4991–4994. IEEE, 2015. Cité page 163.

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181, 2014. Cité page 32.

J. I. Fisher, J. F. Mustard, and M. A. Vadeboncoeur. Green leaf phenology at Landsat resolution: Scaling from the field to the satellite. *Remote Sensing of Environment*, 100 (2):265–279, 2006. Cité page 98.

A. Folleco, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Identifying learners robust to low quality data. *Informatica (Slovenia)*, 33(3):245–259, 2009. Cité pages 96 et 113.

G. M. Foody. On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering % Remote Sensing*, 58(10):1459–1460, 1992. Cité page 48.

G. M. Foody. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1):185–201, 2002. Cité pages 47, 95 , et 211.

G. M. Foody. Classification accuracy comparison: hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, 113(8):1658–1663, 2009. Cité page 49.

G. M. Foody. Ground reference data error and the mis-estimation of the area of land cover change as a function of its abundance. *Remote Sensing Letters*, 4(8):783–792, 2013. Cité page 211.

G. M. Foody and A. Mathur. A relative evaluation of multiclass image classification by Support Vector Machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42 (6):1335–1343, 2004a. Cité page 30.

G. M. Foody and A. Mathur. Toward intelligent training of supervised image classifications: directing training data acquisition for svm classification. *Remote Sensing of Environment*, 93(1):107–117, 2004b. Cité page 30.

G. M. Foody and A. Mathur. The use of small training sets containing mixed pixels for accurate hard image classification: Training on mixed spectral responses for classification by a SVM. *Remote Sensing of Environment*, 103(2):179–189, 2006. Cité page 30.

G. M. Foody, M. Pal, D. Rocchini, C. X. Garzon-Lopez, and L. Bastin. The sensitivity of mapping methods to reference data quality: training supervised image classifications with imperfect reference data. *ISPRS International Journal of Geo-Information*, 5(11): 199, 2016. Cité pages 31, 82 , et 115.

J. Franklin. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*, 9(5):733–748, 1998. Cité page 69.

S. E. Franklin and M. A. Wulder. Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. *Progress in Physical Geography*, 26(2):173–205, 2002. Cité page 20.

B. Frénay and M. Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. Cité pages 96, 120, 121, 158 , et 171.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning (ICML)*, volume 96, pages 148–156. ACM, 1996. Cité page 41.

B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007. Cité page 127.

M. A. Friedl, D. K. McIver, J. C. F. Hodges, X. Y. Zhang, D. Muchoney, A. H. Strahler, C. E. Woodcock, S. Gopal, A. Schneider, A. Cooper, A. Baccini, F. Gao, and C. Schaaf. Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 83(1):287–302, 2002. Cité page 31.

M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 114(1):168–182, 2010. Cité pages 20, 21 , et 31.

S. Fritz, I. McCallum, C. Schill, C. Perger, R. Grillmayer, F. Achard, F. Kraxner, and M. Obersteiner. Geo-Wiki.Org: The use of crowdsourcing to improve global land cover. *Remote Sensing*, 1(3):345–354, 2009. Cité page 18.

S. Fritz, I. McCallum, C. Schill, C. Perger, L. See, D. Schepaschenko, M. van der Velde, F. Kraxner, and M. Obersteiner. Geo-Wiki: An online platform for improving global land cover. *Environmental Modelling & Software*, 31:110–123, 2012. Cité page 18.

D. Gamberger and N. Lavrač. Conditions for occam's razor applicability and noise elimination. *European Conference on Machine Learning (ECML)*, pages 108–123, 1997. Cité pages 123 et 165.

D. Gamberger, N. Lavrač, and C. Groselj. Experiments with noise filtering in a medical domain. In *International Conference on Machine Learning (ICML)*, pages 143–151. ACM, 1999. Cité pages 123, 162, 166 , et 171.

D. Gamberger, N. Lavrač, and S. Dzeroski. Noise detection and elimination in data preprocessing: experiments in medical domains. *Applied Artificial Intelligence*, 14(2): 205–223, 2000. Cité page 158.

B.-C. Gao. NDWI – a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996. Cité page 71.

F. Gao, J. Masek, M. Schwaller, and F. Hall. On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8):2207–2218, 2006. Cité page 98.

L. P. F. Garcia, A. C. P. de Carvalho, and A. C. Lorena. Effect of label noise in the complexity of classification problems. *Neurocomputing*, 160:108–119, 2015. Cité pages 97, 131, 136 , et 160.

GCOS. The global observing system for climate: Implementation needs. *Global Climate Observing System Implementation Plan 2016*, 2016. Cité pages 4 et 7.

S. Gebhardt, T. Wehrmann, M. A. M. Ruiz, P. Maeda, J. Bishop, M. Schramm, R. Kopeinig, O. Cartus, J. Kellndorfer, R. Ressl, L. A. Santos, and M. Schmidt. Madmex: automatic wall-to-wall land cover monitoring for the mexican redd-mrv program using all landsat data. *Remote Sensing*, 6(5):3923–3943, 2014. Cité page 31.

R. Genuer. *Forêts aléatoires: aspects théoriques, sélection de variables et applications*. PhD thesis, Université Paris Sud-Paris XI, 2010. Cité page 45.

R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using Random Forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010. Cité page 45.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. Cité page 46.

B. Ghimire, J. Rogan, V. F. Rodríguez Galiano, P. Panday, and N. Neeti. An evaluation of bagging, boosting, and Random Forests for land-cover classification in Cape Cod, Massachusetts, USA. *GIScience & Remote Sensing*, 49(5):623–643, 2012. Cité page 31.

A. Ghosh, F. E. Fassnacht, P. Joshi, and B. Koch. A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation*, 26:49–63, 2014. Cité page 82.

P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson. Random Forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300, 2006. Cité page 31.

M. Goldstein and A. Dengel. Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and Demo Track*, pages 59–63, 2012. Cité page 121.

M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016. Cité pages 139, 141 , et 146.

C. Gómez, J. C. White, and M. A. Wulder. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:55–72, 2016. Cité pages 32 et 69.

P. Gong, J. Wang, L. Yu, Y. Zhao, Y. Zhao, L. Liang, Z. Niu, X. Huang, H. Fu, S. Liu, C. Li, X. Li, W. Fu, C. Liu, Y. Xu, X. Wang, Q. Cheng, L. Hu, W. Yao, H. Zhang, P. Zhu, Z. Zhao, H. Zhang, Y. Zheng, L. Ji, Y. Zhang, H. Chen, A. Yan, J. Guo, L. Yu, L. Wang, X. Liu, T. Shi, M. Zhu, Y. Chen, G. Yang, P. Tang, B. Xu, C. Giri, N. Clinton, Z. Zhu, J. Chen, and J. Chen. Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing*, 34(7):2607–2654, 2013. Cité pages 20, 27 , et 32.

S. Gopal, C. E. Woodcock, and A. H. Strahler. Fuzzy neural network classification of global land cover from a 1° AVHRR data set. *Remote Sensing of Environment*, 67(2): 230–243, 1999. Cité page 30.

G. Grekousis, G. Mountrakis, and M. Kavouras. An overview of 21 global and 43 regional land-cover mapping products. *International Journal of Remote Sensing*, 36(21):5309–5335, 2015. Cité page 20.

A. Gressin. *Mise à jour d'une base de données d'occupation du sol à grande échelle en milieux naturels à partir d'une image satellite THR*. PhD thesis, Université René Descartes-Paris V, 2014. Cité pages 14 et 63.

A. Gressin, C. Mallet, N. Vincent, and N. Paparoditis. Updating land cover databases using a single very high resolution satellite image. *The ISPRS Workshop on Image Sequence Analysis*, 2(3):13–18, 2013. Cité pages 68 et 82.

F. E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969. Cité page 120.

L. Guo. Margin framework for ensemble classifiers. application to remote sensing data. *Université de Bordeaux*, 2011. Cité page 41.

O. Hagolle, G. Dedieu, B. Mougenot, V. Debaecker, B. Duchemin, and A. Meygret. Correction of aerosol effects on multi-temporal images acquired with constant viewing angles: Application to Formosat-2 images. *Remote Sensing of Environment*, 112(4):1689–1701, 2008. Cité page 57.

O. Hagolle, M. Huc, D. Villa Pascual, and G. Dedieu. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENUS, LANDSAT and SENTINEL-2 images. *Remote Sensing of Environment*, 114(8):1747–1755, 2010. Cité page 57.

O. Hagolle, M. Huc, D. Villa Pascual, and G. Dedieu. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of FormoSat-2, LandSat, VENUS and Sentinel-2 images. *Remote Sensing*, 7(3):2668, 2015a. Cité page 61.

O. Hagolle, S. Sylvander, M. Huc, M. Claverie, D. Clesse, C. Dechoz, V. Lonjou, and V. Poulain. SPOT-4 (Take 5): simulation of Sentinel-2 time series on 45 large sites. *Remote Sensing*, 7(9):12242–12264, 2015b. Cité pages 10 et 52.

J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011. Cité pages 79, 105, 121 , et 123.

M. C. Hansen. Classification trees and mixed pixel training data. In *Remote Sensing of Land Use and Land Cover: Principles and applications*. Taylor & Francis, 2012. Cité page 31.

M. C. Hansen, R. S. DeFries, J. R. G. Townshend, and R. Sohlberg. Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, 21(6-7):1331–1364, 2000. Cité page 20.

R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979. Cité page 68.

R. C. Hasan, D. Ierodiaconou, and J. Monk. Evaluation of four supervised learning methods for benthic habitat mapping using backscatter from multi-beam sonar. *Remote Sensing*, 4(11):3427–3443, 2012. Cité page 82.

D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980. Cité pages 120 et 121.

S. Hawkins, H. He, G. Williams, and R. Baxter. Outlier detection using replicator neural networks. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 170–180. Springer, 2002. Cité page 122.

Z. He, S. Deng, and X. Xu. Outlier detection integrating semantic knowledge. In *International Conference on Web-Age Information Management*, pages 126–131. Springer, 2002. Cité page 128.

Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9):1641–1650, 2003. Cité page 131.

Z. He, X. Xu, J. Z. Huang, and S. Deng. Mining class outliers: concepts, algorithms and applications in crm. *Expert Systems with Applications*, 27(4):681–697, 2004. Cité page 131.

N. M. Hewahi and M. K. Saad. Class outliers mining: Distance-based approach. *International Journal of Intelligent Systems and Technologies*, 2:5, 2007. Cité pages 129, 160 , et 167.

T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. Cité page 41.

V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004. Cité page 121.

R. A. Houghton, J. I. House, J. Pongratz, G. R. van der Werf, R. S. DeFries, M. C. Hansen, C. L. Quéré, and N. Ramankutty. Carbon emissions from land use and land-cover change. *Biogeosciences*, 9(12):5125–5142, 2012. Cité page 4.

C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass Support Vector Machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002. Cité page 39.

C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. 2003. Cité page 38.

C. Huang, L. S. Davis, and J. R. G. Townshend. An assessment of Support Vector Machines for land cover classification. *International Journal of Remote Sensing*, 23(4): 725–749, 2002a. Cité pages 31 et 71.

C. Huang, B. Wylie, L. Yang, C. Homer, and G. Zylstra. Derivation of a tasselled cap transformation based on Landsat 7 at-satellite reflectance. *International Journal of Remote Sensing*, 23(8):1741–1748, 2002b. Cité page 70.

X. Huang and L. Zhang. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):257–272, 2013. Cité page 68.

L. Hubert-Moy, A. Cotonnec, L. Le Du, A. Chardin, and P. Pérez. A comparison of parametric classification procedures of remotely sensed data applied on different landscape units. *Remote Sensing of Environment*, 75(2):174–187, 2001. Cité page 30.

G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968. Cité pages 30 et 31.

N. P. Hughes, S. J. Roberts, and L. Tarassenko. Semi-supervised learning of probabilistic models for ECG segmentation. In *International Conference onEngineering in Medicine and Biology Society (IEMBS)*, volume 1, pages 434–437. IEEE, 2004. Cité page 211.

J. Inglada. PhenOTB, Phenological analysis for image time series, 2016. Cité page 72.

J. Inglada, M. Arias, B. Tardy, O. Hagolle, S. Valero, D. Morin, G. Dedieu, G. Sepulcre, S. Bontemps, P. Defourny, and B. Koetz. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9):12356–12379, 2015. Cité pages 4, 33 , et 60.

J. Inglada, A. Vincent, M. Arias, and C. Marais Sicre. Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. *Remote Sensing*, 8(5):362, 2016. Cité page 210.

J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1):95, 2017. Cité pages 6, 20, 21, 31 , et 91.

IPCC. Intergovernmental panel on climate change. *Climate change*, 2014. Cité page 4.

J. H. M. Janssens. Outlier selection and one-class classification. 2013. Cité pages 120, 127 , et 128.

K. Jia, S. Liang, X. Wei, Y. Yao, Y. Su, B. Jiang, and X. Wang. Land cover classification of Landsat data with phenological features extracted from time series MODIS NDVI data. *Remote Sensing*, 6(11):11518–11532, 2014a. Cité pages 31, 72, 123 , et 165.

K. Jia, S. Liang, X. Wei, Y. Yao, Y. Su, B. Jiang, and X. Wang. Land cover classification of Landsat data with phenological features extracted from time series MODIS NDVI data. *Remote Sensing*, 6(11):11518–11532, 2014b. Cité page 30.

G. H. John. Robust decision trees: Removing outliers from databases. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 174–179. American Association for Artificial Intelligence (AAAI), 1995. Cité page 96.

B. A. Johnson and K. Iizuka. Integrating OpenStreetMap crowdsourced data and landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines. *Applied Geography*, 67:140–149, 2016. Cité pages 18 et 31.

D. Joly, T. Brossard, H. Cardot, J. Cavailhes, M. Hilal, and P. Wavresky. Les types de climats en france, une construction spatiale. *Cybergeo : European Journal of Geography*, 2010. Cartographie, Imagerie, SIG. Cité page 54.

P. Jönsson and L. Eklundh. Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE Transactions on Geoscience and Remote Sensing*, 40(8): 1824–1832, 2002. Cité page 72.

P. Jönsson and L. Eklundh. TIMESAT - a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, 30(8):833–845, 2004. Cité pages 72 et 98.

M. N. Kapp, R. Sabourin, and P. Maupin. An empirical study on diversity measures and margin theory for ensembles of classifiers. In *0th International Conference on Information Fusion*, pages 1–8. IEEE, 2007. Cité page 40.

A. Karmaker and S. Kwek. A boosting approach to remove class label noise1. *International Journal of Hybrid Intelligent Systems*, 3(3):169–177, 2006. Cité page 96.

R. J. Kauth and G. S. Thomas. The tasselled cap - a graphic description of the spectral-temporal development of agricultural crops as seen by landsat. In *LARS Symposia*, page 159, 1976. Cité pages 70 et 71.

M. Khalilia, S. Chakraborty, and M. Popescu. Predicting disease risks from highly imbalanced data using Random Forest. *BMC medical informatics and decision making*, 11(1):51, 2011. Cité page 81.

R. Khatami, G. Mountrakis, and S. V. Stehman. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, 177: 89–100, 2016. Cité pages 20, 22, 28 , et 32.

T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse. An empirical study of learning from imbalanced data using Random Forest. In *IEEE International Conference on Tools with Artificial Intelligence 2007 (ICTAI)*, volume 2, pages 310–317. IEEE, 2007. Cité pages 76, 161 , et 166.

E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer, 1998. Cité pages xii, 123, 124, 125 , et 129.

A. Kolcz and G. V. Cormack. Genre-based decomposition of email class noise. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 427–436. ACM, 2009. Cité page 102.

J. Koplowitz and T. A. Brown. On the relation of performance to editing in nearest neighbor rules. *Pattern Recognition*, 13(3):251–255, 1981. Cité page 163.

H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. Interpreting and unifying outlier scores. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 13–24. SIAM, 2011. Cité page 128.

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. Cité page 133.

L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003. Cité page 40.

N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017. Cité page 210.

V. Labatut and H. Cherifi. Accuracy measures for the comparison of classifiers. *arXiv preprint arXiv:1207.3790*, 2012. Cité page 49.

S. Lallich, F. Muhlenbach, and D. A. Zighed. Improving classification by removing or relabeling mislabeled instances. *Foundations of Intelligent Systems*, pages 5–15, 2002. Cité page 163.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, (1):159–174, 1977. Cité page 48.

P. Lassalle, J. Inglada, J. Michel, M. Grizonnet, and J. Malik. A scalable tile-based framework for region-merging segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5473–5485, 2015. Cité page 31.

R. L. Lawrence, S. D. Wood, and R. L. Sheley. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomforest). *Remote Sensing of Environment*, 100(3):356–362, 2006. Cité page 191.

K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly Of Applied Mathematics*, 1944. Cité page 72.

C. Li, J. Wang, L. Wang, and P. Hu, L.and Gong. Comparison of classification algorithms and training sample sizes in urban land classification with landsat thematic mapper imagery. *Remote Sensing*, 6(2):964–983, 2014. Cité page 30.

A. Liaw and M. Wiener. Classification and regression by Random Forest. *R News*, 2(3): 18–22, 2002. Cité page 87.

H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for Support Vector Machines. *Machine Learning*, 68(3):267–276, 2007. Cité page 40.

W.-J. Lin and J. J. Chen. Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, 14(1):13–26, 2013. Cité page 81.

C. Liu, P. Frazier, and L. Kumar. Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107(4):606–616, 2007. Cité page 48.

F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):3, 2012. Cité pages 131 et 132.

X. Liu, G. Cheng, and J. X. Wu. Analyzing outliers cautiously. *IEEE Transactions on Knowledge and Data Engineering*, 14(2):432–437, 2002. Cité page 120.

T. R. Loveland, B. C. Reed, J. F. Brown, D. O. Ohlen, Z. Zhu, L. Yang, and J. W. Merchant. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, 21(6-7): 1303–1330, 2000. Cité pages 20 et 27.

T. R. Loveland, J. F. Brown, D. O. Ohlen, B. C. Reed, Z. Zhu, L. Yang, and S. Howard. ISLSCP II IGBP DISCover and SiB Land Cover, 1992-1993, 2009. URL daac.ornl.gov. In Hall, Forrest G., G. Collatz, B. Meeson, S. Los, E. Brown de Colstoun, and D. Landis (eds.). ISLSCP Initiative II Collection. Data set. Available on-line. Cité page 20.

F. Löw, U. Michel, S. Dech, and C. Conrad. Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using Support Vector Machines. *ISPRS Journal of Photogrammetry and Remote Sensing*, 85:102–119, 2013. Cité page 82.

R. Lucas, A. Rowlands, A. Brown, S. Keyworth, and P. Bunting. Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping. *ISPRS Journal of photogrammetry and remote sensing*, 62(3):165–185, 2007. Cité page 69.

Z. Lv, P. Zhang, J. Benediktsson, and W. Shi. Morphological profiles based on differently shaped structuring elements for classification of images with very high spatial resolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(12):4644–4652, 2014. Cité page 68.

E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Convolutional Neural Networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, 2017. Cité page 33.

E. Maire, C. Marais Sicre, S. Guillerme, F. Rhoné, J.-F. Dejoux, and G. Dedieu. Télédétection de la trame verte arborée en haute résolution par morphologie mathématique. *Revue Internationale de Géomatique*, 22(4):519–538, 2012. Cité page 4.

Z. Malenovskỳ, H. Rott, J. Cihlar, M. E. Schaepman, G. García-Santos, R. Fernandes, and M. Berger. Sentinels for science: Potential of Sentinel-1,-2, and-3 missions for scientific observations of ocean, cryosphere, and land. *Remote Sensing of Environment*, 120: 91–101, 2012. Cité page 10.

L. M. Manevitz and M. Yousef. One-class SVMs for document classification. *Journal of Machine Learning Research*, 2:139–154, 2001. Cité page 122.

C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008. Cité page 49.

D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(2):431–441, 1963. Cité page 73.

J. F. Mas and J. J. Flores. The application of Artificial Neural Networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*, 29(3):617–663, 2007. Cité page 30.

A. Masse. *Développement et automatisation de méthodes de classification à partir de séries temporelles d'images de télédétection: application aux changements d'occupation des sols et à l'estimation du bilan carbone*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2013. Cité page 4.

S. K. McFeeters. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7): 1425–1432, 1996. Cité pages 70 et 71.

A. Mellor and S. Boukir. Exploring diversity in ensemble classification: Applications in large area land cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 129:151–161, 2017. Cité page 40.

A. Mellor, S. Boukir, A. Haywood, and S. Jones. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing*, 105:155–168, 2015. Cité pages 41, 77, 97, 102 , et 104.

G. Menardi and N. Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, pages 1–31, 2014. Cité page 76.

H. Meyer, M. Kühnlein, T. Appelhans, and T. Nauss. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. *Atmospheric Research*, 169:424–433, 2016. Cité pages 82 et 85.

X. Miao, J. S. Heaton, S. Zheng, D. A. Charlet, and H. Liu. Applying tree-based ensemble algorithms to the classification of ecological zones using multi-temporal multi-source remote-sensing data. *International Journal of Remote Sensing*, 33(6):1823–1849, 2012. Cité page 31.

J. Milgram, M. Cheriet, and R. Sabourin. "one against one" or "one against all": Which one is better for handwriting recognition with SVMs? In *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006. Cité page 39.

G. Mountrakis, J. Im, and C. Ogole. Support Vector Machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011. Cité page 31.

M. Mróz and A. Sobieraj. Comparison of several vegetation indices calculated on the basis of a seasonal SPOT XS time series, and their suitability for land cover and agricultural crop identification. *Technical Sciences*, 7:39–66, 2004. Cité page 70.

F. Muhlenbach, S. Lallich, and D. A. Zighed. Identifying and handling mislabelled instances. *Journal of Intelligent Information Systems*, 22(1):89–109, 2004. Cité page 163.

K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181, 2001. Cité pages 31 et 37.

K. V. S. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998. Cité page 43.

N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204. Curran Associates, Inc., 2013. Cité page 97.

P. Neis and D. Zielstra. Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet*, 6(1):76–106, 2014. Cité page 18.

T. Nery, R. Sadler, M. Solis-Aulestia, B. White, M. Polyakov, and M. Chalak. Comparing supervised algorithms in Land Use and Land Cover classification of a Landsat time-series. In *IEEE International Geoscience and Remote Sensing Symposium 2016 (IGARSS)*, pages 5165–5168, 2016. Cité page 82.

D. F. Nettleton, A. Orriols-Puig, and A. Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4):275–306, 2010. Cité pages 94 et 96.

L. Nezvalová, L. Popelínskỳ, L. Torgo, and K. Vaculík. Class-based outlier detection: Staying zombies or awaiting for resurrection? In *International Symposium on Intelligent Data Analysis*, pages 193–204. Springer, 2015. Cité pages 128 et 155.

S. Oliveira, F. Oehler, J. San-Miguel-Ayanz, A. Camia, and J. M. Pereira. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *Forest Ecology and Management*, 275:117–129, 2012. Cité page 87.

J. Osman, J. Inglada, and J.-F. Dejoux. Assessment of a markov logic model of crop rotations for early crop mapping. *Computers and Electronics in Agriculture*, 113:234–243, 2015. Cité page 69.

M. Paget, A. Gressin, and C. Mallet. Multi-temporal optical VHR image fusion for land-cover mapping. In *IEEE International Geoscience and Remote Sensing Symposium 2015 (IGARSS)*, pages 1913–1916, 2015. Cité page 68.

M. Pal. Random Forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005. Cité pages 31, 82 , et 85.

M. Pal and P. M. Mather. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, 86(4):554–565, 2003. Cité page 31.

H. Pang, A. Lin, M. Holford, B. E. Enerson, B. Lu, M. P. Lawton, E. Floyd, and H. Zhao. Pathway analysis using Random Forests classification and regression. *Bioinformatics*, 22(16):2028–2036, 2006. Cité page 160.

S. Papadimitriou and C. Faloutsos. Cross-outlier detection. *Advances in Spatial and Temporal Databases*, pages 199–213, 2003. Cité page 128.

M. Pechenizkiy, A. Tsymbal, S. Puuronen, and O. Pechenizkiy. Class noise and supervised learning in medical domains: The effect of feature extraction. In *IEEE Symposium on Computer-Based Medical Systems 2006 (CBMS)*, pages 708–713. IEEE, 2006. Cité page 97.

F. Petitjean. *Dynamic Time Warping : Apports théoriques pour l'analyse de données temporelles. Application à la classification de séries temporelles d'images satellites.* PhD thesis, Université de Strasbourg, 2012. Cité page 148.

R. A. Pielke. Land use and climate change. *Science*, 310(5754):1625–1626, 2005. Cité page 4.

M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. Cité page 121.

J. Piper. Variability and bias in experimentally measured classifier error rates. *Pattern Recognition Letters*, 13(10):685–692, 1992. Cité page 30.

K. Pittman, M. C. Hansen, I. Becker-Reshef, P. V. Potapov, and C. O. Justice. Estimating global cropland extent with multi-year modis data. *Remote Sensing*, 2(7):1844–1863, 2010. Cité page 72.

J. Platt. Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999. Cité page 40.

R. G. Pontius Jr and M. Millones. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429, 2011. Cité page 48.

T. Postadijan, A. Le Bris, H. Sahbi, and C. Mallet. Investigating the potential of deep neural networks for large-scale classification of very high resolution satellite images. 2017. Cité page 33.

A. B. Potgieter, A. Apan, P. Dunn, and G. Hammer. Estimating crop area using seasonal time series of Enhanced Vegetation Index from MODIS satellite imagery. *Australian Journal of Agricultural Research*, 58(4):316–325, 2007. Cité page 68.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986. Cité page 43.

J. R. Quinlan. *C4. 5: Programs for Machine Learning*, volume 1. Morgan Kaufmann, 1993. Cité page 43.

J. Radoux, C. Lamarche, E. Van Bogaert, S. Bontemps, C. Brockmann, and P. Defourny. Automated training sample extraction for global land cover mapping. *Remote Sensing*, 6(5):3965, 2014. Cité page 30.

S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In *ACM International Conference on Management of Data (SIGMOD)*, volume 29, pages 427–438, 2000. Cité pages xii, 125, 126, 160 , et 167.

U. Rebbapragada. *Strategic targeting of outliers for expert review*. PhD thesis, Tufts University, 2010. Cité page 211.

U. Rebbapragada and C. E. Brodley. Class noise mitigation through instance weighting. In *European Conference on Machine Learning (ECML)*, pages 708–715. Springer, 2007. Cité pages 104, 123 , et 163.

P. Rebours. *Partitioning filter approach to noise elimination: An empirical study in software quality classification*. PhD thesis, Florida Atlantic University, Boca Raton, FL, 2004. Cité page 166.

M. Robnik-Sikonja. Improving Random Forests. In *European Conference on Machine Learning (ECML)*, volume 3201, pages 359–370. Springer, 2004. Cité pages 43 et 46.

V. F. Rodríguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. An assessment of the effectiveness of a Random Forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93 – 104, 2012. Cité pages 20, 31, 68, 69, 87, 113, 160 , et 191.

J. Rogan, J. Franklin, and D. A. Roberts. A comparison of methods for monitoring multitemporal vegetation change using Thematic Mapper imagery. *Remote Sensing of Environment*, 80(1):143–156, 2002. Cité pages 20 et 68.

L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, 2010. Cité page 40.

L. Rokach. Decision forest: Twenty years of research. *Information Fusion*, 27:111–125, 2016. Cité page 46.

J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering. Monitoring Vegetation Systems in the Great Plains with ERTS. *Third ERTS Symposium, Washington, NASA*, pages 309–317, 1973. Cité page 71.

J. A. Sáez, M. Galar, J. Luengo, and F. Herrera. Tackling the problem of classification with noisy data using multiple classifier systems: Analysis of the performance and robustness. *Information Sciences*, 247:1–20, 2013. Cité page 96.

J. A. Sáez, M. Galar, J. Luengo, and F. Herrera. Analyzing the presence of noise in multi-class problems: alleviating its influence with the one-vs-one decomposition. *Knowledge and Information Systems*, 38(1):179–206, 2014. Cité page 94.

J. A. Sáez, M. Galar, J. Luengo, and F. Herrera. INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control. *Information Fusion*, 27: 19–32, 2016. Cité pages 136, 164 , et 166.

S. Sasikala, S. Bharathidason, and C. J. Venkateswaran. Improving classification accuracy based on random forest model through weighted sampling for noisy data with linear decision boundary. *Indian Journal of Science and Technology*, 8(S8):614–619, 2015. Cité page 46.

J. Scepan. Thematic validation of high-resolution global land-cover data sets. *Photogrammetric Engineering % Remote Sensing*, 65(9):1051–1060, 1999. Cité page 20.

R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Aannals of Sstatistics*, 26(5): 1651–1686, 1998. Cité page 40.

B. Schölkopf and A. J. Smola. *Learning with kernels: Support Vector Machines, regularization, optimization, and beyond*. MIT Press, 2002. Cité pages 37 et 38.

B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems (NIPS)*, pages 582–588, 2000. Cité page 122.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. Cité page 122.

E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On evaluation of outlier rankings and outlier scores. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 1047–1058. SIAM, 2012. Cité page 139.

E. Schubert, A. Zimek, and H.-P. Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*, 28(1):190–237, 2014. Cité page 130.

L. See, D. Schepaschenko, M. Lesiv, I. McCallum, S. Fritz, A. Comber, C. Perger, C. Schill, Y. Zhao, V. Maus, M. A. Sirajh, F. Albrechti, A. Ciprianij, M. Vakolyuka, A. Garcial, A. H. Rabiam, K. Singhan, A. A. Marcarinio, T. Kattenbornp, R. Hazarikas, M. Schepaschenkoq, M. van der Veldea, F. Kraxnera, and M. Obersteinera. Building a hybrid land cover map with crowdsourcing and geographically weighted regression. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:48–56, 2015. Cité page 18.

M. R. Segal. Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics*, 2004. Cité page 113.

C. Senf, P. J. Leitão, D. Pflugmacher, S. van der Linden, and P. Hostert. Mapping land cover in complex Mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery. *Remote Sensing of Environment*, 156:527–536, 2015. Cité page 70.

D. Sheeren, A. Masse, D. Ducrot, M. Fauvel, F. Collard, and S. May. La télédétection pour la cartographie de la trame verte en milieu agricole. *Revue internationale de Géomatique*, 22(4):539–563, 2012. Cité page 4.

T. Shi and S. Horvath. Unsupervised learning with Random Forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138, 2006. Cité page 133.

P. P. Shingare, P. M. Hemane, and D. S. Dandekar. Fusion classification of multispectral and panchromatic image using improved decision tree algorithm. In *International Conference on Signal Propagation and Computer Technology 2014 (ICSPCT)*, pages 598–603. IEEE, 2014. Cité page 71.

N. G. Silleos, T. K. Alexandridis, I. Z. Gitas, and K. Perakis. Vegetation indices: advances made in biomass estimation and vegetation monitoring in the last 30 years. *Geocarto International*, 21(4):21–28, 2006. Cité page 70.

B. Sluban. *Ensemble-based noise and outlier detection.* PhD thesis, Jozef Stephan Inernational Postgraduate School, 2014. Cité page 169.

B. Sluban, D. Gamberger, and N. Lavrač. Advances in class noise detection. In *European Conference on Artificial Intelligence 2010 (ECAI)*, pages 1105–1106. IOS Press, 2010. Cité page 162.

M. R. Smith and T. Martinez. Using classifier diversity to handle label noise. In *IEEE International Joint Conference on Neural Networks 2015 (IJCNN)*, pages 1–8. IEEE, 2015. Cité pages 136, 161, 162, 163 , et 164.

R. J. Stapenhurst. *Diversity, margins and non-stationary learning.* PhD thesis, University of Manchester, 2012. Cité page 41.

S. V. Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77–89, 1997. Cité page 48.

B. Sun, S. Chen, J. Wang, and H. Chen. A robust multi-class AdaBoost algorithm for mislabeled noisy data. *Knowledge-Based Systems*, 102:87–102, 2016. Cité page 96.

Y. Sun, A. K. C. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, 2009. Cité page 76.

T. Takenouchi, S. Eguchi, N. Murata, and T. Kanamori. Robust boosting algorithm against mislabeling in multiclass problems. *Neural Computation*, 20(6):1596–1630, 2008. Cité page 102.

R. Tateishi, N. T. Hoan, T. Kobayashi, B. Alsaaideh, G. Tana, and D. X. Phong. Production of global land cover data–GLCNMO2008. *Journal of Geography and Geology*, 6(3):99, 2014. Cité page 20.

K. Tatsumi, Y. Yamashiki, M. A. C. Torres, and C. L. R. Taipe. Crop classification of upland fields using Random Forest of time-series Landsat 7 ETM+ data. *Computers and Electronics in Agriculture*, 115:171–179, 2015. Cité page 87.

D. M. J. Tax and R. P. W. Duin. Support Vector Domain Description. *Pattern Recognition Letters*, 20(11):1191–1199, 1999. Cité page 122.

D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54 (1):45–66, 2004. Cité page 122.

C.-M. Teng. Correcting noisy data. In *International Conference on Machine Learning (ICML)*, pages 239–248. ACM, 1999. Cité pages 102 et 104.

C.-M. Teng. A comparison of noise handling techniques. In *FLAIRS Conference*, pages 269–273, 2001. Cité page 163.

J. Thomas, P.-E. Jouve, and N. Nicoloyannis. Optimisation and evaluation of Random Forests for imbalanced datasets. In *International Symposium on Methodologies for Intelligent Systems*, pages 622–631. Springer, 2006. Cité page 77.

I. Tomek. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):448–452, 1976. Cité pages 129 et 165.

R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown, I. N. Travera, P. Deghayea, B. Duesmanna, B. Rosicha, N. Mirandaa, C. Brunob, M. L'Abbateb, R. Crocib, A. Pietropaolob, M. Huchlerc, and F. Rostanc. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012. Cité page 10.

W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. F. T. van Hijum. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Briefings in Bioinformatics*, page bbs034, 2012. Cité page 160.

R. Trias-Sanz. Texture orientation and period estimator for discriminating between forests, orchards, vineyards, and tilled fields. *IEEE Transactions on Geoscience and Remote Sensing*, 44(10):2755–2760, 2006. Cité page 68.

S. Tsuji, Y. Midorikawa, T. Takahashi, K. Yagi, T. Takayama, K. Yoshida, Y. Sugiyama, and H. Aburatani. Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis. *British Journal of Cancer*, 106(1):126–132, 2012. Cité page 160.

A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Dynamic integration with Random Forests. In *European Conference on Machine Learning (ECML)*, pages 801–808. Springer, 2006. Cité page 46.

C. J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979. Cité page 70.

D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011. Cité page 211.

S. Valero, D. Morin, J. Inglada, G. Sepulcre, M. Arias, O. Hagolle, G. Dedieu, S. Bontemps, P. Defourny, and B. Koetz. Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. *Remote Sensing*, 8(1):55, 2016. Cité pages 4 et 72.

T. G. Van Niel, T. R. McVicar, and B. Datt. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. *Remote Sensing of Environment*, 98(4):468–480, 2005. Cité page 30.

V. N. Vapnik. *The nature of statistical learning theory.* Springer Science & Business Media, 1995. Cité page 35.

V. N. Vapnik. *Statistical learning theory.* Wiley New York, 1998. Cité page 35.

S. Verbaeten and A. Van Assche. Ensemble methods for noise elimination in classification problems. In *International Workshop on Multiple Classifier Systems*, pages 317–325. Springer, 2003. Cité pages 162 et 171.

M. A. Vieira, A. R. Formaggio, C. D. Rennó, C. Atzberger, D. A. Aguiar, and M. P. Mello. Object based image analysis and data mining applied to a remotely sensed landsat time-series to map sugarcane over large areas. *Remote Sensing of Environment*, 123:553–562, 2012. Cité page 69.

F. Waldner, G. S. Canto, and P. Defourny. Automated annual cropland mapping using knowledge-based temporal features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 110:1–13, 2015a. Cité pages 69 et 72.

F. Waldner, M.-J. Lambert, W. Li, M. Weiss, V. Demarez, D. Morin, C. Marais-Sicre, O. Hagolle, F. Baret, and P. Defourny. Land cover and crop type classification along the season based on biophysical variables retrieved from multi-sensor high-resolution time series. *Remote Sensing*, 7(8):10400–10424, 2015b. Cité page 69.

J. Wang, Y. Zhao, C. Li, L. Yu, D. Liu, and P. Gong. Mapping global land cover in 2001 and 2010 with spatial-temporal consistency at 250m resolution. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:38–47, 2015. Cité page 20.

B. Waske and J. A. Benediktsson. Fusion of Support Vector Machines for classification of multisensor data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12): 3858–3866, 2007. Cité pages 191 et 210.

B. Waske and S. van der Linden. Classifying multilevel imagery from SAR and optical sensors by decision fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 46 (5):1457–1466, 2008. Cité pages 82 et 210.

C. Weber and A. Puissant. Urbanization pressure and modeling of urban growth: example of the tunis metropolitan area. *Remote Sensing of Environment*, 86(3):341–352, 2003. Cité page 10.

G. M. Weiss and F. Provost. Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003. Cité page 77.

A. K. Whitcraft, I. Becker-Reshef, and C. O. Justice. A framework for defining spatially explicit earth observation requirements for a global agricultural monitoring initiative (GEOGLAM). *Remote Sensing*, 7(2):1461–1481, 2015. Cité page 4.

G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu. A comparative study of RNN for outlier detection in data mining. In *IEEE International Conference on Data Mining 2002 (ICDM)*, pages 709–712. IEEE, 2002. Cité page 122.

D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, (3):408–421, 1972. Cité page 129.

E. H. Wilson and S. A. Sader. Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sensing of Environment*, 80(3):385–396, 2002. Cité page 71.

T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004. Cité page 40.

H. Xiao, H. Xiao, and C. Eckert. Adversarial label flips attack on Support Vector Machines. In *European Conference on Artificial Intelligence 2012 (ECAI)*, pages 870–875, 2012. Cité page 104.

H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, and F. Roli. Support Vector Machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015a. Cité page 97.

T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015b. Cité page 102.

X. Xiao, S. Boles, J. Liu, D. Zhuang, S. Frolking, C. Li, W. Salas, and B. Moore. Mapping paddy rice agriculture in southern china using multi-temporal MODIS images. *Remote Sensing of Environment*, 95(4):480–492, 2005. Cité page 68.

H. Xiong, G. Pandey, M. Steinbach, and V. Kumar. Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, 18(3):304–319, 2006. Cité page 120.

H. Xu. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14): 3025–3033, 2006. Cité page 71.

H. Xu. A new index for delineating built-up land features in satellite imagery. *International Journal of Remote Sensing*, 29(14):4269–4276, 2008. Cité page 71.

K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 320–324. ACM, 2000. Cité page 121.

J. Yeom, Y. Han, and Y. Kim. Separability analysis and classification of rice fields using KOMPSAT-2 high resolution satellite imagery. *Research Journal of Chemistry and Environment*, 17:12, 2013. Cité page 70.

L. Yu, J. Wang, N. Clinton, Q. Xin, L. Zhong, Y. Chen, and P. Gong. FROM-GC: 30 m global cropland extent derived through multisource data integration. *International Journal of Digital Earth*, 6(6):521–533, 2013a. Cité page 20.

L. Yu, J. Wang, and P. Gong. Improving 30 m global land-cover map FROM-GLC with time series MODIS and auxiliary data sets: a segmentation-based approach. *International Journal of Remote Sensing*, 34(16):5851–5867, 2013b. Cité page 20.

L. Yu, J. Wang, X. Li, C. Li, Y. Zhao, and P. Gong. A multi-resolution global land cover dataset through multisource data aggregation. *Science China Earth Sciences*, 57(10): 2317–2329, 2014. Cité page 20.

W. Yuan, D. Guan, Q. Zhu, and T. Ma. Novel mislabeled training data detection algorithm. *Neural Computing and Applications*, pages 1–11, 2016. Cité page 162.

X. Zeng and T. Martinez. An algorithm for correcting mislabeled data. *Intelligent Data Analysis*, 5(6):491–502, 2001. Cité page 163.

Y. Zha, J. Gao, and S. Ni. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24 (3):583–594, 2003. Cité pages 70 et 71.

X. Zhang, M. A. Friedl, C. B. Schaaf, A. H. Strahler, J. C. F. Hodges, F. Gao, B. C. Reed, and A. Huete. Monitoring vegetation phenology using MODIS. *Remote Sensing of Environment*, 84(3):471–475, 2003. Cité pages 72 et 98.

F. Zhou and A. Zhang. Optimal subset selection of time-series MODIS images and sample data transfer with Random Forests for supervised classification modelling. *Sensors*, 16 (11):1783, 2016. Cité page 160.

X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210, 2004. Cité pages 94, 95, 96 , et 104.

X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *International Conference on Machine Learning (ICML)*, pages 920–927. ACM, 2003. Cité pages 104, 162, 166 , et 171.

Acronymes

ACP Analyse par Composantes Principales. 63

ANN *Artificial Neural Network* ou réseau de neurones artificiels en français. 28, 29

ASTER *Advanced Spaceborne Thermal Emission and Reflection Radiometer*. 10

AUC *Area Under the Curve*. 132–136, 138, 140, 141, 145, 146

AVHRR *Advanced Very High Resolution Radiometer* ou radiomètre avancé à très haute résolution en français. 8, 18, 19, 28

CART *Classification And Regression Trees*. 41

CBLOF *Cluster-Based Local Outlier Factor*. 123

CCI-LC *Climate Change Initiative - Land Cover*. 19

CES OSO Centre d'Expertise Scientifique de l'Occupation des Sols. 3, 86

CESBIO Centre d'Études Spatiales de la BIOsphère. 19, 30, 60, 67

CESSC *Centre for Earth System Science China*. 19

CLC CORINE *(Coordination of Information on the Environment) Land Cover*. 4, 5, 14–16, 59

CNES Centre National d'Études Spatiales. 9

CNN *Convolutional Neural Network* ou réseau de neurones convolutif en français. 30, 31

COF *Class Outlier Factor*. 121

DISCover *Data and Information System Cover*. 18

DTW *Dynamic Time Warping* ou déformation temporelle dynamique en français. 142

EEA *European Environment Agency* ou agence européenne pour l'environnement en français. 14

EM Espérance-Maximisation. 115

ENN *Edited Nearest Neighbor*. 121, 122, 130, 149, 154, 160–167

ENVISAT *ENVIronment SATellite*. 8

ESA *European Space Agency* ou agence spatiale européenne en français. 9, 19, 50, 51, 189

FAO *Food and Agriculture Organization of the United Nations* ou organisation des Nations unies pour l'alimentation et l'agriculture en français. 18

GCOS *Global Climate Observing System*. 3, 6, 20

GEOGLAM *GEO Global Agricultural Monitoring*. 3

GIEC Groupe d'experts Intergouvernemental sur l'Évolution du Climat ou *Intergovernmental Panel on Climate Change* (IPCC) en anglais. 3

GLC *Global Land Cover*. 18

GLCC *Global Land Cover Characterization*. 18

GLP *Global Land Programme*, ex-*Global Land Project*. 3

GMES *Global Monitoring for Environment and Security*. 9

GMM *Gaussian Mixture Model* ou modèle de mélanges gaussiens en français. 28, 114

GPS *Global Positioning System* ou géo-positionnement par satellite en français. 6, 14, 17

HRL *High Resolution Layers*. 14

HRVIR Haute Résolution Visible et Infra-Rouge. 49

IGBP *International Geosphere-Biosphere Program* ou programme international pour la géosphère et la biosphère en français. 18, 19

IGN Institut National de l'Information Géographique et Forestière. 3, 13–15, 59, 60

INFFC *Iterative Noise Filter based on the Fusion of Classifiers*. 155

INSPIRE *Infrastructure for Spatial Information in the European Community* ou infrastructure d'information géographique dans la communauté européenne en français. 14

ISODATA *Iterative Self-Organizing Data Analysis Technique yAy*. 25

ITSS *Iterated Training Sample Selection*. 154

LCCS *Land Cover Classification System*. 18, 194

LCML *Land Cover Macro Language*. 194

LiDAR *Light Detection And Ranging*. 7

LOF *Local Outlier Factor*. 122, 123, 130, 136

LULCC *Land Use Land Cover Change*. 3

MACCS *Multi-Sensor Atmospheric Correction and Cloud Screening*. 58

MCD12Q1 *Collection 5 MODIS Land Cover Type Product*. 18, 20, 29

MERIS *Medium Resolution Imaging Spectrometer*. 8, 18, 28

MNT Modèle Numérique de Terrain. 51, 53, 64

MODIS *Moderate Resolution Imaging Spectroradiometer* ou radiomètre spectral pour imagerie de résolution moyenne en français. 8, 9, 18, 20, 28, 29, 67, 185

MOS Mode d'Occupation des Sols. 15

MSI *Multi-Spectral Instrument*. 50

MUSCATE MUlti-Satellite, multi-Capteurs pour des données multi-TEmporelles. 193

NASA *National Aeronautics and Space Administration*. 8, 49

NASG *National Administration of Surveying, Mapping and Geoinformation*. 19

NDBI *Normalized Difference Built-up Index*. 65

NDVI *Normalized Difference Vegetation Index*. 11, 65, 67–69, 82–84, 92–96, 101, 102, 104, 105, 107–109, 128, 162, 169, 198

NDWI *Normalized Difference Water Index*. 65

NICD *Noise Identification using Classifier Diversity*. 151, 153

NOAA *National Oceanic and Atmospheric Administration* ou agence américaine d'observation océanique et atmosphérique en français. 8, 18

OA *Overall Accuracy* ou taux de bonnes classification en français. 36, 45, 46, 72, 74, 75, 77, 79–82, 85, 96, 99, 101, 104–108, 132, 158, 160–171, 173

OCS-GE OCcupation des Sols à Grande Échelle. 3, 5, 13, 58, 59, 70, 89

OC-SVM *One Class - Support Vector Machine*. 115

OLI *Operationbal Land Imager*. 8, 49

OOB *Out Of Bag*. 43, 74

OSM *OpenStreetMap*. 17

PAC Politique Agricole Commune. 59

PACA Provence-Alpes-Côte d'Azur. 15

PEPS Plateforme d'Exploitation des Produits Sentinel. 55

PPV Plus Proches Voisins. 90, 126, 142, 150, 151, 153

PWEM *Pair-Wise Expectation Maximization*. 115, 153

RADAR *Radio Detection And Ranging*. 7

RBF *Radial Basis Function*. x, xiv, 35, 36, 91, 92, 98–109

RENN *Repeated Edited Nearest Neighbor*. 122, 130, 154, 160, 166

RF *Random Forest* ou forêt aléatoire en français. x, 29–31, 37, 38, 42, 43, 63, 69, 71, 72, 74–77, 79, 80, 82, 84–86, 90–92, 98–108, 110, 113, 124–128, 130, 131, 133, 140, 142–148, 150, 151, 153–158, 160–167, 169, 171, 173, 174, 177, 178, 197, 198

RNN *Replicator Neural Network*. 115

ROC *Receiver Operating Characteristic*. 132–136, 138, 140, 141, 144–146

RPG Registre Parcellaire Graphique. 58–60, 70, 89, 95, 129, 130, 195

SIG Système d'Information Géographique. 14, 17, 113

SOM *Self-Organizing Map*. 25

SOS *Stochastic Outlier Selection*. 120, 121, 130, 131, 140, 141, 143, 144

SPOT Satellites Pour l'Observation de la Terre. 8, 9, 18

SVDD *Support Vector Data Description*. 115

SVM *Support Vector Machine* ou Séparateur à Vaste Marge en français. x, xiv, 29–37, 65, 69, 71, 72, 74–77, 79, 80, 86, 91, 92, 98–110, 115, 116, 126, 177

Teruti-Lucas UTIlisation des TERitoires - *Land Use / Cover Area frame statistical Survey*. 16

TIRS *Thermal Infra-Red Sensor*. 49

TOA *Top of Atmosphere* désigne les réflectances exo-atmosphérique. 53, 57, 193

TOC *Top of Canopy* désigne les réflectances au sol. 53, 57, 58, 69, 92, 193

UMC Unité Minimale de Collecte. 4–6, 14, 15

URSS Union des Républiques Socialistes Soviétiques. 8

USGS *United States Geological Survey*. 9, 51, 57, 189

UTM *Universal Transverse Mercator*. 193

VCE Variable Climatique Essentielle ou *Essential Climate Variable* (ECV) en anglais. 3

Annexes

Annexe A

Données satellitaires et données de référence

A.1 Dates des images satellitaires utilisées

Les dates d'acquisition disponibles p our la série temporelle composée d'images SPOT-4 et Landsat-8 en 2013 (zone rouge de la Figure 3.1a) sont indiquées dans le Tableau A.1.

De manière similaire, les dates des acquisitions utilisées pour la série temporelle composée uniquement de données Landsat-8 (zone verte de la Figure 3.1a) sont montrées dans le Tableau A.2 pour chacune des huit tuiles. Le nom des tuiles suit la nomenclature de l'USGS, et est visible sur la Figure 3.2a.

Les dates des acquisitions utilisées pour les six tuiles sont indiquées dans le Tableau A.3. Les premières dates, en novembre et décembre, correspondent à des acquisitions de 2015, les autres dates correspondent aux acquisitions de 2016. Le nom des tuiles est visible sur la Figure 3.2a en suivant la nomenclature de l'ESA.

Tableau A.1 – Images disponibles pour la série temporelle composée d'images SPOT-4 et Landsat-8 en 2013.

	SPOT-4	**Landsat-8**
Dates	16/02	14/04
	21/02	19/07
	03/03	04/08
	03/08	20/08
	03/18	05/09
	03/23	07/10
	12/04	23/10
	17/04	10/12
	22/04	
	12/05	
	17/05	
	27/05	
	06/06	
	11/06	
	16/06	

Tableau A.2 – Images disponibles pour la série temporelle composée d'images Landsat-8 en 2013 pour huit tuiles (nomenclature USGS).

	D5H4	**D6H4**	**D4H3**	**D5H3**	**D6H3**	**D4H2**	**D5H2**	**D4H1**
Dates	14/04	16/04	14/04	14/04	14/04	14/04	04/14	14/04
	23/04	23/04	10/07	23/04	16/04	10/07	23/04	16/05
	16/05	02/05	19/07	25/05	23/04	19/07	25/05	10/06
	17/06	03/06	26/07	17/06	02/05	26/07	17/06	19/07
	26/06	17/06	04/08	26/06	18/05	04/08	26/06	20/08
	10/07	26/06	11/08	12/07	03/06	11/08	12/07	05/09
	12/07	05/07	20/08	19/07	26/06	20/08	19/07	14/10
	19/07	12/07	05/09	04/08	05/07	05/09	28/07	15/11
	26/07	19/07	28/09	13/08	12/07	12/09	04/08	01/12
	04/08	21/07	07/10	20/08	19/07	28/09	13/08	
	11/08	04/08	23/10	29/08	21/07	07/10	20/08	
	13/08	13/08	30/10	05/09	04/08	14/10	29/08	
	20/08	20/08	01/12	07/10	06/08	23/10	05/09	
	29/08	22/08	10/12	23/10	13/08	30/10	14/09	
	05/09	29/08	26/12	03/12	22/08	10/12	07/10	
	07/10	05/09		10/12	29/08	17/12	23/10	
	23/10	23/09			23/09		03/12	
	30/10	07/10			23/10		10/12	
	01/12	23/10			03/12			
	03/12	03/12			12/12			
	10/12	10/12						

Tableau A.3 – Images disponibles pour la série temporelle composée d'images Sentinel-2A acquises fin 2015 et sur l'année 2016 pour six tuiles (nomenclature ESA).

	T30TYN	**T30TYP**	**T31TCH**	**T31TCJ**	**T31TDH**	**T31TDJ**
Dates	03/12	03/12	30/11	30/11	30/11	30/11
	06/12	06/12	03/12	03/12	03/12	03/12
	23/12	23/12	23/12	23/12	23/12	23/12
	26/12	26/12	30/12	30/12	30/12	30/12
	05/01	02/01	12/01	12/01	12/01	12/01
	12/01	04/02	19/01	08/02	12/03	12/03
	25/01	15/03	12/03	22/03	22/03	22/03
	04/02	22/03	22/03	11/04	29/03	29/03
	14/02	01/04	29/03	14/04	08/04	08/04
	12/03	11/04	01/04	18/04	11/04	11/04
	15/03	14/04	08/04	28/04	18/04	18/04
	22/03	01/05	11/04	01/05	28/04	28/04
	01/04	04/05	18/04	04/05	01/05	01/05
	11/04	21/05	28/04	18/05	18/05	18/05
	14/04	24/05	01/05	21/05	21/05	21/05
	24/04	20/06	18/05	28/05	28/05	28/05
	01/05	23/06	21/05	20/06	07/06	07/06
	04/05	03/07	28/05	23/06	10/06	10/06
	14/05	10/07	07/06	27/06	20/06	20/06
	21/05	23/07	20/06	07/07	27/06	27/06
	24/05	30/07	27/06	10/07	07/07	07/07
	20/06	02/08	07/07	17/07	10/07	10/07
	23/06	12/08	10/07	27/07	17/07	17/07
	03/07	19/08	17/07	30/07	20/07	20/07
	10/07	22/08	20/07	06/08	27/07	27/07
	13/07	01/09	27/07	12/08	30/07	30/07
	23/07	11/09	30/07	16/08	06/08	06/08
	30/07	21/09	06/08	22/08	16/08	16/08
	02/08	28/09	16/08	26/08	19/08	19/08
	12/08	11/10	19/08	01/09	26/08	26/08
	19/08		26/08	05/09	29/08	29/08
	22/08		05/09	11/09	05/09	05/09
	01/09		08/09	15/09	08/09	15/09
	08/09		15/09	28/09	15/09	28/09
	11/09		18/09	15/10	25/09	08/10
	18/09		28/09		28/09	15/10
	21/09		08/10		05/10	
	28/09		15/10		08/10	
	01/10				15/10	
	08/10					
	11/10					

A.2 Niveau de traitements des données satellitaires

Le Tableau A.4 montre les différents niveaux de correction appliqués aux images SPOT-5 qui étaient effectués par l'organisme distributeur Spot Image.

	Corrections	**Précision de localisation**
Niveau 1A	Aucune correction géométrique. Correction des défauts radiométriques provenant des écarts de sensibilité entre les détecteurs de l'instrument de prise de vue. Image quasiment brute.	50 m
Niveau 1B	Identique au niveau 1A avec en plus les corrections géométriques.	50 m
Niveau 2A	La scène est rectifiée dans la projection cartographique standard - UTM WGS 84 -, l'image est géo-référencée. Cette correction est réalisée sans utiliser de points d'appui.	30 m
Niveau 2B	Identique au niveau 2A mais les corrections sont effectuées en prenant des points d'appui, mesurés sur une carte ou issus de relevés topographiques. Le produit peut être utilisé dans les régions avec un peu de relief.	30 m
Niveau 3	Identique au niveau 2B mais les erreurs résiduelles de parallaxe dues au relief sont corrigées grâce à l'utilisation d'un Modèle Numérique d'Elevation (MNE). Produit ortho-rectifié.	10 m

Tableau A.4 – Niveaux de pré-traitements effectués sur les images SPOT-5 par l'organisme distributeur Spot Image.

Concernant les images Sentinel-2, les produits fournis par le pôle Theia à travers l'atelier de production MUlti-Satellite, multi-Capteurs pour des données multi-TEmporelles (MUSCATE) sont les suivants :

Niveau 1C Données ortho-rectifiées en réflectance TOA.

Niveau 2A Données ortho-rectifiées en réflectance TOC après correction atmosphérique avec les masques de nuages (et de leurs ombres), des surfaces en eau et de la neige.

Niveau 3A Synthèses bi-mensuelles ou mensuelles de réflectance TOC constituées de la moyenne pondérée des réflectances TOC de surface des pixels non-nuageux obtenues au cours de la période.

A.3 Tuilage

La Figure A.1 montre les systèmes de tuilage utilisés pour les images Landsat-8 en vert et pour les images Sentinel-2 en bleu. Les images Landsat-8 sont projetées dans le système Lambert 93, tandis que les images Sentinel-2 sont fournies initialement en projection *Universal Transverse Mercator* (UTM) (UTM 30, 31 et 32). La reprojection des images Sentinel-2 en Lambert 93 explique les inclinaisons des tuiles observées à l'Est et à l'Ouest de la France sur la Figure A.1.

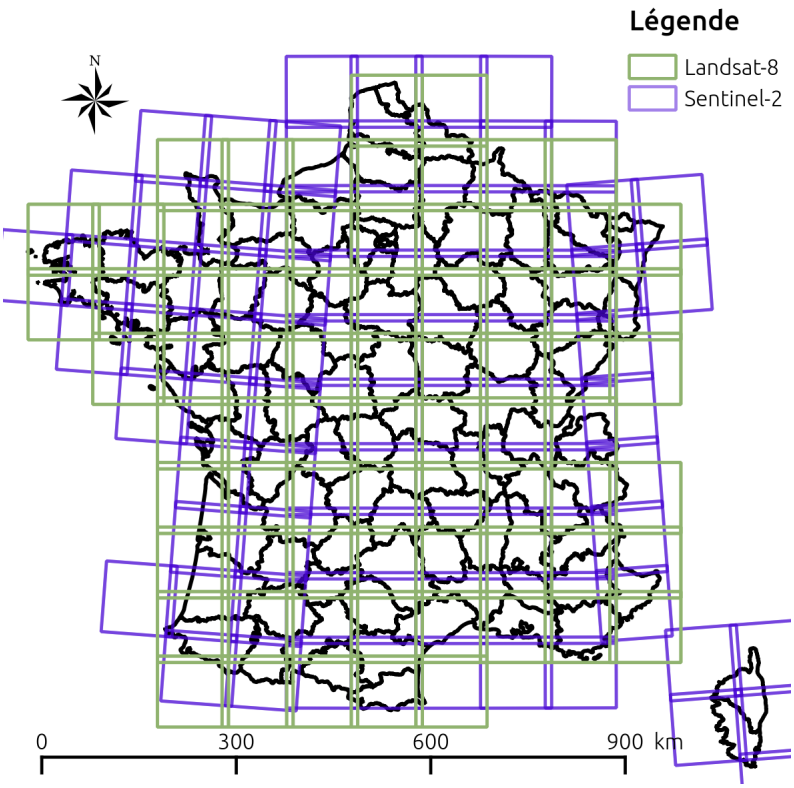


FIGURE A.1 – Systèmes de tuilage utilisés lors de la distribution des images Landsat-8 (en vert) et Sentinel-2 (en bleu).

A.4 Nomenclature

A.4.1 *Land Cover Classification System*

Le système LCCS s'appuyant sur le langage *Land Cover Macro Language* (LCML) est considéré comme l'un des standards pour la classification. Chaque classe d'occupation des sols est définie par la combinaison d'un ensemble d'attributs indépendants. LCCS permet de décrire l'ensemble des classes d'occupation des sols quelque soit la zone géographique étudiée.

La définition de la nomenclature LCCS repose sur deux étapes : (1) une étape dichotomique, et (2) une étape modulaire hiérarchique. La Figure A.2 montre l'ensemble de la nomenclature ainsi formée.

La première étape dichotomique décrit huit classes d'occupation des sols en niveau hiérarchique en se basant sur la présence de végétation ou non, le type de milieu et le degré d'artificialisation. Contrairement à d'autres nomenclatures hiérarchiques, celle du LCCS

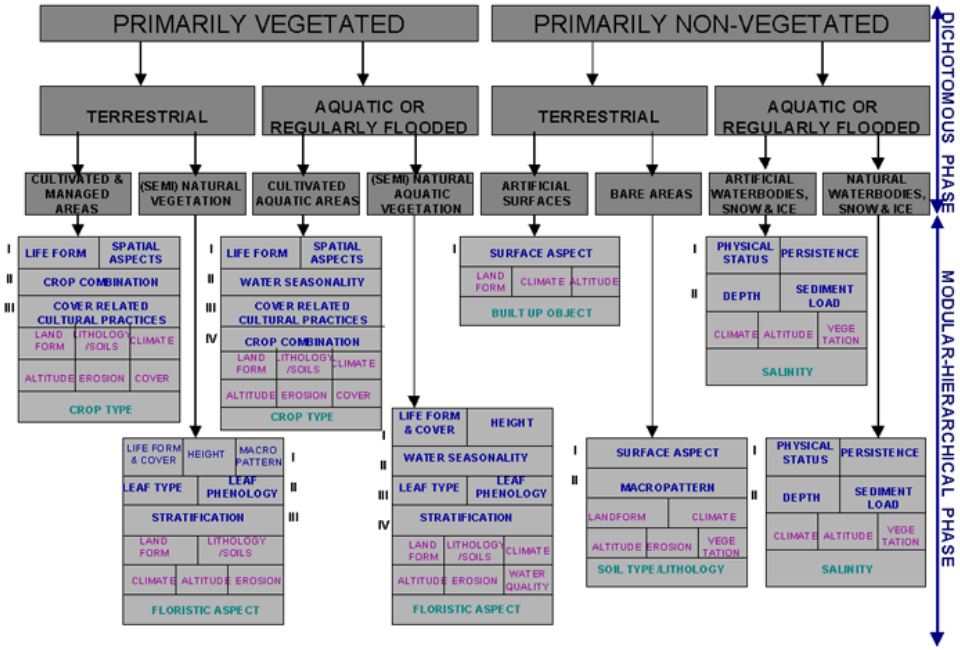


FIGURE A.2 – Nomenclature complète LCCS Di Gregorio and Jansen [2000].

permet de classer toute la surface terrestre sans aucun conflit : chaque point du globe n'appartient qu'à une seule catégorie. Cette nomenclature est décrite par le Tableau A.5.

TABLEAU A.5 – Nomenclature hiérarchique LCCS [Di Gregorio and Jansen, 2000]

Niveau 1	**Niveau 2**	**Niveau 3**
A Principalement végétalisée	**A1** Terrestre	**A11** Zone terrestre cultivée et gérée
		A12 Végétation terrestre naturelle et semi-naturelle
	A2 Aquatique ou régulièrement inondée	**A23** Aquatique cultivée ou régulièrement inondée
		A24 Aquatique naturelle et semi-naturelle ou régulièrement inondée
B Principalement non-végétalisée	**B1** Terrestre	**B15** Surface artificielle et zone associée
		B16 Surface nue
	B2 Aquatique ou régulièrement inondée	**B27** Plan d'eau artificiel, neige et glace
		B28 Plan d'eau naturel, neige et glace

La deuxième étape modulaire hiérarchique permet d'obtenir une nomenclature plus fine que celle du troisième niveau. Elle repose sur l'utilisation de trois attributs qui sont spécifiques à chacune des huit classes. Ces attributs concernent l'occupation des sols (en bleu sur la Figure A.2), l'environnement (en magenta sur la Figure A.2), et des aspects techniques spécifiques (en vert sur la Figure A.2).

A.4.2 Registre Parcellaire Graphique

La nomenclature du RPG est donnée par le Tableau A.6.

Tableau A.6 – Nomenclature du Registre Parcellaire Graphique

Blé tendre
Maïs grain et ensilage
Orge
Autres céréales
Colza
Tournesol
Autres oléagineux
Protéagineux
Plantes à fibres
Semences
Autres gels
Légumineuses à grains
Fourrage
Estives landes
Prairies permanentes
Prairies temporaires
Vergers
Vignes
Fruits à coque
Autres cultures industrielles
Légumes-fleurs
Arboriculture
Divers

A.4.3 Légende

La Figure A.3 donne la légende des cartes d'occupation des sols présentées au Chapitre 1.

Source : amcarto.alwaysdata.net/demo/test/compa_ocs.html

FIGURE A.3 – Légende

Annexe B

Compléments sur le *Random Forest*

B.1 Tirages aléatoires

B.1.1 Tirage aléatoire avec remise

La probabilité de tirer k fois un échantillon lors de n tirages avec remise est donnée par la loi binomiale suivante :

$$P(X = k) = \binom{k}{n} p^k (1-p)^{n-k}. \tag{B.1}$$

Dans le cas particulier du tirage des échantillons *bootstrap* pour la construction du modèle du RF, le nombre de tirages n est égal au nombre totale d'échantillons d'apprentissage N. De plus, chaque échantillon à la même probabilité d'être tiré au sort, la probabilité de succès p est donc égal à $1/N$.

La probabilité qu'un échantillon ne soit jamais tiré est donc :

$$P(X = 0) = \binom{0}{N} \frac{1}{N}^0 (1 - \frac{1}{N})^{N-0} \tag{B.2}$$

$$= (1 - \frac{1}{N})^N \tag{B.3}$$

Posons :

$$L = \lim_{N \to +\infty} (1 - \frac{1}{N})^N \tag{B.4}$$

Or :

$$\ln(L) = \lim_{N \to +\infty} N \ln(1 - \frac{1}{N}) \tag{B.5}$$

$$= \lim_{N \to +\infty} \frac{\ln(1 - \frac{1}{N})}{\frac{1}{N}} \tag{B.6}$$

Posons $x = 1/N$, alors

$$\ln(L) = \lim_{x \to 0} \frac{\ln(1-x)}{x} \tag{B.7}$$

Posons $f(x) = ln(1-x)$, et $g(x) = 0$. On a alors $f(0) = g(0) = 0$ et $g'(0) \not\!\!/ 0$, donc d'après la règle de l'Hôpital :

$$\ln(L) = \frac{g'(0)}{f'(0)} \tag{B.8}$$

$$= -1 \tag{B.9}$$

D'où,

$$L = e^{-1} \tag{B.10}$$

Finalement, la probabilité qu'un échantillon ne soit jamais tiré au sort pour un N très grand est :

$$P(X = 0) = e^{-1} = 0.3679 \tag{B.11}$$

Statistiquement, il y a bien environ un tiers des échantillons qui ne sont pas utilisés dans la construction de chacun des arbres du RF.

De la même manière, il est possible de calculer que :

— 36.79 % des échantillons sont inclus une seule fois

— 18.39 % des échantillons sont inclus exactement deux fois

— 6.13 % des échantillons sont inclus exactement trois fois

— 1.53 % des échantillons sont inclus exactement quatre fois

— *etc.*

B.1.2 Tirage aléatoire sans remise

La probabilité de tirer k variables parmi K variables d'intérêt sur l'ensemble des P variables en effectuant m tirages est donnée par la loi hypergéométrique :

$$P(X = k) = \frac{\binom{k}{K}\binom{m-k}{p-K}}{\binom{m}{p}}. \tag{B.12}$$

Cette formule permet de connaître la probabilité de tirer une des variables dans l'ensemble des variables d'intérêt lors de la construction d'un nœud dans le modèle du RF. Par exemple, si les échantillons sont représentés par 200 primitives ($p = 200$) dont 180 sont issues des bandes spectrales et 20 du NDVI ($K = 20$). Si l'on souhaite connaître, la probabilité qu'une variable de type NDVI soit sélectionnée au moins une fois lors de la construction d'un nœud pour potentiellement être utilisée dans la règle de décision, il faut calculer :

$$P(X > 1) = 1 - P(X = 0) \tag{B.13}$$

$$= 1 - \frac{\binom{0}{20}\binom{m-0}{200-20}}{\binom{m}{200}} \tag{B.14}$$

$$= 1 - \frac{\binom{m}{180}}{\binom{m}{200}} \tag{B.15}$$

Si $m = \sqrt{p} \sim 14$ (paramétrage classique pour le RF), on obtient

$$P(X > 1) = 0.7831 \tag{B.16}$$

Mais si l'on choisit $m = 2$, la probabilité devient :

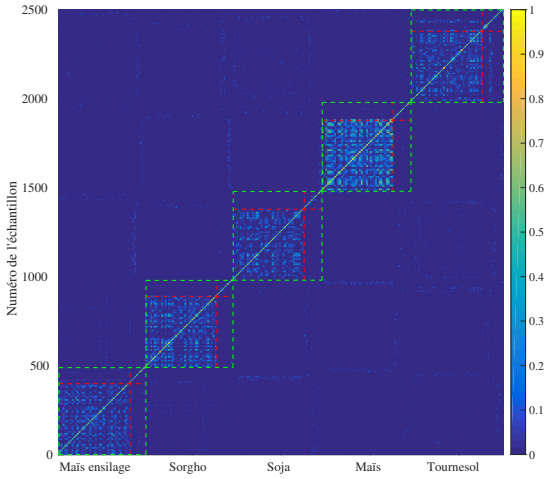
$$P(X > 1) = 0.1905 \tag{B.17}$$

B.2 Matrice de proximité

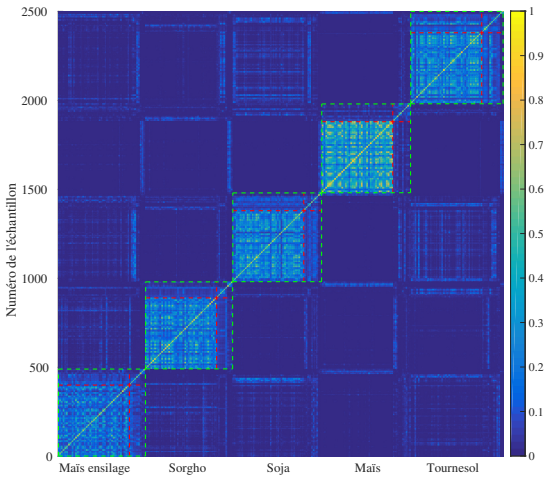
Afin de mieux visualiser la différence entre les trois mesures de proximité étudiées, la Figure B.1 montre les matrices de proximité calculées pour les données simulées à cinq classes décrites par les profils de NDVI. Un bruit aléatoire de 20 % est ajouté (Chapitre 5).

Pour faciliter la visualisation, les échantillons ont été ordonnés par classe (les échantillons de 1 à 500 appartiennent à la classe maïs ensilage, de 501 à 1000 à la classe sorgho, *etc*). Chaque carré vert représente les échantillons qui appartiennent à une même classe. De plus, les échantillons correctement étiquetés sont affichés avant ceux mal étiquetés. Les lignes en pointillés rouges représentent cette séparation.

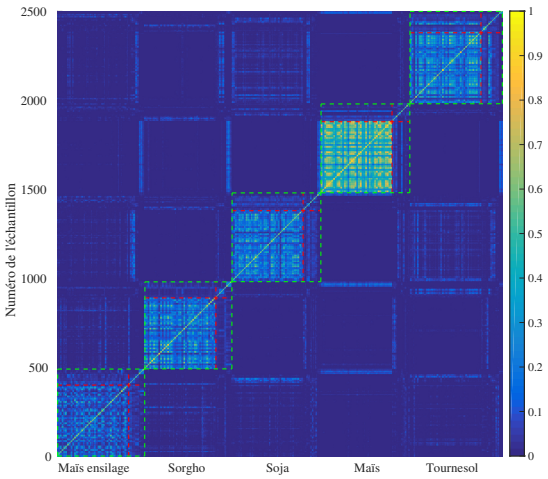
La Figure B.1a montre la matrice de proximité Breiman qui est caractérisé par de nombreuses valeurs nulles pour les échantillons appartenant à différentes classes. Les mesures de proximité proposées (DistanceLCA et PuretyLCA) étant moins sévères, elles ont généralement des valeurs de proximité plus élevées.



(a) Proximité Breiman



(b) Proximité DistanceLCA



(c) Proximité PuretyLCA

FIGURE B.1 – Illustration de différentes matrices de proximité.

Annexe C

Liste des publications

C.1 Journal international à comité de lecture

S. Ferrant, A. Selles, M. Le Page, P-A. Herrault, C. Pelletier, A. Al-Bitar, S. Mermoz, S. Gascoin, A. Bouvet, M. Saqalli, B. Dewandel, Y. Caballero, S. Ahmed, J-C. Maréchal & Y. Kerr. *Detection of irrigated crops from Sentinel-1 and Sentinel-2 data to estimate seasonal groundwater use in South-India*, Submitted to Remote Sensing MDPI (minor revisions).

C. Pelletier, S. Valero, J. Inglada, N. Champion, C. Marais-Sicre & G. Dedieu. *Effect of training class label noise on classification performances for land cover mapping with satellite image time series*, Remote Sensing MDPI, 2017.

C. Pelletier, S. Valero, J. Inglada, N. Champion & G. Dedieu. *Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas*, Remote Sensing of Environment, 2016.

C.2 Conférence internationale à comité de lecture

C. Pelletier, S. Valero, J. Inglada, N. Champion & G. Dedieu, *New iterative learning strategy to improve classification systems by using outlier detection*, IGARSS 2017, Fort Worth

C. Pelletier, S. Valero, J. Inglada, N. Champion & G. Dedieu, *Filtering mislabeled data for improving time series classification*, MultiTemp 2017, Bruges

J. Grimaldi, W. Trambouze, T. Dufourcq, M. Vergne, R. Fieuzal, C. Pelletier, T. Houet & V. Bustillo, *Can intercropped trees mitigate heat and drought effects on grapevines ? A study of microclimate patterns in agroforestry vineyards*, Southern France, AGRI 2017, Sitges

J. Inglada, B. Tardy, D. Derksen, C. Pelletier, A. Vincent, S. Valero, J. Michel & V. Thierion, *Operational land cover map production using Sentinel image time series supervised classification with out-of-date reference data*, WorldCover 2017, Esrin

C. Pelletier, S. Valero, J. Inglada, N. Champion & G. Dedieu. *An assessment of image features and Random Forest for land cover mapping over large areas using high resolution satellite image time series*, IGARSS 2016

S. Valero, C. Pelletier & M. Bertolino. *Patch-based reconstruction of high resolution satellite image time series with missing values using spatial, spectral and temporal similarities*, IGARSS 2016.

J. Grimaldi, R. Fieuzal, C. Pelletier, V. Bustillo, T. Houet & D. Sheeren. *Microclimate patterns in an agroforestry intercropped vineyard : first results*, EURAF 2016.

S. May & C. Pelletier. *Primal sketch of image series with edge preserving filtering. Application to change detection*, Multitemp 2015.

Résumé

L'étude des surfaces continentales est devenue ces dernières années un enjeu majeur à l'échelle mondiale pour la gestion et le suivi des territoires, notamment en matière de consommation des terres agricoles et d'étalement urbain. Dans ce contexte, les cartes d'occupation du sol caractérisant la couverture biophysique des terres émergées jouent un rôle essentiel pour la cartographie des surfaces continentales.

La production de ces cartes sur de grandes étendues s'appuie sur des données satellitaires qui permettent de photographier les surfaces continentales fréquemment et à faible coût. Le lancement de nouvelles constellations satellitaires – Landsat-8 et Sentinel-2 – permet depuis quelques années l'acquisition de séries temporelles à hautes résolutions. Ces dernières sont utilisées dans des processus de classification supervisée afin de produire les cartes d'occupation du sol. L'arrivée de ces nouvelles données ouvre de nouvelles perspectives, mais questionne sur le choix des algorithmes de classification et des données à fournir en entrée du système de classification.

Outre les données satellitaires, les algorithmes de classification supervisée utilisent des échantillons d'apprentissage pour définir leur règle de décision. Dans notre cas, ces échantillons sont étiquetés, *i.e.* la classe associée à une occupation des sols est connue. Ainsi, la qualité de la carte d'occupation des sols est directement liée à la qualité des étiquettes des échantillons d'apprentissage. Or, la classification sur de grandes étendues nécessite un grand nombre d'échantillons, qui caractérise la diversité des paysages. Cependant, la collecte de données de référence est une tâche longue et fastidieuse. Ainsi, les échantillons d'apprentissage sont bien souvent extraits d'anciennes bases de données pour obtenir un nombre conséquent d'échantillons sur l'ensemble de la surface à cartographier. Cependant, l'utilisation de ces anciennes données pour classer des images satellitaires plus récentes conduit à la présence de nombreuses données mal étiquetées parmi les échantillons d'apprentissage. Malheureusement, l'utilisation de ces échantillons mal étiquetés dans le processus de classification peut engendrer des erreurs de classification, et donc une détérioration de la qualité de la carte produite.

L'objectif général de la thèse vise à améliorer la classification des nouvelles séries temporelles d'images satellitaires à hautes résolutions. Le premier objectif consiste à déterminer la stabilité et la robustesse des méthodes de classification sur de grandes étendues. Plus particulièrement, les travaux portent sur l'analyse d'algorithmes de classification et la sensibilité de ces algorithmes vis-à-vis de leurs paramètres et des données en entrée du système de classification. De plus, la robustesse de ces algorithmes à la présence des données imparfaites est étudiée. Le second objectif s'intéresse aux erreurs présentes dans les données d'apprentissage, connues sous le nom de données mal étiquetées. Dans un premier temps, des méthodes de détection de données mal étiquetées sont proposées et étudiées. Dans un second temps, un cadre méthodologique est proposé afin de prendre en compte les données mal étiquetées dans le processus de classification. L'objectif est de réduire l'influence des données mal étiquetées sur les performances de l'algorithme de classification, et donc d'améliorer la carte d'occupation des sols produite.

Abstract

Land surface monitoring is a key challenge for diverse applications such as environment, forestry, hydrology and geology. Such monitoring is particularly helpful for the management of territories and the prediction of climate trends. For this purpose, mapping approaches that employ satellite-based Earth Observations at different spatial and temporal scales are used to obtain the land surface characteristics.

More precisely, supervised classification algorithms that exploit satellite data present many advantages compared to other mapping methods. In addition, the recent launches of new satellite constellations – Landsat-8 and Sentinel-2 – enable the acquisition of satellite image time series at high spatial and spectral resolutions, that are of great interest to describe vegetation land cover. These satellite data open new perspectives, but also interrogate the choice of classification algorithms and the choice of input data.

In addition, learning classification algorithms over large areas require a substantial number of instances per land cover class describing landscape variability. Accordingly, training data can be extracted from existing maps or specific existing databases, such as crop parcel farmer's declaration or government databases. When using these databases, the main drawbacks are the lack of accuracy and update problems due to a long production time. Unfortunately, the use of these imperfect training data lead to the presence of mislabeled training instance that may impact the classification performance, and so the quality of the produced land cover map.

Taking into account the above challenges, this Ph.D. work aims at improving the classification of new satellite image time series at high resolutions. The work has been divided into two main parts. The first Ph.D. goal consists in studying different classification systems by evaluating two classification algorithms with several input datasets. In addition, the stability and the robustness of the classification methods are discussed. The second goal deals with the errors contained in the training data. Firstly, methods for the detection of mislabeled data are proposed and analyzed. Secondly, a filtering method is proposed to take into account the mislabeled data in the classification framework. The objective is to reduce the influence of mislabeled data on the classification performance, and thus to improve the produced land cover map.