



HAL
open science

Méthodes d'exploitation de données historiques pour la production de cartes d'occupation des sols à partir d'images de télédétection et en absence de données de référence de la période à cartographier

Benjamin Tardy

► **To cite this version:**

Benjamin Tardy. Méthodes d'exploitation de données historiques pour la production de cartes d'occupation des sols à partir d'images de télédétection et en absence de données de référence de la période à cartographier. Géophysique [physics.geo-ph]. Université Paul Sabatier - Toulouse III, 2019. Français. NNT: 2019TOU30261 . tel-02966097

HAL Id: tel-02966097

<https://theses.hal.science/tel-02966097>

Submitted on 13 Oct 2020

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

Présentée et soutenue par
Benjamin TARDY

Le 19 décembre 2019

**Méthodes d'exploitation de données historiques pour la
production de cartes d'occupation des sols à partir d'images de
télédétection et en absence de données de référence de la
période à cartographier.**

Ecole doctorale : **SDUZE - Sciences de l'Univers, de l'Environnement et de
l'Espace**

Spécialité : **Surfaces et interfaces continentales, Hydrologie**

Unité de recherche :
CESBIO - Centre d'Etudes Spatiales de la Biosphère

Thèse dirigée par
Jordi INGLADA

Jury

Mme Agnès BEGUE, Rapporteur
M. Clément MALLET, Rapporteur
M. Nicolas COURTY, Examineur
M. Eric CESCHIA, Examineur
M. Mathieu FAUVEL, Examineur
M. Jordi INGLADA, Directeur de thèse

Résumé

L'étude des surfaces continentales constitue un enjeu majeur à l'échelle mondiale pour le suivi et la gestion des territoires, notamment en matière de répartition entre l'expansion urbaine, terres agricoles et espaces naturels. Dans ce contexte, les cartes d'Occupation des Sols (OCS) caractérisant la couverture biophysique des terres émergées sont un atout essentiel pour l'analyse des surfaces continentales. Les algorithmes de classification supervisée permettent, à partir de séries temporelles annuelles d'images satellites et de données de référence, de produire automatiquement la carte de la période correspondante. Cependant, les données de référence sont une information coûteuse à obtenir surtout sur de grandes étendues. En effet, les campagnes de relevés terrain requièrent un fort coût humain, et les bases de données sont associées à de longs délais de mises à jour. De plus, ces données de référence disposent d'une validité limitée à la période correspondante, en raison des changements d'OCS. Ces changements concernent essentiellement l'expansion urbaine au détriment des surfaces naturelles, et les terres agricoles soumises à la rotation des cultures. L'objectif général de la thèse vise à proposer des méthodes de production de cartes d'OCS sans exploiter les données de référence de la période correspondante. Les travaux menés s'appuient sur un historique d'OCS. Cet historique regroupe toutes les informations disponibles pour la zone concernée : cartes d'OCS, séries temporelles, données de référence, modèles de classification, etc. Une première partie des travaux considère que l'historique ne contient qu'une seule période. Ainsi, nous avons proposé un protocole de classification naïve permettant d'exploiter un classifieur déjà entraîné sur une nouvelle période. Les performances obtenues ont montré que cette approche se révèle insuffisante, requérant ainsi des méthodes plus performantes. L'adaptation de domaine permet d'aborder ce type de problématique. Nous avons considéré deux approches : la projection de données via une analyse canonique des corrélations et le transport optimal. Ces deux approches permettent de projeter les données de l'historique afin de réduire les différences avec l'année à traiter. Néanmoins ces approches offrent des résultats équivalents à la classification naïve pour des coûts de production bien plus significatifs. Une seconde partie des travaux considère que l'historique contient au moins deux périodes de données. À partir des cartes supervisées de ces périodes précédentes, nous proposons une approche de mise à jour de la carte la plus récente, en modélisant les transitions des classes d'OCS. Nous avons également proposé l'utilisation d'un classifieur unique entraîné à partir de plusieurs périodes de l'historique. L'objectif de ce classifieur consiste à pouvoir s'adapter aux variations entre les années. Enfin nous avons mis en place des systèmes de vote afin de réaliser une fusion de classifieurs, chacun entraîné sur une période différente de l'historique. Ces systèmes offrent l'avantage d'être toujours plus performants que chaque classifieur individuellement. Nous avons comparé les performances de plusieurs approches allant du simple vote majoritaire à des fusions plus complexes : vote par confiance, vote par probabilités, vote Dempster-Shafer ainsi qu'une inférence bayésienne. Ces approches produisent des performances similaires, mais pour des coûts de production variables. Nous avons expérimenté ces approches sur deux jeux de données, l'un constitué de sept années d'images Formosat-2 et l'autre de trois années d'images Sentinel-2. Le premier offre une très bonne diversité temporelle mais sur une faible emprise spatiale. Inversement, le second couvre une large zone mais pour un historique limité. Nous avons conclu que les approches du classifieur unique ainsi qu'un simple vote majoritaire offrent de bonnes performances pour des faibles coûts indépendamment du jeu de données.

Abstract

The study of continental surfaces is a major global challenge for the monitoring and management of territories, particularly in terms of the distribution between urban expansion, agricultural land and natural areas. In this context, land cover maps characterizing the biophysical cover of land are an essential asset for the analysis of continental surfaces. Supervised classification algorithms allow, from annual time series of satellite images and reference data, to automatically produce the map of the corresponding period. However, reference data is expensive information to obtain, especially over large areas. Indeed, field survey campaigns require a high human cost, and databases are associated with long update times. In addition, these reference data are valid only for the corresponding period due to changes in land use. These changes mainly concern urban expansion at the expense of natural areas, and agricultural land subject to crop rotation. The general objective of the thesis is to propose methods for producing land cover maps without exploiting the reference data of the corresponding period. The work carried out is based on the creation of a land cover history. This history includes all the information available for the area of interest: land cover maps, image time series, reference data, classification models, etc. A first part of the work considers that the history contains only one period. Thanks to this history we proposed a “naïve” classification approach allowing to use a classifier already trained, over a new period. The performances obtained shown that this approach is insufficient, thus requiring more efficient methods. Domain adaptation makes it possible to address this type of problem. We considered two approaches: data projection via canonical correlation analysis and optimal transport. These two approaches allow the historical data to be projected in order to reduce differences with the year to be processed. Nevertheless, these approaches offer results equivalent to the naive classification for much more significant production costs. A second part of the work considers that the history contains at least two periods of data. Based on the supervised maps of these previous periods, we propose an approach to updating the most recent map by modelling land cover class transitions. We also proposed the use of a single classifier trained from several periods of the history. The objective of this classifier is to be able to adapt to variations between years. Finally, we set up voting systems in order to merge classifiers, each trained over a different historical period. These systems offer the advantage of always being more efficient than each classifier individually. We compared the performance of several approaches ranging from simple majority voting to more complex fusion: confidence voting, probability voting, Dempster-Shafer fusion and Bayesian inference. These approaches produce similar performances, but for variable production costs. We experimented with these approaches on two datasets, one consisting of seven years of Formosat-2 images and the other of three years of Sentinel-2 images. The first one offers a very good temporal diversity but on a small spatial footprint. Conversely, the second one covers a large area but for a limited history. We concluded that the single classifier approaches as well as a simple majority vote offer good performance at low costs regardless of the data set.

Remerciements

L'écriture de ces remerciements, une épreuve supplémentaire dans ce long processus, est l'occasion de clôturer trois années de travail intensif (parfois) et passionné (toujours). Une fois n'est pas coutume, et pour éviter les jalousies, je vais commencer par me remercier d'avoir été un des piliers de ce travail. Je dois cependant partager ce podium avec mon ordinateur qui a supporté tant bien que mal la rédaction de ce manuscrit et qui signale son épuisement à l'écriture de ces lignes.

Si à ce stade du projet il est facile de fanfaronner, cependant l'histoire aurait été différente sans l'aide de vous tous, encadrants, amis et famille. Tout d'abord mon inestimable directeur de thèse, Jordi Inglada que je ne peux que remercier pour tous ses conseils, sa grande disponibilité, sa patience et son soutien infaillible au cours de ces années, du stage de fin d'étude à la soutenance de ces travaux de thèse. Merci également à Julien Michel, qui a co-encadré ma thèse, pour ses conseils, son soutien, sa disponibilité et ses remarques constructives lors du suivi, des publications et de la rédaction de ce manuscrit.

Je remercie le CNES et la région Occitanie pour avoir financé ces travaux.

Merci aussi aux membres du jury, Agnès Bégué, Eric Ceschia, Nicolas Courty, Mathieu Fauvel, Clément Mallet, d'avoir évalué avec pertinence et objectivité mes recherches, de m'avoir ouvert des pistes différentes. Je ne peux oublier le rôle des rapporteurs, Agnès et Clément qui, par leurs remarques constructives m'ont accompagné jusqu'au bout.

Je tiens particulièrement à associer toutes les personnes qui ont contribué de près ou de loin à l'aboutissement de cette thèse, qu'ils soient de la famille, amis ou collègues. Merci Arthur parce que *iota*² ça marche du premier coup, Claire pour m'avoir expliqué encore et encore comment étaient récoltées les données de référence et l'utilisation de (l'outil du mal) qgis, Dawa puisque la science n'avance que pendant les déjeuners au CNES, Mireille pour ses vastes connaissances sur les dataframes et son esprit critique sur le monde et les pratiques scientifiques, Myriam pour avoir rendu le bureau aussi vivant et plus feng-shui, Silvia pour son suivi et ses recommandations au fil des années, Vincent pour m'avoir offert ma première et dernière expérience d'ingénieur d'études au sein du CNRS, et tous ceux que j'oublie (in)volontairement.

Je remercie toutes les personnes qui m'ont soutenu durant la période sombre de la thèse, la rédaction, avec l'éternelle question *Alors ça avance ?*.

Je tiens encore plus à remercier celles et ceux qui m'ont permis de traverser une période encore plus sombre et qui ont contribué à mon retour plus vivant que jamais (dans l'ordre fourni par shuffle de python pour les mécontents) : Stéphanie parce que le meilleur des remèdes c'est le rire, Piboune, Léa, Alex, Aurore, Cathy, Cathy et Cathy, Sarah, Ludo, Clément, Lucie, Christine, Mireille, Claire, Vincent, Jordi, et tous les membres du pavillon de neuro et de l'hémato 3A.

Enfin, je tiens à remercier les membres de ma famille et mes proches pour leur soutien au cours de ces longues années d'études. En particulier mes parents qui m'ont toujours soutenu dans mes projets. Une dédicace particulière à mon père qui a pris de son précieux temps de vacances pour relire et corriger l'intégralité de ce manuscrit

À tous ceux qui auraient été, involontairement, oubliés ou insuffisamment remerciés, cet espace vous est dédié :

... *Insérez vos remerciements ici* ...

À tous les oubliés de ces remerciements, un grand merci.

I	Introduction	13
1	Contexte général	14
1.1	La télédétection au service de l'étude de la biosphère	14
1.2	La télédétection au service de l'occupation des sols	16
1.2.1	La production manuelle	16
1.2.2	La production automatique	18
1.3	Position du problème	20
1.3.1	Données de référence	20
1.3.2	Problématique	23
II	Présentation du problème	26
2	Production de cartes d'occupation du sol	27
2.1	Présentation des données	27
2.1.1	Données de référence	27
2.1.2	Séries temporelles d'images Formosat-2	30
2.1.3	L'utilisation des données passées	32
2.2	Classification Supervisée	34
2.2.1	Les Random Forests	34
2.2.2	Chaîne de production d'une carte d'occupation des sols	40
2.3	Valider une carte d'occupation des sols	41
2.3.1	Mesures de qualité d'une classification	41
2.3.2	Validation des méthodes de vote	43
3	La classification Naïve	44
3.1	Principe de l'approche	44
3.2	Résultats	45

3.2.1	Performances globales	45
3.2.2	Performances par classe	47
3.3	Séparation entre classes annuelles et pérennes	53
3.4	Synthèse	55
III Propositions de méthodes exploitant un unique domaine Source		57
4	Exploiter un historique	58
4.1	Transfert d'apprentissage	58
4.2	L'adaptation de domaine	60
4.2.1	Adaptation semi-supervisée de classifieur	61
4.2.2	Adaptation de classifieur par apprentissage actif	62
4.2.3	Sélection de primitives invariantes	63
4.2.4	Adaptation de données	63
4.3	Autres approches d'exploitation de l'historique	64
4.3.1	Détection de changements	64
4.3.2	Mise à jour des cartes d'occupation des sols	66
4.4	Synthèse	66
5	Analyse Canonique des Corrélations	69
5.1	Introduction	69
5.2	Principe de l'analyse canonique des corrélations	70
5.3	Une approche d'adaptation de domaine	71
5.3.1	Supprimer toute source de perturbations potentielles	72
5.3.2	Conservation du maximum d'informations	75
5.3.3	Mesurer l'impact des changements sur la projection ACC	78
5.4	Une approche de détection de changements	80
5.4.1	Principe du MAD	81
5.4.2	Expérimentations	81
6	Transport Optimal	88
6.1	Introduction	88
6.1.1	Le Transport Optimal à travers les siècles	88
6.1.2	Formulation de Monge-Kantorovitch	89
6.2	Transport Optimal et adaptation de domaine	93
6.2.1	La formulation discrète	93
6.2.2	L'utilisation des connaissances sur les deux domaines	97
6.3	Utilisation du Transport Optimal pour la production d'OCS	98
6.4	Application du Transport Optimal à un vaste jeu de données	99
6.4.1	Études de sensibilité	99
6.4.2	Production opérationnelle	104
6.5	La place du transport optimal dans la production opérationnelle	113

IV	Propositions de méthodes exploitant de multiples domaines Source	117
7	Modélisation de l'historique	118
7.1	Modéliser des connaissances	118
7.1.1	La logique de Markov	119
7.2	Modélisation des transitions de classes	122
7.2.1	Principe	122
7.2.2	Expérimentations	124
7.2.3	Limitations et perspectives d'amélioration	126
7.3	Modélisation des signatures temporelles	128
7.3.1	Le classifieur unique	128
7.3.2	Résultats	129
8	La fusion de classifications	132
8.1	Les systèmes de vote	132
8.1.1	Généralités	133
8.1.2	La fusion de classifieurs par vote	134
8.1.3	Conception du système de vote	137
8.1.4	Choix des entrées	137
8.1.5	Règles de décision	138
8.1.6	Le problème des non-décisions	138
8.1.7	Performances du vote majoritaire	140
8.1.8	Comparaison des systèmes de vote pondéré	143
8.2	La fusion par Dempster-Shafer	144
8.2.1	Théorie de Dempster-Shafer	145
8.2.2	Résultats	147
8.3	Approche de fusion bayésienne	148
8.3.1	Modélisation	148
8.3.2	Résultats	149
9	Comparaison des approches multi-annuelles	152
9.1	Comparaison statistique	152
9.2	Analyse des coûts de production	159
9.2.1	La composition de l'historique	159
9.2.2	La qualité des données	161
9.2.3	Temps de calcul	162
V	Mise en oeuvre des méthodes pour une production opérationnelle sur de grandes étendues	165
10	Production sur des grandes étendues avec Sentinel-2	166
10.1	iota^2	166
10.2	Données disponibles	168

10.3	Expérimentations	169
10.3.1	Préparation des données	169
10.3.2	Utilisation de la chaîne	170
10.4	Comparaison des différentes approches	172
10.4.1	Performances obtenues	172
10.4.2	Rendu visuel des approches de classification	175
10.4.3	Évaluation des coûts	180
10.5	Un début prometteur	181
VI	Conclusions	183
11	Conclusion générale	184
11.1	Bilan	184
11.2	Perspectives	187
	Bibliographie	189
	Glossaire	201
A	Bandes spectrales	205
B	La classification Naïve	207
B.1	Illustration de la PG naïve	207
B.2	Matrice de confusion	209
C	L'analyse canonique des corrélations	210
C.1	Résultats	210
D	Transport Optimal	211
D.1	Résultats annexes de l'étude de sensibilité	211
E	Synthèse des comparatifs de vote	219
F	Visualisation des performances iota^2	225

Liste des figures

1.1	Corine Land Cover millésime 2018.	17
1.2	Cas simple de classification non supervisée	19
1.3	Exemple d'un protocole de classification supervisée pour la production d'une carte d'OCS.	20
1.4	Nomenclature hierarchiqe utilisée lors des campagnes terrains du CESBIO	22
1.5	Utilisation de l'historique pour la production de cartes d'OCS	24
1.6	Présentation générale des différentes méthodes proposées dans ce manuscrit ayant conduis à des publications (blocs orange)	25
2.1	Données de référence obtenues lors de la campagne terrain	28
2.2	Dates d'acquisition et couverture nuageuse	31
2.3	Distorsions possibles entre profils NDVI d'années différentes	33
2.4	Exemple d'arbre de décision	35
2.5	Effet des divisions successives au sein d'un arbre de décision	37
2.6	Chaîne de traitement pour la production de cartes d'occupation des sols	40
3.1	Chaîne de traitement utilisée dans le cadre de la classification naïve.	44
3.2	FScore par classe obtenu pour la classification naïve	48
3.3	Exemple de matrice de confusion pour une classification naïve	50
3.4	Exemple de matrice de confusion pour une classification supervisée	51
3.5	Classifications supervisée et naïve pour le millésime 2011.	51
3.6	Chaîne de traitement utilisée dans le cadre de la classification SAP.	53
3.7	Performances de la classification SAP	54
3.8	Différences de FScore entre le cas naïf et la SAP	55
4.1	Une vue d'ensemble des différentes configurations du transfert d'apprentissage.	60
5.1	Principe de l'analyse canonique des corrélations exploitant des séries temporelles.	71
5.2	Précision globale de l'ACC en masquant toutes sources de bruit potentielles	73
5.3	Précision par classe de l' ACC en masquant toutes sources de bruit potentielles	74

5.4	Précision globale de l'ACC en masquant les pixels altérés (bord et nuages)	75
5.5	Précision par classe de l'ACC en masquant les pixels altérés (bord et nuages)	76
5.6	FScore cas particuliers de la projection ACC	77
5.7	Performances du MAD pour la production de cartes	83
5.8	Détection de changements entre demi-série de l'année 2009	85
6.1	Exemple de permutation respectant la contrainte d'approvisionnement.	91
6.2	Échantillons, à deux dimensions, du domaine Source et Cible générés aléatoirement selon des distributions gaussiennes. Trois classes sont représentées par des couleurs différentes.	94
6.3	Matrice de coût associée au jeu de données test.	95
6.4	Matrices de couplage et de permutations	96
6.5	Échantillons transportés pour les deux algorithmes de résolution EMD (gauche) et Sinkhorn (droite).	97
6.6	Matrice de coût et utilisation des étiquettes.	98
6.7	Chaîne de traitement utilisant le Transport Optimal pour la production de cartes d'OCS.	99
6.8	Matrices de performances de l'EMD supervisé	101
6.9	FScore par classe obtenu par le EMD	102
6.10	PG du Sinkhorn avec les échantillons issus des données de référence.	103
6.11	PG de l'EMD avec les échantillons issus des données de référence et l'utilisation des étiquettes.	104
6.12	PG du Sinkhorn avec les échantillons issus des données de référence et l'utilisation des étiquettes.	104
6.13	Exemple de frontière entre objets associés à différentes classes.	106
6.14	Performances de l'EMD avec la sélection par position	107
6.15	FScore obtenus pour le EMD avec la sélection par position	108
6.16	Résultat du K-moyennes sur la série temporelle de la classe Blé. Le filtrage est réalisé sur la classification naïve.	110
6.17	Matrices de performances globales de l'EMD SAP.	111
6.18	FScore obtenu par le TO EMD combiné à la procédure SAP.	112
6.19	Matrices de performances du Sinkhorn avec clustering	114
6.20	FScore obtenus par le TO Sinkhorn avec clustering.	115
7.1	Performances globales de la mise à jour de la carte	124
7.2	FScore par classe obtenus par la mise à jour de la carte	126
7.3	Conception de la chaîne de traitement pour le classifieur unique.	128
7.4	PG du classifieur unique.	130
7.5	FScore par classe obtenus par le classifieur unique pour différentes tailles d'historique.	130
8.1	Conception du système de vote pondéré.	137
8.2	Impact des non-décisions sur les résultats de PG	140
8.3	PG vote avec corrections des non-décisions	141
8.4	FScore du VM et des corrections de non-décisions	142
8.5	PG obtenues par les différentes méthodes de vote pondéré.	143
8.6	FScore obtenu par les différentes méthodes de vote pondéré, pour le cas 2.	144
8.7	PG obtenue pour les différentes approches de fusion par Dempster-Shafer	147

8.8	PG bayésien avec l'exploitation optionnelle du TO et des proportions π .	150
8.9	FScore obtenu pour les classifieurs bayésien	151
9.1	PG obtenue pour toutes les méthodes abordées dans ce chapitre.	153
9.2	FScore obtenus pour toutes les méthodes proposées.	156
9.3	Illustration des cartes produites selon différentes méthodes à partir des données Formosat-2.	157
9.4	Illustration des cartes produites selon différentes méthodes à partir des données Formosat-2.	158
9.5	Évaluation des performances en fonction du coût d'apprentissage	164
10.1	Illustration des tuiles Sentinel-2, en projection UTM	167
10.2	Illustration des régions éco-climatiques sur le grand sud	170
10.3	FScore obtenus par les différentes expérimentations.	174
10.4	Nomenclature à 23 classes pour la classification des images Sentinel-2	177
10.5	Illustration des cartes générées à partir des données Sentinel-2 et les différentes méthodes proposées (Classifications).	178
10.6	Illustration des cartes générées à partir des données Sentinel-2 et les différentes méthodes proposées (Fusion de classifieurs).	179
10.7	Carte d'occupation des sols de 2018 couvrant le sud de la France métropolitaine continentale.	180
10.8	Évaluation des coûts pour différentes méthodes en fonction du temps d'un apprentissage et différents nombres d'années utilisées.	181
B.1	Matrice de PG de la classification naïve et supervisée	208
B.2	Matrice de confusion de la classification naïve de 2008 (cible) par 2007 (source).	209
D.1	FScore par classe obtenu par le Sinkhorn avec sélection des échantillons à partir des données de référence des deux domaines.	212
D.2	FScore par classe obtenus par l'EMD avec exploitation des étiquettes des deux domaines.	213
D.3	FScore par classe obtenus par le Sinkhorn avec exploitation des étiquettes des deux domaines.	214
D.4	Matrices de performances pour le Sinkhorn avec la sélection par position.	215
D.5	FScore obtenus par le Sinkhorn avec la sélection par position.	216
D.6	Matrices de performance pour le TO EMD par classe.	217
D.7	FScore obtenu par l'EMD par classe. Toutes les classes sont transportées.	218
E.1	Comparatif du FScore pour le cas 3.	220
E.2	Comparatif du FScore pour le cas 4.	221
E.3	Comparatif du FScore pour le cas 5.	222
E.4	Comparatif du FScore pour le cas 6.	223
E.5	Intervalles de confiance associés au FScore pour le cas 4.	224
F.1	Carte de différence entre les classifications produites par le classifieur unique et bayésien	225

Liste des tables

2.1	Nomenclature et nombre d'échantillons dans la donnée de référence.	29
2.2	Bandes spectrales et longueurs d'onde associées pour Formosat-2.	30
2.3	Exemples de vecteurs de probabilités par classes fournis par Shark	39
2.4	Exemple de matrice de confusion	42
3.1	Performances de la classification naïve	46
3.2	Matrice de confusion après un regroupement hiérarchique des classes	52
3.3	Matrice de confusion supervisée après un regroupement hiérarchique des classes	52
5.1	PG obtenues à partir des demi séries temporelles.	80
5.2	Proportions de changements détectés par le MAD entre les demi-séries de 2009	85
5.3	Cohérences des changements détectés par le MAD et la référence pour 2009	85
5.4	Proportion de changements (chgts) artificiels (art) détectés par le MAD.	86
6.1	Exemple de temps de trajet entre les boulangeries x_i (lignes) et les cafés y_j . Les valeurs vertes indiquent un exemple de choix associés à une permutation s	90
7.1	Exemple d'une Matrice de Transition.	123
8.1	Exemple de pondération des votants par la méthode de Borda.	134
8.2	Classifications générées pour réaliser la fusion Dempster-Shafer	147
9.1	Évaluation des méthodes selon le critère de composition de l'historique.	160
9.2	Évaluation des méthodes selon le critère de qualité des données de l'historique.	162
9.3	Évaluation des coûts de production des différentes méthodes	163
10.1	Répartition du nombre d'images moyen disponible par tuile et par année	168
10.2	Nomenclature et nombre d'échantillons disponibles par année.	169
10.3	Performances obtenues par les différentes méthodes proposées sur la zone Grand Sud	175

10.4	Évaluation de l'effort de production des différentes méthodes	181
A.1	Bandes spectrales du Return Beam Vidicon (RBV)	205
A.2	Bandes spectrales du Multispectral Scanner (MSS)	205
A.3	Bandes spectrales du capteur optique Operational Land Imager (OLI)	206
A.4	Bandes spectrales Sentinel-2	206
C.1	Matrice d'écart entre l'ACC utilisant le masque de pixels altérés et le masque de classe annuelles	210
F.1	Répartition du nombre d'images par tuile et par année	226

Première partie

Introduction

1.1 La télédétection au service de l'étude de la biosphère

Au cours du siècle dernier, l'humanité a pris conscience de l'impact de ses activités sur le monde qui l'entoure. Impact d'autant plus visible que la population ne cesse d'augmenter, entraînant un besoin toujours plus grand de ressources et d'espace. Cette prise de conscience se traduit par la multiplication des études centrées sur la biosphère. La biosphère, telle que définie par Eduard Suess en 1875, désigne tous les écosystèmes abritant la vie sur Terre, c'est-à-dire la basse atmosphère, l'hydrosphère (océan mondial) et la lithosphère (surfaces continentales émergées). La définition du terme "biosphère" a évolué pour représenter les analyses systémiques de ces trois couches, ainsi que les interactions entre êtres vivants et milieux naturels. De nombreuses initiatives ont vu le jour, à l'image du programme [Programme sur l'homme et la biosphère ou *Man and Biosphere en anglais* \(MAB\)](#) de l'UNESCO¹ dont l'objectif principal vise la préservation des réserves naturelles de biodiversité. Le [Groupe d'experts intergouvernemental sur l'évolution du climat \(GIEC\)](#)² (ou *Intergovernmental Panel on Climate Change* (IPCC) en anglais) constitue un autre exemple axé sur l'analyse des causes et des risques liés au changement climatique. L'objectif commun de ces initiatives se concentre sur la prévention de la perte de biodiversité, tout en répondant aux contraintes économiques, sociales et écologiques. Ces projets reposent sur la communauté scientifique en charge d'analyser l'évolution et les changements au fil du temps. L'étude de la biosphère réunit de nombreuses disciplines scientifiques, de la biologie aux mathématiques en passant par la physique, et tout autant de thématiques couvrant les trois sphères de la biosphère. À titre d'exemple d'étude de la biosphère, nous pouvons citer les cycles biogéochimiques comme le cycle de l'eau, de l'azote ou encore de l'oxygène, la circulation océanique, la déforestation, etc. La plupart de ces études requièrent des relevés de terrain, mais également l'exploitation de données de télédétection comme les images satellites.

La télédétection, dans son acception la plus large, désigne l'acquisition d'informations relatives à un

¹<http://www.unesco.org/new/fr/natural-sciences/environment/ecological-sciences/>

²<https://www.ipcc.ch/>

phénomène ou à un objet terrestre, au moyen d'un outil qui n'est pas en contact direct avec l'observation. Selon cette définition, les sismographes, les appareils photographiques ou autres représentent des appareils de télédétection. L'essor des satellites dédiés à l'étude de la biosphère au cours des dernières années a contribué à l'amélioration des techniques d'analyse en fournissant des données précieuses en continu. Un des exemples les plus connus du grand public reste sans conteste la météorologie qui utilise en quasi temps réel des données acquises par satellites. La météorologie constitue également un bon exemple de pluralité des données utilisées, dans la mesure où les prédictions sont réalisées en combinant des relevés au sol, des analyses par ballon sonde et images satellites.

Parmi les différents types existants, pour l'étude de la biosphère, et plus particulièrement des surfaces continentales, les scientifiques utilisent essentiellement les satellites optiques. La télédétection spatiale est une discipline relativement récente, qui a réellement débuté en 1972 avec le tout premier satellite optique dédié à l'observation terrestre : Landsat-1, issu du programme Américain éponyme. Landsat-1 disposait de deux capteurs, un *Return Beam Vidicon* (RBV) et un *MultiSpectral Scanner* (MSS), acquérant des données tous les 18 jours à une résolution spatiale de 80m, et sur un total de 7 bandes spectrales (cf annexe A.1 et A.2).

De nombreux satellites ont succédé à ce précurseur, chacun apportant son lot d'améliorations, soit en termes de spectre étudié, soit de résolution spatiale ou temporelle (temps entre deux acquisitions).

Landsat-8, actuellement en fonctionnement depuis 2013, assure une source constante de données précieuses pour l'observation des surfaces émergées. Avec une acquisition de la même scène tout les 16 jours, sur une emprise de $185km \times 180km$, avec une résolution de 30m. Il dispose de neuf bandes spectrales (cf annexe A.3). Ces caractéristiques offrent des éléments appréciables pour l'étude des surfaces continentales.

Le dernier en date, Sentinel-2 est une constellation de deux satellites jumeaux, permettant une acquisition tous les cinq jours de l'ensemble des surfaces émergées, sur 13 bandes spectrales (cf annexe A.4) avec une résolution de 10m à 60m. Sentinel-2, au même titre que Landsat, appartient au groupe des satellites à acquisition systématique. L'acquisition couvre la totalité des terres émergées sur un segment de pôle à pôle. Ces acquisitions sont découpées en segments. Ceux-ci sont projetés en *Universal Transverse Mercator* (UTM), définissant ainsi les tuiles Sentinel-2. Ces aspects seront vu plus en détail par la suite. Sentinel-2 a été conçu spécifiquement pour l'étude des surfaces émergées, permettant le suivi d'évolution d'évènements sur de longue période comme l'agriculture, ou un cours laps de temps pour du suivi de catastrophe naturelle comme le passage d'un ouragan³.

Il existe également des satellites à acquisition à la demande. Les utilisateurs demandent une acquisition sur la zone souhaitée au fournisseur d'images, qui optimise la programmation de manière à satisfaire le maximum d'utilisateurs. L'utilisateur n'est pas certain de d'obtenir une acquisition. C'est le cas pour le satellite Formosat-2 que nous utilisons dans cette thèse. Cependant, ce type d'acquisition ne permet pas d'assurer une acquisition régulière des données, inversement aux satellites à acquisition systématique.

Pour l'étude de la biosphère et particulièrement des surfaces continentales, les cartes d'*OCcupation des Sols* (OCS) constituent un élément essentiel pour de nombreuses disciplines. Une carte d'OCS est une base de données géoréférencée où chaque point du territoire est associé à un identifiant de classe dans la nomenclature. Le terme "géoréférencé" indique que le produit possède des coordonnées dans un système de projection, ce qui permet de situer avec exactitude la zone couverte. Ces cartes sont utilisées pour l'analyse du cycle de l'eau [Sterling et al., 2012], du carbone [Jung et al., 2006] ou encore les bilans d'énergie [Claverie et al., 2012]. Elles représentent également un intérêt pour l'aménagement du territoire [Dewan and Yamaguchi,

³<https://labo.obs-mip.fr/multitemp/saint-barthelemy-apres-irma/>

2009], l'évolution des forêts [Hansen et al., 2013]. La production de cartes d'occupation des sols, sur de vastes étendues, repose sur l'exploitation des images de télédétection.

1.2 La télédétection au service de l'occupation des sols

En parallèle des évolutions technologiques et méthodologiques, les besoins des utilisateurs (géologues, hydrologues, agronomes, services d'aménagement du territoire par exemple) ont également évolué. La cartographie de l'OCS se caractérise par un intérêt commun pour une vaste communauté hétérogène (scientifique, socio-économique, politiques publiques, etc).

Une haute résolution (10m ou 1m) est souvent requise pour des enjeux liés aux suivis de cultures agricoles ou de la gestion de territoire. À plus grande échelle, une telle carte constituerait un support pour des enjeux environnementaux de grande envergure, comme la gestion des ressources forestières ou le suivi de l'évolution des déserts. Une carte d'OCS permet de distinguer deux variables distinctes pour l'étude de la biosphère : l'occupation (*land cover*) et l'utilisation des sols (*land use*). En effet, selon les définitions de l'OCS, l'occupation témoigne de la présence d'une classe, par exemple une rivière ou un bâtiment. Alors que l'utilisation représente l'usage, par exemple l'agriculture ou une zone industrielle. En fonction de l'utilisation souhaitée, l'un ou l'autre produit suffit généralement. Ces deux variables ne disposent pas forcément des mêmes pré-requis de mise à jour. On estime qu'une production tous les cinq ans suffit pour l'occupation alors qu'un rafraîchissement annuel est requis pour l'utilisation des sols. Néanmoins, la plupart des cartes d'OCS mélangent ces deux aspects.

La production automatique de cartes d'OCS est stimulée par l'arrivée de Sentinel-2 qui représente un atout majeur pour le suivi des surfaces émergées. L'utilisation de données Sentinel-2 apporte son lot de défis. Le plus conséquent se caractérise par le volume important de données à traiter, avec une image tous les cinq jours et une fauchée (distance couverte par une acquisition) de 290 km. Ces caractéristiques imposent des contraintes sur les méthodes à considérer en termes de complexité de traitement.

1.2.1 La production manuelle

Photo-Interprétation

La photo-interprétation ou interprétation humaine est un processus lent et fastidieux durant lequel un intervenant humain analyse une image pour en déterminer le contenu. En règle générale, pour la production de cartes d'OCS, l'identification des classes requiert des images issues de différentes périodes de l'année. L'exemple de la classe Colza est démonstratif car sa floraison homogène rend particulièrement facile son identification, sous réserve d'avoir au moins une image exploitable en avril.

Corine Land Cover (CLC) constitue l'exemple de cartes d'OCS le plus étendu produit par photo-interprétation. Ce produit européen fournit une carte d'OCS, couvrant 39 pays européens. CLC représente des objets d'une surface minimale de 25 hectares ou de 100 m minimum pour les objets ligneux (routes, ...), ce qui représente un bon compromis entre coût de production et finesse de l'information de l'OCS.

CLC possède une nomenclature de 44 classes regroupées dans une nomenclature hiérarchique à trois niveaux (le terme de nomenclature hiérarchique est détaillé dans la section 1.3.1). Cette nomenclature a été définie pour le premier CLC en 1990 et demeure inchangée depuis. La description des 44 classes a été significativement améliorée aux cours des productions successives pour offrir une meilleure interprétation des

classes [Kosztra et al., 2019]. La figure 1.1 présente une illustration de CLC centrée sur la France et les 44 classes de la nomenclature de niveau 3.

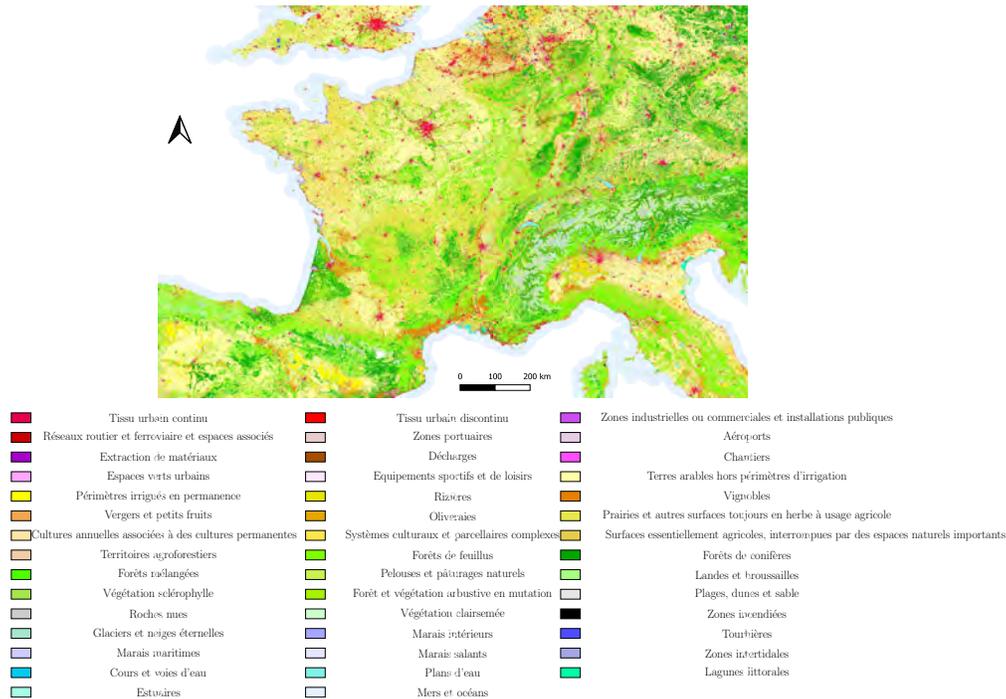


FIGURE 1.1 : Corine Land Cover millésime 2018.

La production de CLC exploite des images non nuageuses issues de différents capteurs, initialement les images Landsat-4/5 puis [Satellite pour l'observation de la Terre \(SPOT\) 4/5](#) et enfin Sentinel-2 et Landsat-8 pour la production 2018. Malgré l'amélioration de la résolution des images utilisées, la carte finale conserve l'unité minimale de collecte de 25 ha. Cinq CLC ont été produites, la première en 1990, puis les suivantes avec un rythme régulier de 6 ans à partir de 2000, la dernière date de l'année 2018. Selon les informations issues de la note technique du produit 2018⁴, les temps de production diminuent d'année en année, avec un temps de production de 13 ans pour le produit de 1990 et un record de 1.5 ans pour le produit 2018. L'amélioration des temps de production résulte de l'utilisation de techniques de détection de changements, permettant de réduire le volume de données à classer par photo-interprétation.

Les relevés terrains

En France, depuis 1946 le service statistique du ministère de l'Agriculture, au sein du projet "Agreste", réalise une enquête auprès des agriculteurs. Cette initiative appelée Teruti puis Teruti-Lucas relève l'OCS soit à l'échelle du point (cercles de 3m à 40m de diamètre) ou du segment homogène d'environ 1.5 km². Cette enquête représente un échantillonnage statistique de l'OCS qui peut être exploité pour produire une carte. Cependant le mode de collecte soulève des problèmes d'exhaustivité et de précision.

A l'heure du développement des initiatives communautaires, la production participative (ou *crowdsourcing*

⁴<https://land.copernicus.eu/user-corner/technical-library>

en anglais) réunit de plus en plus d'adeptes. Grâce à la prolifération des [Système d'Information Géographique \(SIG\)](#), chaque individu peut participer à l'identification de l'OCS, sous réserve d'un accès à des images aéroportées ou de données GPS, afin d'assurer la correspondance du géoréférencement.

Une des plus importantes initiatives communautaires à ce jour, [OpenStreetMap \(OSM\)](#), lancée en 2004, permet après une simple inscription via leur site⁵ de renseigner, corriger ou compléter l'information de la topographie pour une zone voulue. Le principe général consiste en ce que chaque participant puisse, au fil des trajets quotidiens, signaler des changements et ainsi assurer une mise à jour constante de la base de données. Un des avantages majeurs de la production participative réside en son très faible coût lié à la maintenance des serveurs recueillant les informations et la création ou mise à jour d'outils dédiés aux relevés. Cependant, ce faible coût compense difficilement le taux d'erreurs important dans ces relevés, en partie en raison de participants non experts en SIG ou en analyse terrain, induisant des imprécisions ou des confusions dans les relevés. À cette fin, OSM propose désormais des outils d'assurance qualité⁶ permettant de limiter les erreurs aberrantes de géométrie, par exemple. L'exactitude d'OSM constitue une interrogation constante des utilisateurs, ainsi [[Zielstra and Zipf, 2010](#)] quantifie les erreurs géométrique dans les relevés sur plusieurs zones de France. [[Girres and Touya, 2010](#)] réalisent une analyse similaire sur l'Allemagne. Le constat commun de ces travaux concerne l'amélioration de la qualité des relevés grâce aux outils de corrections. Ils mettent également en évidence que les relevés pédestres jouissent d'une meilleure précision.

Les bases de données topographiques peuvent être exploitées pour déduire des polygones de validation dans le cadre de la production de cartes d'OCS.

1.2.2 La production automatique

Bien que les méthodes manuelles, notamment la photo-interprétation offrent des précisions correctes, elles induisent des forts coûts de production, proportionnels à la surface considérée. Aussi, pour de grandes étendues, on privilégie les méthodes automatiques. L'objet de cette section est d'introduire les bases de la classification supervisée. Nous détaillerons les méthodes exploitées dans les travaux dans la suite du manuscrit. La production automatique de cartes d'occupation du sol passe par l'exploitation des algorithmes d'[Apprentissage Automatique \(AA\)](#). L'AA regroupe un ensemble de méthodes permettant de prendre des décisions de façon automatique à partir de données sans en expliciter les règles.

Au sein de la grande famille d'algorithmes de classification de l'AA, deux grands groupes prédominent, les algorithmes non supervisés et supervisés.

Classification Non Supervisée

En vision par ordinateur et reconnaissance d'objets, on utilise fréquemment les algorithmes de classification non supervisés ou algorithmes de regroupement (*clustering* en anglais). À partir d'un ensemble d'objets non différenciés, l'algorithme cherche une séparation optimale entre N groupes ou *clusters* généralement définis par l'utilisateur. Une fois les règles de séparation estimées, l'ensemble des échantillons sont classés. La décision prise par le classifieur consiste à attribuer un identifiant de groupe à chaque échantillon. Ensuite l'utilisateur associe à chaque groupe une classe d'OCS correspondante, ou si on possède des objets étiquetés, celle de

⁵<https://openstreetmap.org>

⁶https://wiki.openstreetmap.org/wiki/FR:Assurance_qualit%C3%A9

la majorité. La figure 1.2 illustre une classification non-supervisée, qui consiste à grouper les échantillons proches en trois groupes (*k-means* [Hartigan and Wong, 1979]).

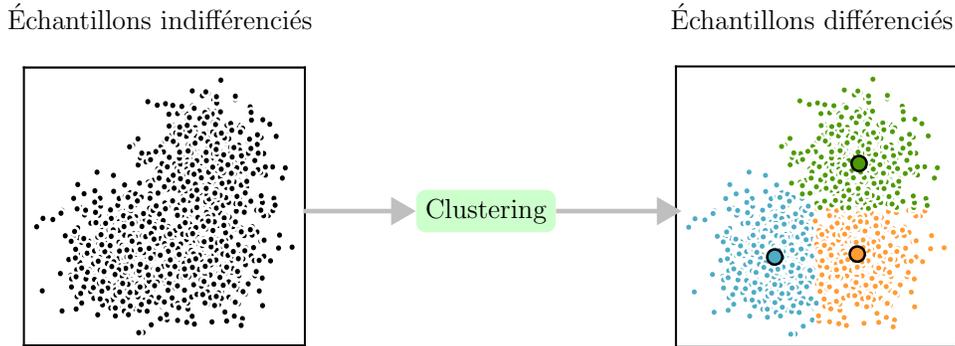


FIGURE 1.2 : Cas simple de classification non supervisée. À partir d'un ensemble d'échantillons non différenciés (sans étiquettes) et du paramètre $k=3$, il fournit la solution présentée à droite.

Ce genre d'algorithme se révèle efficace pour des questions de classification avec peu de classes ou de faible dimension. À plus grande dimension, les mesures de similarités s'avèrent moins adaptées, ce qui impacte la pertinence des frontières de décisions. Dans le cadre de la production de cartes d'OCS, les approches non-supervisées ont rapidement montré leurs limites.

Classification Supervisée

A l'inverse de la classification non supervisée, on différencie les objets utilisés lors de la phase d'apprentissage, c'est à dire qu'ils appartiennent à une classe identifiée par une étiquette. Dans ce manuscrit nous n'aborderons pas les approches semi-supervisées exploitant des ensembles d'objets partiellement étiquetés.

Grâce aux étiquettes des objets, la phase d'apprentissage consiste alors à trouver une frontière optimale entre les différentes classes. On doit relever ici l'importance de la notion de frontière optimale. Trois cas de figures peuvent apparaître lors de l'apprentissage, le sous-apprentissage, le sur-apprentissage et enfin celui recherché qui est le meilleur compromis entre les deux précédents. Cette problématique est connu sous le terme de compromis (ou dilemme) entre biais et variance.

En AA, le compromis biais-variance représente le problème de minimiser simultanément deux sources d'erreurs qui impacte la capacité de généralisation d'un classifieur supervisé.

La variance correspond à l'erreur due à la sensibilité aux légères variations du jeu d'apprentissage. Une variance élevée peut entraîner un sur-apprentissage. Un sur-apprentissage modélise le bruit aléatoire des échantillons d'apprentissage plutôt que les sorties prévues.

Le biais correspond à l'erreur provenant de règles de décisions fausses dans l'algorithme d'apprentissage. Un biais élevé peut être lié à un algorithme qui manque de relations pertinentes entre les données en entrée et les sorties prévues. On parle alors de sous-apprentissage, qui apparaît dès qu'un classifieur est trop simple pour modéliser un problème.

On définit alors l'erreur associée à un classifieur comme étant :

$$Err = biais^2 + variance + \text{erreur irréductible}$$

L'erreur irréductible correspond à l'erreur qui sera toujours réalisée par le classifieur. Ce genre d'erreur peut venir d'une erreur d'étiquetage des données ou des échantillons très proches qu'il n'est pas possible de distinguer par classification automatique.

Par conséquent, la seule solution viable résulte d'un compromis entre les deux cas. Il faut alors que les règles de décision séparent efficacement les différentes classes, tout en acceptant qu'une certaine quantité d'échantillons ne soient pas correctement classés. Cette solution autorise alors les erreurs d'étiquetage des échantillons d'apprentissage (qui sont fréquentes).

La frontière de séparation, ou règle de décision estimée dépend de l'algorithme utilisé.

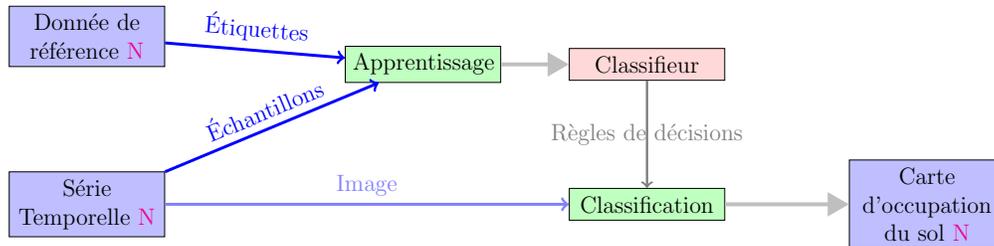


FIGURE 1.3 : Exemple d'un protocole de classification supervisée pour la production d'une carte d'OCS.

Le point commun de tous les algorithmes supervisés réside en l'utilisation d'étiquettes pour identifier les objets lors de l'apprentissage. Ces étiquettes proviennent de données de référence constituées au préalable par un expert ou par l'utilisateur. La figure 1.3 présente un exemple simplifié de chaîne de classification supervisée. Généralement, la quantité et la qualité des données de référence impactent fortement les performances obtenues par le classifieur. Le nombre de dimensions des échantillons influe également sur les temps d'exécution et l'efficacité de l'algorithme. Cette dépendance aux données de référence et à leur qualité explique la grande efficacité des algorithmes supervisés pour résoudre des problèmes de classification complexes.

1.3 Position du problème

Devant la demande croissante de cartes d'OCS de bonne qualité, et dans des délais de production toujours plus rapides, les méthodes actuelles ne parviennent pas à satisfaire ces deux critères conjointement. D'une part les méthodes non supervisées permettent de faibles délais de production mais avec une qualité moindre, voire très insuffisante. D'autre part les méthodes supervisées satisfont les exigences de qualité mais le délai de production est soumis à la mise à disposition des données de référence correspondantes.

1.3.1 Données de référence

Les algorithmes de classification supervisés utilisent des données de référence pour étiqueter une partie des échantillons que l'on souhaite classer. Cela permet d'obtenir de très bonnes performances, généralement supérieures aux approches non supervisées, surtout pour la production de cartes d'OCS.

Dans le cadre de l'OCS, les données de références représentent des bases de données géoréférencées, mises en relation avec les séries temporelles d'images satellites. Les données de référence respectent la hiérarchie induite par la nomenclature. Une nomenclature correspond à un inventaire exhaustif des éléments appartenant à l'objet d'étude. Dans le cadre de l'OCS, elle recense les différentes classes que nous souhaitons représenter

dans la carte produite. Lors de la production, son rôle consiste à associer une étiquette, fournie par la classification, à un nom de classe, représentatif de l'OCS.

La nomenclature définit le degré de finesse sémantique que l'on peut espérer produire à partir des algorithmes d'AA. Généralement, on utilise une nomenclature hiérarchique, offrant une représentation de l'OCS à finesse variable.

L'élaboration d'une nomenclature valide relève de procédés simples en apparences : il s'agit de répartir les classes d'OCS en groupes thématiques représentatifs. Cependant, son élaboration soulève des questions complexes, notamment en termes d'universalité et de réutilisation. Ainsi, de grands groupes comme l'organisation pour l'alimentation et l'agriculture⁷ (FAO), ou EAGLE⁸ proposent des méthodologies et outils permettant de créer une nomenclature répondant à ces questions. L'objectif de ces groupes consiste à proposer des outils sémantiques afin d'harmoniser la terminologie utilisée et donc pas conséquent la représentation des classes. Ainsi, le transfert d'information entre le fournisseur de la carte et l'utilisateur se réalise sans pertes.

La figure 1.4 illustre un exemple de la nomenclature utilisée lors des campagnes terrain du [Centre d'Études Spatiales de la BIOSphère \(CESBIO\)](#). Les cercles colorés indiquent la nomenclature à 17 classes que nous utilisons dans ces travaux.

La constitution d'une base de données de référence constitue un travail de longue haleine et un coût important (financier et humain). Une autre approche consiste à mener des campagnes terrains. On réalise généralement ces routines sur de petites étendues, car elles requièrent, dans le protocole du [CESBIO](#), le déplacement d'au moins deux agents pour réaliser un relevé exhaustif de l'OCS sur un ensemble de parcelles préalablement décidées (à partir d'une image satellites haute résolution, d'un cadastre ou de mesure sur la parcelle elle même). Les campagnes terrains requièrent à minima deux passages dans l'année, pour relever l'état des sols en été et en hiver (surtout pour les terres agricoles).

Ces deux approches compensent le fort coût par un très faible taux d'erreur lors des relevés ou de l'interprétation. Cependant, pour couvrir une zone étendue il devient évident que la charge de travail devient trop importante pour satisfaire les délais de production, tout en conservant la qualité maximale.

Heureusement, il existe plusieurs bases de données annexes exploitables. Par exemple, la BD TOPO de l'[Institut Géographique National \(IGN\)](#) constitue une base de données, divisée en plusieurs thèmes (hydrologie, bâti, végétation, etc), qui recense une grande diversité d'objets pouvant être exploités comme données de référence.

La [Politique Agricole Commune \(PAC\)](#) présente un avantage majeur pour la production de cartes d'OCS couvrant la France. En effet, dans le cadre de ce programme, les agriculteurs français souhaitant bénéficier de cette aide doivent déclarer toutes leurs parcelles agricoles en y identifiant les mises en cultures. On consigne ces informations dans une base de données géoréférencée mise à jour chaque année : le [Registre Parcellaire Graphique \(RPG\)](#). Les agriculteurs effectuent directement la complétion de cette base de données sur un support, où les parcelles sont définies à partir d'images aériennes. Avant 2016, les déclarations pouvaient être réalisées par îlots (groupement de parcelles contiguës exploitées par le même agriculteur) en explicitant le pourcentage de chaque culture semée. Désormais, chaque parcelle est déclarée séparément, ce qui réduit les erreurs d'exploitation du [RPG](#) lors de l'utilisation. Le [RPG](#) n'est cependant pas exhaustif à l'échelle du territoire dans la mesure où seuls les bénéficiaires de la [PAC](#) sont soumis aux déclarations. Malgré une mise à jour annuelle de la base de données, celle-ci n'est transmise aux institutions qu'au terme de l'année suivante,

⁷<http://www.fao.org/>

⁸<https://land.copernicus.eu/eagle/>

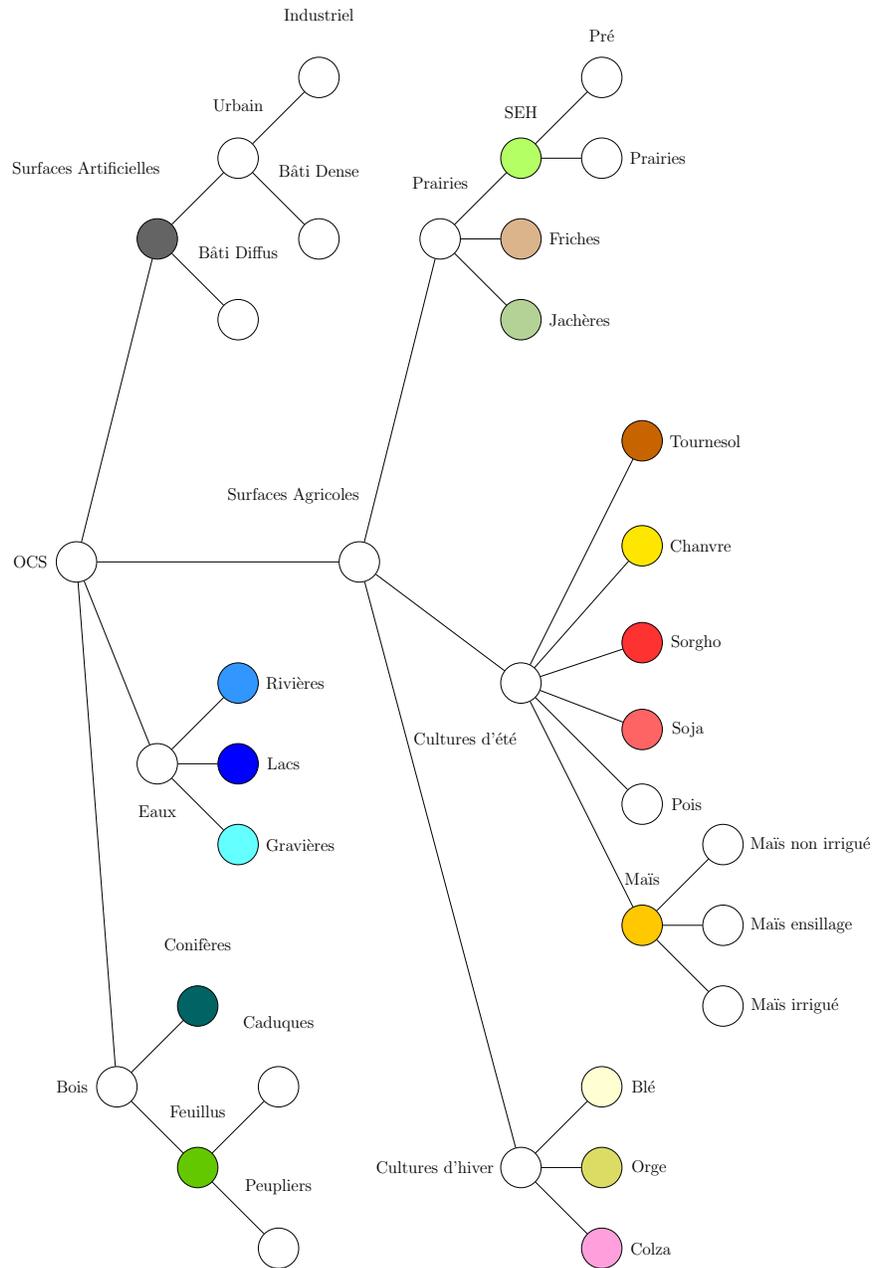


FIGURE 1.4 : Nomenclature hiérarchique utilisée lors des campagnes terrains du [CESBIO](#). Les cercles colorés indiquent les classes retenues composant la nomenclature utilisée dans la suite des travaux.

lorsqu'il n'y a pas de retard. Ce qui ajoute un temps supplémentaire pour la production du millésime dans des délais courts.

Quelque soit l'origine des données de référence, les coûts associés sont importants. Un des principaux problèmes liés aux données de référence se caractérise par leurs délais de mises à jour, souvent largement

supérieurs à la vitesse des changements de l'OCS. Ainsi, les bases de données représentent une information annuelle, qui ne permet pas de connaître un changement intermédiaire, par exemple une parcelle de sol nu après la récolte d'une culture.

Dès lors, il paraît impossible que les relevés d'une année M soient valables pour l'année N . On peut nuancer ces propos, car ces éléments dépendent en grande partie de la nomenclature choisie et de la fréquence de mise à jour souhaitée. Dans notre cas, nous visons une fréquence de production annuelle, nous pouvons alors différencier deux types de classes : les classes annuelles et pérennes.

Par définition une classe pérenne désigne une classe géographiquement stable, par exemple un lac ou une forêt. Ces parcelles peuvent subir des variations minimales comme une baisse du niveau de l'eau en raison d'une sécheresse. Pour ces classes, observer un changement requiert un suivi sur plusieurs années. À l'inverse, pour les classes annuelles le changement est la définition même de la classe. On va retrouver parmi celles-ci toutes les classes de cultures agricoles, soumises aux cycles de cultures, du sol nu à la récolte. La rotation des cultures permet de préserver les propriétés du sol, en imposant une alternance entre cultures d'été et culture d'hiver. Ainsi d'une année sur l'autre, seule l'information associée à l'identification d'une parcelle agricole peut être valide mais pas son contenu. Ces variations rendent obsolètes les données de référence pour les classes annuelles dès que l'année de leur acquisition est terminée. Cependant, elles demeurent utilisables pour une nouvelle production sur l'année correspondante.

Malgré cette validité éphémère, l'acquisition des données de référence en continu permet de constituer un historique sur la zone couverte, témoignant ainsi de l'évolution de l'OCS au fil du temps.

1.3.2 Problématique

Les méthodes utilisées actuellement pour la production de cartes d'OCS reposent sur des algorithmes de classification supervisées, car ils produisent de bien meilleurs résultats que les algorithmes non supervisés. Cependant, l'inconvénient majeur des approches supervisées se caractérise par leur dépendance aux données de référence très coûteuses à acquérir.

Les méthodes non-supervisées nécessitent, dans une moindre mesure, des données de référence pour associer une étiquette thématique (nom de classe) au résultat de la classification.

L'objectif de cette thèse est de proposer des méthodes de production de cartes d'OCS ne nécessitant pas les données de référence de la période à cartographier. Nous implémenterons ces méthodes dans une approche opérationnelle et elles devront donc satisfaire les exigences de qualité, de délais de production et de modularité (capteurs et classes variables).

Dans ce contexte opérationnel, la notion d'historique représente un avantage considérable. L'historique peut être constitué à partir de toutes les sources disponibles sur la zone d'étude. Ainsi, les séries temporelles des années précédentes et leurs données de référence constituent la base de l'historique auquel peut s'ajouter toutes les cartes d'OCS créées au fil du temps et toute base de données mise à jour entre temps. Mais il peut également contenir des connaissances expertes comme une analyse des changements entre deux cartes, les similarités climatiques entre années. Le principe général, illustré dans la figure 1.5, consiste à la substitution de la phase d'apprentissage dans le processus de classification par l'historique.

À partir de cet historique nous pouvons envisager plusieurs axes, comme les méthodes de transfert de connaissances visant à transposer un savoir acquis à une nouvelle situation, d'exploitation de connaissances à priori déduites automatiquement, comme les transitions entre les classes d'OCS ou encore les approches de fusion de données ou de classification. Ainsi, l'adaptation de domaine semble apporter des pistes de réponses.

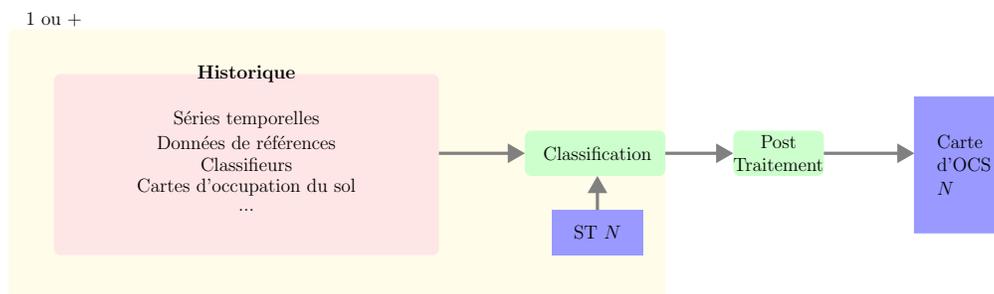


FIGURE 1.5 : Utilisation de l'historique pour la production de cartes d'OCS

En effet, ces méthodes proposent d'exploiter un domaine Source, pour lequel nous disposons de données de référence, afin de traiter un domaine Cible, qui en est dépourvu.

Ce manuscrit suit le plan suivant afin de présenter les travaux menés au cours de cette thèse.

Partie I Cette première partie décrit le contexte des travaux de thèse, présentant rapidement les concepts et enjeux de la cartographie de l'occupation des sols. Elle introduit les problèmes engendrés par la dépendance aux données de référence, des approches de production actuelles. Enfin, les objectifs généraux de la thèse sont présentés.

Partie II Cette partie présente les données et les méthodes exploitées. Le chapitre 2 présente les données de référence et les séries temporelles, ainsi que le Random Forest qui est un algorithme de classification supervisé. Le chapitre 3 propose une approche simple permettant de contourner le problème de dépendance aux données de référence.

Partie III Cette partie est consacrée aux approches d'adaptation de domaine, réponse de la littérature à notre problème. Le chapitre 4 présente les méthodes existantes présentant un intérêt pour la production de cartes d'occupation des sols. Le chapitre 5 propose d'exploiter l'analyse canonique des corrélations pour réduire les différences entre les séries temporelles de deux années différentes. Le chapitre 6 est consacré à l'application du transport optimal et des stratégies d'échantillonnage requises pour assurer des performances acceptables.

Partie IV Cette partie propose d'exploiter plusieurs années de données passées, en combinant les informations disponibles. Le chapitre 7 propose de modéliser les connaissances issues de l'historique. Ainsi, nous proposons une mise à jour des cartes existantes, et l'apprentissage d'un classifieur unique. Le chapitre 8 se consacre à la fusion des classifications naïves, à l'aide de systèmes de vote de complexité variables, de la fusion par Dempster-Shafer ainsi qu'une inférence bayésienne. Le chapitre 9 est consacré à une analyse comparative de ces méthodes.

Partie V Cette partie propose un passage à l'échelle des méthodes proposées. Elle exploite trois années de données Sentinel-2, sur une zone couverte par trente tuiles différentes.

Partie VI Cette partie offre des conclusions générales sur les travaux de thèse. Les perspectives méthodologiques et applicatives sont également discutées.

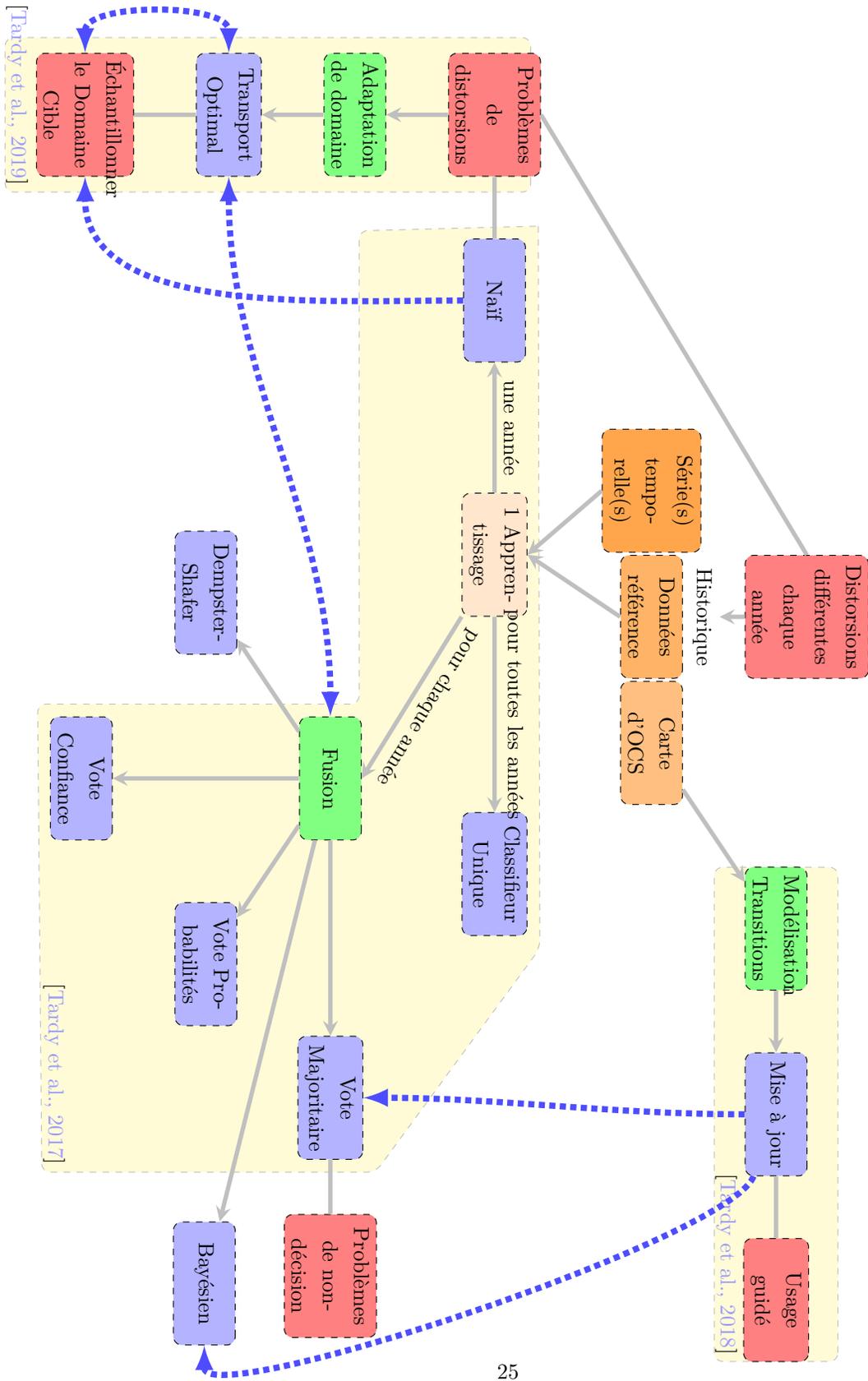


FIGURE 1.6 : Présentation générale des différentes méthodes proposées dans ce manuscrit ayant conduis à des publications (blocs orange)

Deuxième partie

Présentation du problème

Production de cartes d'occupation du sol

Au cours de ce chapitre, nous allons introduire les concepts liés à l'apprentissage et la classification supervisée qui représentent les meilleures alternatives de production de cartes d'OCS à l'heure actuelle. La production de cartes requiert en effet une étape de classification des images de la période concernée, qui permet d'attribuer à chaque pixel une étiquette d'OCS. Nous réalisons un apprentissage supervisé qui requiert des données de référence. Ce chapitre présente dans un premier temps les données de référence, ainsi que les images utilisées pour la classification, ce qui nous permet d'établir les contraintes inhérentes à la manipulation de telles données. Dans un second temps, nous présentons la classification supervisée et notamment les *Random Forests* ou *forêts aléatoires en français* (RF), qui constituent un outil performant pour la production de cartes d'OCS. Ce chapitre terminera sur le problème complexe de la validation de cartes d'OCS.

2.1 Présentation des données

2.1.1 Données de référence

Depuis 2007, des campagnes de relevés terrains d'OCS ont été menées par le CESBIO, sur un ensemble de plus de 1300 parcelles diverses (agricoles, naturelles ou artificielles). Ces campagnes sont menées trois à quatre fois par an, afin de couvrir à la fois les cultures d'été et d'hiver. La figure 2.1 illustre sur une image Formosat-2 en vraies couleurs, les parcelles visitées chaque années.

Cette zone relativement restreinte, en comparaison de l'emprise Sentinel-2, requiert une journée entière sur le terrain.

Cette campagne terrain, ne couvre pas l'intégralité de l'emprise Formosat-2. Néanmoins, elle est, dans son ensemble, représentative de l'OCS agricole sur la zone définie comme l'Observatoire Spatial Régional (OSR) MiPy¹, qui couvre une zone d'environ $24km \times 24km$ au sud-ouest de Toulouse.

¹<https://osrmipy.kalideos.fr/drupal/fr>

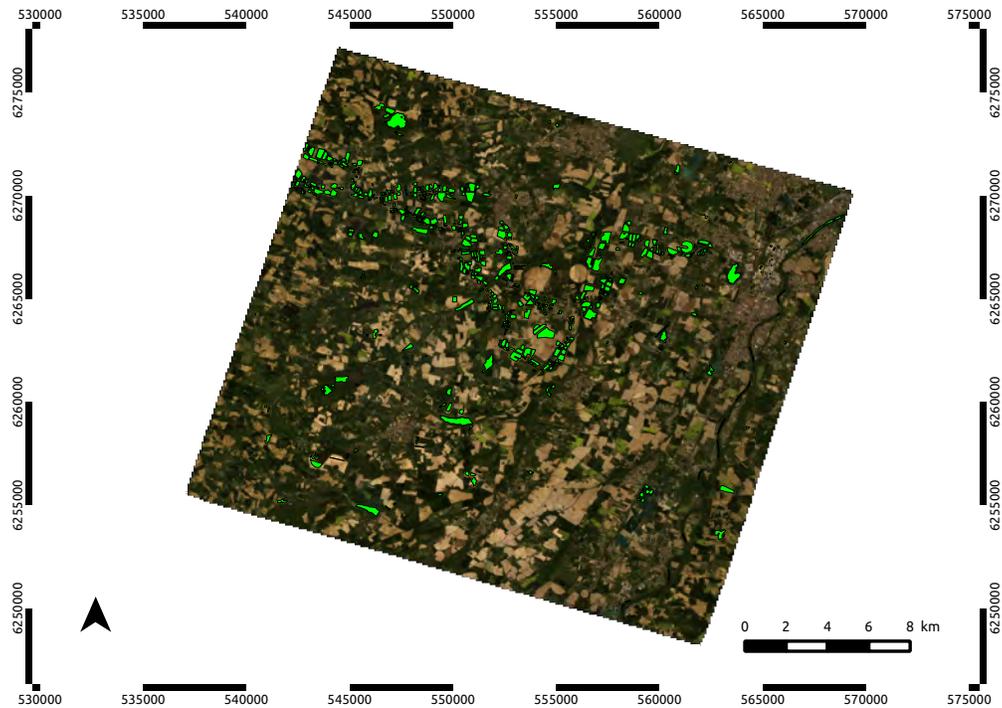


FIGURE 2.1 : Données de référence obtenues lors de la campagne terrain du CESBIO.

Une fois acquises, ces données sont nettoyées et vérifiées si besoin par photo-interprétation. Une analyse exhaustive de la cohérence des relevés terrain permet, à l'issue de chaque campagne, d'assurer un faible taux d'erreur d'environ 2 %. Les données de référence sont fournies sous forme de fichier vectoriel géoréférencé, aisément réutilisable. L'information fournie par ces données est double, avec d'une part l'emplacement des terres agricoles (forme des parcelles) et d'autre part une étiquette de classe associée.

Les campagnes terrains du CESBIO concernent essentiellement des parcelles agricoles. Les classes pérennes, caractérisé par une stabilité géographique pluri-annuelle, ne sont pas relevées systématiquement. Néanmoins, elles font l'objet d'une vérification annuelle.

Nous disposons de 7 années de données de référence acquises en continu entre 2007 et 2013. Les pratiques d'acquisition et les personnels en charge de ces campagnes ayant évolué aux cours de ces années, les données acquises ne sont pas homogènes sur la période.

Pour pouvoir les utiliser sans problèmes, il a fallu tout d'abord harmoniser ces données en assurant la cohérence des étiquettes. La figure 1.4 présente la nomenclature extraite à partir des 7 campagnes terrains. Dans cette figure, nous avons fait apparaître la nomenclature harmonisée utilisée dans ces travaux par les cercles colorés, pour un total de 17 classes d'OCS. Certaines classes ne sont pas présentes pour toutes les années, pour différentes raisons. Par exemple, le Chanvre n'apparaît qu'à partir de 2009 date d'implantation d'une chanvrière dans la région. À l'inverse, la distinction entre Friches et Jachères n'est plus systématiquement réalisée depuis 2010.

Nous avons été contraints de faire des choix lors de l'harmonisation des classes sur l'ensemble de la période. Certaines années, des classes supplémentaires sont relevées pour des besoins précis, comme la classe de bi-

culture, celles-ci sont supprimées. Par contre, des classes apparaissant de manière discontinue comme le Soja, celles-ci sont conservées. En effet, leur absence signifie sans doute un effet de la rotation des cultures. Leur conservation est donc justifiée. Enfin certaines classes sont décomposées pour des besoins thématiques par exemple les cultures irriguées et non irriguées. Ces classes sont regroupées sous une étiquette unique correspondant à la culture. La nomenclature hiérarchique finale est représentée dans la figure 1.4 par les cercles de couleurs.

Une fois que les données sont harmonisées, une vérification d’erreurs est réalisée, en superposant toutes les données de référence des 7 années. Le but de cette manipulation est de vérifier qu’une erreur de relevé ne ce soit pas glissée parmi les données malgré les étapes précédentes de préparation des données de référence. À partir de ces données nettoyées et harmonisées, la séparation entre jeux d’apprentissage et de validation est réalisée. Ce besoin est expliqué en détail dans la partie 2.3.1.

La table 2.1, présente pour chaque année disponible la quantité d’échantillons (pixels) disponibles par classe dans le jeu d’apprentissage. Les polygones sont ensuite discrétisés afin de fournir des échantillons représentant les pixels comme expliqué dans la section 2.2.2.

TABLE 2.1 : Nomenclature et nombre d’échantillons (pixel) dans la donnée de référence (apprentissage et validation).

Année Classes	2007	2008	2009	2010	2011	2012	2013
Feuillus	33 659	39 060	40 905	28 702	39 743	39 743	39 989
Conifères	10 160	13 112	6 486	3 703	3 611	3 611	3 611
Blé	66 116	49 848	23 854	66 047	340 803	58 476	97 825
Colza	27 651	12 933	25 937	13 869	67 104	9 885	40 508
Orge	1 937	5 908	3 564	1 203	35 799	12 055	20 270
Maïs	58 438	39 185	49 570	54 858	142 214	29 063	105 107
Tournesol	5 851	19 952	19 489	24 215	237 662	23 107	29 544
Sorgho	2 040	1 746	10 696	9 829	8 806	0	362
Soja	754	7 921	8 816	6 497	12 482	0	2 308
Surf.Art.	1 550	1 047	1 047	1 339	2 089	1 426	1 496
Jachères	16 148	5 145	3 396	0	35 110	0	0
Friches	1 089	1 299	9 954	4 142	10 357	10 357	14 208
Rivières	5 806	9 092	6 825	6 736	13 298	8 850	10 071
Lacs	14 294	9 997	10 090	20 070	4 615	4 440	4 508
Gravières	14 659	12 919	12 919	11 496	12 894	12 894	12 894
Chanvre	0	0	960	1 806	5 881	670	279
SEH	42 656	11 900	13 571	18 379	120 299	21 182	25 858
Total	302 808	241 064	267 568	272 891	1 092 767	235 759	408 840

2.1.2 Séries temporelles d’images Formosat-2

Formosat-2

Au cours de l’introduction, nous avons présenté les mérites des images de télédétection pour la production de cartes d’OCS, que ce procédé soit manuel ou automatique. Nous avons également abordé l’intérêt d’utiliser plusieurs images d’une même année afin de couvrir au maximum les différentes phases de la végétation sur un cycle phénologique. Cet intérêt motive le choix d’exploiter des séries temporelles d’images Formosat-2, qui concentre dans une série unique l’ensemble de l’information annuelle. Dans cette série, les différents canaux représentent les bandes spectrales aux différentes dates d’acquisition.

Formosat-2 est un satellite lancé par l’agence spatiale nationale de Taïwan en 2004. Pourvu d’une durée de vie estimée à 5 ans, il a cependant acquis des images durant 12 ans avant d’être mis hors service au cours de l’été 2016. Il disposait d’une caméra *Remote Sensing Intrument (RSI)* permettant l’acquisition d’images à une résolution de 8 m sur 24 km d’emprise (tuile) avec la possibilité d’une acquisition au mieux quotidienne. Cette caméra offre une acquisition sur quatre bandes spectrales regroupés dans la table 2.2

TABLE 2.2 : Bandes spectrales et longueurs d’onde associées pour Formosat-2.

Bande Spectrale	Longueur d’onde (nm)
Bleu (B)	450-520
Vert (V)	520-600
Rouge (R)	630-690
Proche InfraRouge (PIR)	760-900

Contrairement à Landsat et Sentinel-2, l’accès aux données était payant. Les images disponibles pour ces travaux ont donc été achetées pour les besoins du [CESBIO](#), avant le début de cette thèse. Cependant, plus de huit années de données sont disponibles avec une répartition disparate telle que le montre la figure 2.2. Dans cette figure, pour les années de 2006 à 2013 les images disponibles en fonction du jour de l’année sont indiquées par les étoiles colorées. Les bandes colorées indiquent les différentes saisons. Chaque étoile est placée sur un axe vertical indiquant le pourcentage de nuages présents sur l’image. Notons, dans un premier constat, que certaines périodes souffrent d’une absence totale d’images, car ces images n’ont pas été achetées en raison de la forte couverture nuageuse ou du faible taux de pixels exploitables.

Ce jeu de données a été privilégié par rapport à des jeux de données Landsat pour plusieurs raisons : le nombre d’années disponibles en continu, que nous n’avons pas sur les données Sentinel-2 ou Landsat 5 ou 8, ainsi que la résolution spatiale et temporelle meilleures que Landsat. Nous présenterons en fin de manuscrit des résultats obtenus avec les données Sentinel-2, afin d’évaluer le passage à l’échelle des méthodes.

Les images disponibles sont pré-traitées par l’algorithme [Multi-sensor Atmospheric Correction and Cloud Screening \(MACCS\)](#) [[Hagolle et al., 2010](#)], qui permet de corriger les effets de l’atmosphère sur les valeurs de réflectances (produit N2A) et de fournir des masques de nuages, de bords, et de saturation.

Harmonisation des données

Dans la mesure où huit années de données peuvent être utilisées, nous devons procéder à l’harmonisation des nouvelles données à l’instar des données de référence. Nous réalisons tout d’abord un masque d’emprise commune en effectuant la somme de tous les masques de bords disponibles. Ensuite nous appliquons un seuil

sur cette somme afin de conserver une emprise maximale. Ce seuil a été arbitrairement fixé à 20 % du total d'images disponibles. Puis nous appliquons ce masque d'emprise commune à toutes les images, ainsi qu'à ceux de nuages et saturation. À partir de ces images, nous constituons les séries temporelles de chaque année. Pour couvrir l'intégralité du cycle agricole, le choix a été fait de réaliser des années de 14 mois du 1^{er} octobre de l'année $N - 1$ au 31 décembre de l'année N (N est considérée comme l'année courante). Nous utilisons les images de 2006 uniquement pour compléter la série temporelle de 2007, ne disposant pas de données en 2005 d'une part, et d'autre part, les données de référence ne sont disponibles qu'à partir de 2007.

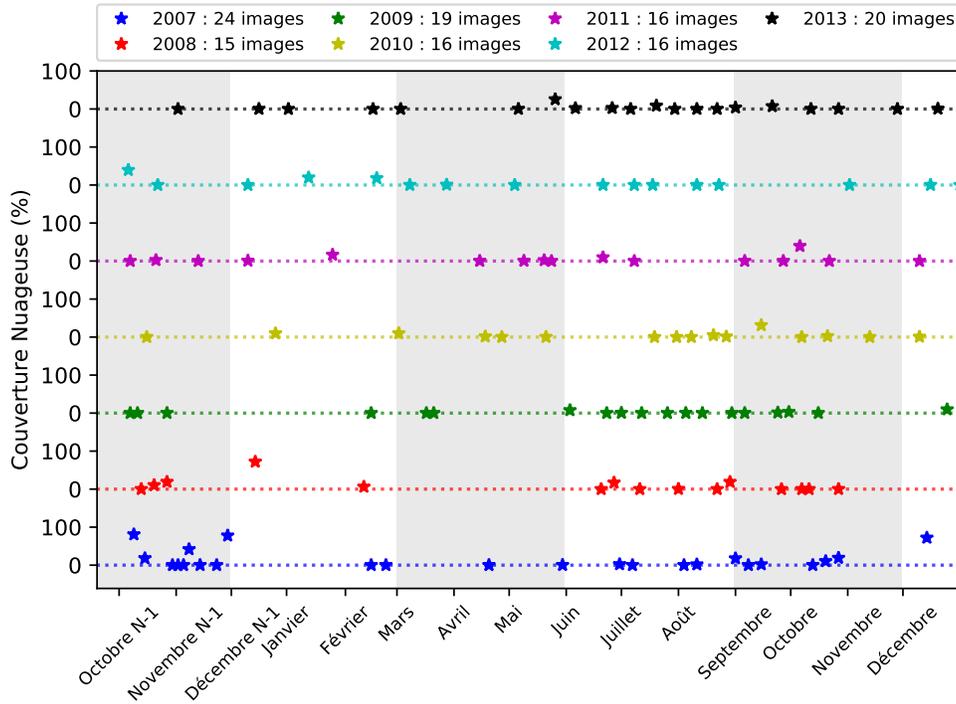


FIGURE 2.2 : Répartition des images (étoiles) Formosat-2 disponibles par année agronomique (octobre $N - 1$ à décembre N). Les bandes colorées indiquent un changement de saison. Les lignes en pointillés indiquent une faible proportion de nuage. L'écart entre une étoile et la ligne colorée correspondante implique une couverture nuageuse plus importante.

Une fois que l'on dispose des images avec la même emprise, nous souhaitons reconstruire des profils temporels continus (sans nuages). Nous utilisons une interpolation temporelle pixel à pixel permettant de corriger les valeurs aberrantes dans les séries temporelles. Il utilise une série temporelle de masque pour identifier tous les pixels nuageux, saturés et les éventuelles bordures communes. Pour un pixel masqué, sa nouvelle valeur est calculée par interpolation linéaire entre la valeur valide précédente et la suivante dans la série temporelle. Pour les extrémités de la série temporelle, la dernière valeur valide est dupliquée. Le choix de l'interpolation temporelle linéaire a été réalisé après des études de sensibilité sur les résultats de classification, en comparant l'interpolation linéaire avec une interpolation avec des splines. Un effet de la reconstruction des profils temporels est illustré dans la figure 2.3.

L’interpolation offrant des performances acceptables, nous ré-échantillons les profils temporels sur une nouvelle grille d’acquisition commune à toutes les années. L’utilisation du ré-échantillonnage n’introduit qu’un faible taux d’erreur qui n’altèrent pas les performances des cartes ainsi obtenues [Inglada et al., 2015].

Cependant, comme toute méthode d’interpolation, la qualité du signal reconstruit dépend en grande partie de celle des données en entrée. Ainsi, pour les séries temporelles souffrant d’une longue absence consécutive d’images, l’interpolation sera moins pertinente que pour les séries temporelles dont les images sont mieux réparties sur le cycle phénologique. La figure 2.2 permet de déduire que l’interpolation pour les années 2007, 2008, 2009 sera de moins bonne qualité que les quatre autres.

Dans notre cas, nous cherchons à simuler une image valide mensuelle, en fixant un pas de temps à 28 jours. Le ré-échantillonnage temporel est optionnel. Mais celui-ci permet d’offrir des séries temporelles pour lesquelles la correspondance entre les canaux est assurée quelque soit les années considérées, puisque nous disposons des mêmes dates virtuelles d’acquisition. Cette astuce facilitera l’application de la plupart des méthodes abordées dans ce manuscrit.

Indices spectraux

À partir des séries temporelles ré-échantillonnées, nous calculons des indices spectraux. Les bandes spectrales de Formosat-2 nous permettent d’en calculer deux parmi les trois généralement utilisés pour la production de cartes d’OCS [Inglada et al., 2015] : le *Normalised Difference Vegetation Index* (NDVI) et la brillance.

Le troisième indice est le *Normalised Difference Water Index* (NDWI). Il permet de mettre en évidence la présence d’eau par une combinaison des bandes spectrales PIR et Moyen InfraRouge (MIR). Cependant, Formosat-2 ne dispose pas de la bande MIR.

La brillance est calculée comme la norme des bandes spectrales, cet indice est très utilisé pour caractériser les zones urbaines témoignant d’un fort contraste (toit/asphalte).

$$brillance = \sqrt{B^2 + V^2 + R^2 + PIR^2}.$$

Le NDVI est utilisé pour caractériser la présence de végétation en combinant les bandes rouges et PIR :

$$NDVI = \frac{PIR - R}{PIR + R}.$$

Cet indice normalisé est compris entre -1 et 1. Une valeur proche de 0 indique un sol nu et une valeur proche de 1 la présence de végétation. Les valeurs négatives correspondent souvent à des pixels d’ombre ou d’eau.

Ces indices sont calculés pour chaque date de la série temporelle après ré-échantillonnage temporel. Au total la série temporelle de réflectance et les indices spectraux associés constituent une image de 108 bandes.

2.1.3 L’utilisation des données passées

Comme abordé précédemment, une production opérationnelle de carte d’occupation des sols nécessite une acquisition constante de données au fil des ans. Provenant de bases de données issues d’organismes comme le Ministère de l’agriculture, ou bien de campagnes terrain, les données de référence deviennent disponibles, mais bien souvent après la période à cartographier concernée. Les images satellites sont acquises régulièrement, notamment avec Landsat 8 et Sentinel-2. Cependant, toutes les images ne sont pas exploitables, soit à cause des nuages ou d’une erreur matérielle. Dans le cadre de cette thèse, nous nous plaçons dans le cas où les images de la période à produire sont disponibles, mais pas les données de référence correspondantes. Dans

le cadre des travaux de thèse, nous disposons des données de référence, nous les utilisons donc pour réaliser la validation des différentes cartes produites. Cependant, la validation sans données de référence demeure un problème conséquent, que nous avons choisi de ne pas aborder dans ces travaux.

Dans le cas de production opérationnelle, celle-ci peut s'envisager en deux temps : la livraison de la carte puis des métriques d'évaluation une fois les données de référence disponibles. Au moment de la production, nous disposons de toutes les données précédemment acquises sur la zone à traiter.

Cependant, utiliser des données issues de l'historique requiert des conditions très précises, car deux types de distorsions empêchent une utilisation directe.

La figure 2.3 illustre, sur un profil moyen de NDVI, les distorsions pouvant être observées entre des années. La courbe rouge représentant 2008 met en évidence le manque d'image couvrant le printemps et donc le cycle du Blé. Par exemple, le climat impacte directement tous les profils temporels des classes de végétation. Cette distorsion impacte automatiquement les approches exploitant des seuils numériques puisqu'elle fait varier l'amplitude du signal acquis.

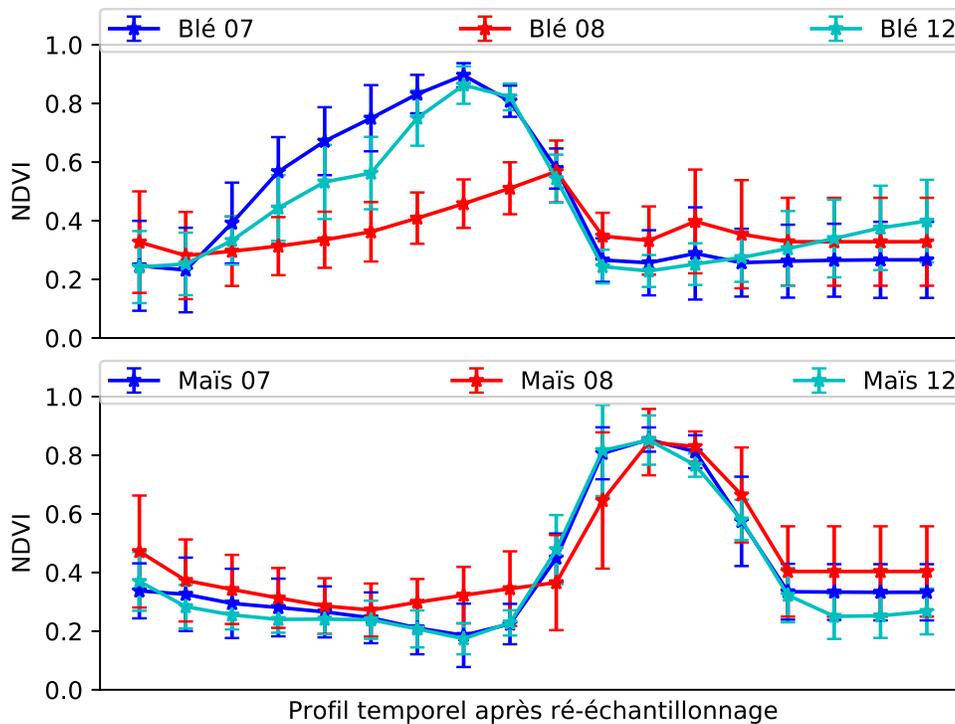


FIGURE 2.3 : Distorsions observables entre profils moyens de NDVI d'années différentes pour une classe de culture d'été et d'hiver. L'effet du ré-échantillonnage est clairement visible pour la culture d'hiver de 2008. Les barres d'erreurs représentent l'écart type mesuré sur l'ensemble des échantillons de la classe.

2.2 Classification Supervisée

2.2.1 Les Random Forests

Breiman et Cutler ont formalisé les *Random Forests* ou *forêts aléatoires en français* (RF) en 2001 [Breiman, 2001]. Cet algorithme combine les notions de *bagging* et de sous-espaces aléatoires. La phase d'apprentissage d'un *Random Forests* ou *forêts aléatoires en français* (RF) consiste en l'apprentissage de multiples arbres de décision entraînés chacun à partir d'un sous-ensemble des échantillons disponibles

Bagging

Le *Bootstrap aggregating* (*bagging*) est une technique utilisée pour améliorer la stabilité et la précision des algorithmes d'AA. L'objectif du bagging consiste à optimiser le compromis biais/variance. Pour rappel, un fort biais entraîne un sous-apprentissage, car l'algorithme "vise toujours à côté" du point valide. Inversement, une forte variance provoque un sur-apprentissage car la moindre petite variation dans les données d'entrée entraîne une grande variation sur les sorties attendues. À partir des échantillons du jeu d'apprentissage A de taille n , le *bagging* constitue M sous-ensembles d'échantillons A_1, \dots, A_m de taille n' par tirage aléatoire avec remise. Si $n = n'$ et que A est suffisamment grand, alors A_i contient au mieux $1/1 - e = 63.2\%$ échantillons uniques, le reste est alors constitué d'échantillons dupliqués. L'utilisation du bagging en entrée de l'apprentissage du RF force des variations dans les données d'apprentissage pour chaque arbre de décision. L'agrégation des prédictions de chaque arbre permet ainsi de réduire la variance. L'exploitation des différents paramètres du RF, détaillés plus tard dans cette section, permet de limiter le biais de chaque arbre de décision. Ainsi, l'ensemble d'arbres de décision entraîné permet d'offrir un bon compromis entre le biais et la variance du RF.

Arbres de décision

Les arbres de décision, ou CART pour *Classification And Regression Tree* en anglais, sont une des formes de classification les plus simples et intuitives.

À l'origine, les arbres de décision constituent une représentation de l'attitude humaine face à une question complexe, en simulant la décomposition en plusieurs problèmes simplifiés. Lier les différentes solutions entre elles permet de créer un arbre de décision. L'arbre de décision le plus utilisé est l'arbre de décision binaire, qui n'autorise que deux réponses à une question fermée. C'est le principe d'un raisonnement où l'on pèse le pour et le contre de chaque argument.

La figure 2.4 est une illustration d'un problème de la vie courante : est ce que je dois utiliser mon vélo ou prendre le bus pour mes déplacements du lendemain ? Dans ce genre de situation, le nombre de paramètres entrant en jeu peut être très important et provenir de sources très différentes. Ainsi dans la balance, nous retrouvons des paramètres arithmétiques comme la distance, des facteurs extérieurs comme la météo ou des facteurs humains comme l'affluence dans les moyens de transports en commun. Pour cette exemple nous retiendrons trois critères que l'on peut formuler sous forme de question :

1. Est ce qu'il va faire beau ? (météo ok ? dans la figure 2.4)
2. Est ce que la distance à parcourir est acceptable ? (distance ok ?)
3. Est ce que je possède des vêtements de pluie ? (vêtements ok ?)

La réponse à ces trois questions est forcément oui ou non. En fonction de la question et de la réponse, la décision prend une orientation différente. Dans la figure 2.4, trois cas différents de raisonnements sont illustrés à partir de ce problème simple. Chacun commence le raisonnement avec une des questions ci-dessus (en vert dans la figure), puis enchaîne les autres questions (en bleu) en fonction des réponses (indiquées sur les flèches) jusqu'à aboutir à une décision finale (en rouge).

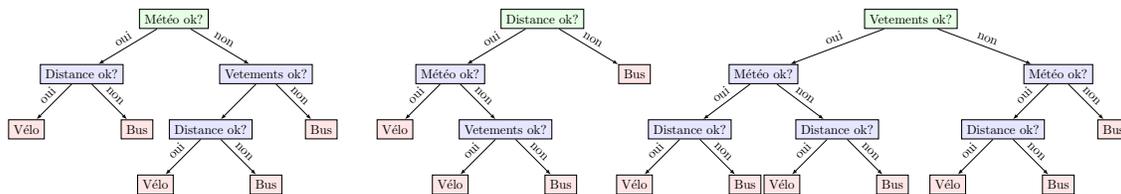


FIGURE 2.4 : Illustration de l'utilisation d'un arbre de décision pour un problème quotidien. Il représente la capacité humaine d'analyse et d'association de facteurs hétérogènes.

Cet exemple illustre bien le principe des arbres de décision en introduisant les concepts qui font sa force. Alors que seulement deux choix sont disponibles, le nombre de chemins possibles pour atteindre cette décision est bien plus grand. Le choix de la question initiale a également un impact sur la structure de l'arbre de décision obtenu. Cette différence est bien illustrée par les deux arbres les plus à droite dans la figure 2.4. Alors que l'arbre du milieu n'utilise qu'une seule fois chaque question, l'arbre de droite les utilise plusieurs fois. Ces particularités seront détaillées par la suite. Cet exemple permet surtout d'illustrer les concepts ayant amené à la création des arbres de décision. En effet, cet exemple représente une inférence, en d'autres termes une application de l'arbre de décision sans aucun apprentissage. Dans le cas d'une application concrète en informatique, on considère des échantillons représentatifs d'un problème. Ces échantillons sont séparés en sous-ensemble lors de l'apprentissage, permettant ainsi de construire des arbres différents.

Un arbre de décision est constitué de deux éléments : les noeuds et les branches. Trois types de noeuds existent, la racine représente l'état initial. C'est le point de départ du parcours au sein de l'arbre. On distingue ensuite les noeuds terminaux, ou feuilles, qui représentent la décision prise par l'arbre. On distingue la feuille des autres noeuds par le fait qu'elles n'ont pas de descendant, c'est-à-dire que lorsque l'on arrive dans une feuille le parcours est terminé. Les autres noeuds ont tous un parent et deux fils, le fils gauche et le fils droit qui représentent la séparation des échantillons en deux nouveaux sous-ensembles. Un arbre de décision se décompose en niveaux. Par défaut, la racine correspond au niveau 0, ses descendants directs (droit et gauche) sont au niveau 1 et ainsi de suite. La feuille ayant le niveau le plus haut définit la profondeur maximale, qui est souvent un critère d'arrêt lors de l'apprentissage. Cependant, au sein d'un même arbre, les différentes feuilles peuvent avoir des profondeurs différentes. Les branches représentent une réponse au test réalisé dans le noeud, le nombre de branches pour chaque noeud est soit nul (feuille) soit de deux.

Les noeuds et les branches constituent les briques d'un arbre de décision. Pour le construire trois éléments sont requis :

- Une règle permettant d'affirmer qu'un noeud est une feuille
- Une règle permettant d'attribuer à une feuille une étiquette de classe. Généralement, c'est la classe ayant le plus d'échantillons dans cette feuille qui est choisie.
- Une fonction de séparation permettant de répartir les échantillons entre les deux fils d'un noeud.

La fonction de séparation est une fonction simple, amenant une réponse binaire.

Pour déterminer cette fonction, l'algorithme passe en revue les différentes primitives et cherche celle offrant la meilleure séparation entre les échantillons présents dans le noeud. Afin de choisir la meilleure fonction de séparation, on calcule un critère de pureté. Ce critère est libre de choix avec peu d'incidence sur l'arbre construit, de manière générale pour les arbres de classification on utilise l'indice de Gini. Cet indice est calculé à partir d'un ensemble d'échantillons p de taille n ayant des étiquettes à valeurs discrètes entre $1, \dots, K$ correspondant aux K classes présentes dans le noeud considéré. Au sein de chaque noeud, la quantité n_k représente le nombre d'échantillons appartenant à la classe k . La formulation de l'indice de Gini est alors :

$$I_{Gini}(p) = \sum_{k=1}^K p_k \sum_{i \neq k} p_i = \sum_{k=1}^K p_k (1 - p_k) = 1 - \sum_{k=1}^K p_k^2$$

où $\sum_{i \neq k} p_i = (1 - p_k)$ représente la probabilité que l'échantillon soit mal étiqueté, si son étiquette était choisie aléatoirement parmi celles disponibles dans le noeud. La valeur p_k représente le ratio entre les échantillons étiquetés par la classe k par rapport à l'ensemble des échantillons du noeud soit $p_k = \frac{n_k}{n}$. L'indice d'impureté de Gini est maximal lorsque toutes les classes sont équitablement représentées dans le noeud, et minimal lorsque le noeud est pur (c'est-à-dire qu'une seule classe est représentée). La séparation retenue est celle qui maximise l'écart entre la pureté du noeud actuel et celles de ses descendants.

La construction d'un arbre de décision se poursuit jusqu'à satisfaire au moins un critère d'arrêt, en règle générale au nombre de trois et initialisés par l'utilisateur :

1. La profondeur maximale
2. Une stabilité du critère de pureté ou une valeur minimale de pureté
3. Un nombre d'échantillons minimal dans un noeud

Une fois au moins l'un de ces critères atteint, le noeud est transformé en feuille et une étiquette de classe lui est attribuée. L'arbre est construit par divisions successives en commençant par le noeud le plus à gauche. Une fois que les divisions successives produisent une feuille à une profondeur d , l'algorithme poursuit les séparations à partir du noeud droit du niveau $d - 1$. L'arbre est définitivement construit lorsque le noeud le plus profond à droite est transformé en feuille. Les critères d'arrêts concernent chaque feuille indépendamment. Ainsi, on peut obtenir des feuilles à différents niveaux au sein du même arbre. De même, les séparations sont indépendantes entre chaque noeud, il n'est donc pas impossible qu'une primitive apparaisse plusieurs fois dans un même arbre. Les critères d'arrêt 2 et 3, les plus importants, sont complémentaires. Ils permettent d'éviter le cas de sur-apprentissage, où une feuille ne contiendrait qu'un seul échantillon, mais écarte également le risque de diviser inutilement un noeud déjà pur. Ces deux paramètres impactent la capacité de généralisation de l'algorithme, et sa capacité à classer correctement un nouvel échantillon. Le sujet de la profondeur maximale fait débat : celle-ci permet essentiellement de gagner du temps lors de la classification, puisque le parcours de l'arbre est plus rapide. Cependant, elle peut aussi bloquer une séparation de classes très proches en termes de primitives.

La construction d'un arbre de décision peut également être vu comme un algorithme de divisions successives. La figure 2.5 illustre les divisions successives d'un problème factice de classification à 4 classes en deux dimensions. Ces données sont générées à partir de fonction gaussiennes. Le principe de cette figure permet d'offrir une évaluation des différentes étapes du processus de création de l'arbre de décision.

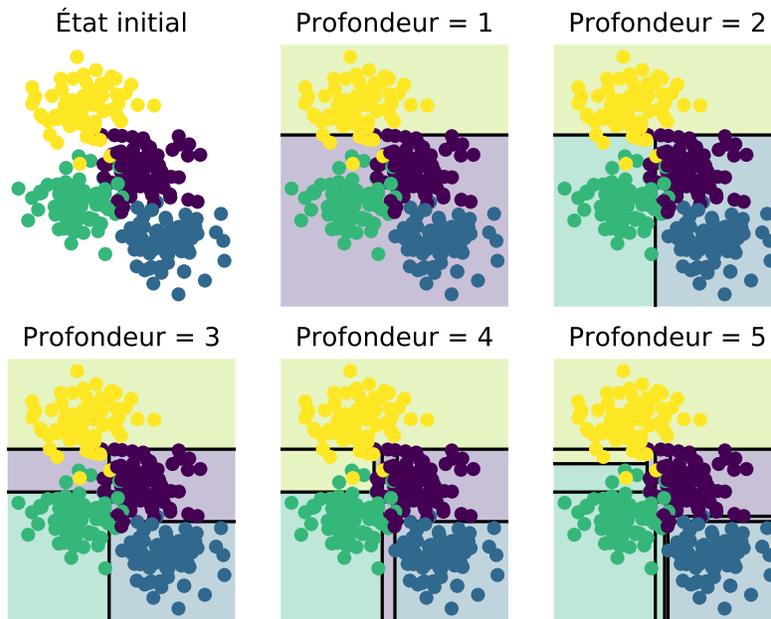


FIGURE 2.5 : Visualisation des séparations successives d'échantillons tests en 2D par un arbre de décision. Les couleurs différencient les classes considérées (4), pour des profondeurs dans l'arbre variant de la racine à une profondeur maximale de 5.

L'état initial représente les données que l'on cherche à classer. Les couleurs indiquent des classes différentes. Dans cette figure, les différents noeuds sont illustrés en fonction de leur réponse à la fonction de séparation. Pour la profondeur 1, les échantillons sont divisés en deux sous ensembles. À ce stade, seules deux classes peuvent être prédites par l'arbre de décision, la classe jaune et la classe violette indiquées par la couleur des rectangles, tout pixel classé dans un de ces deux espaces se verra attribuer l'étiquette correspondante. La division supérieure est très bonne avec une seule classe présente à l'intérieur, il représente une feuille puisque jamais divisé par la suite. À l'inverse la partie basse contient des échantillons des quatre classes. Pour chaque profondeur, nous pouvons voir les frontières entre les classes, où chaque noeud est divisé en deux nouveaux noeuds. Nous voyons qu'à partir d'une profondeur 4, l'algorithme détermine de grandes zones pures (une seule classe à l'intérieur) et des zones exiguës avec peu d'échantillons à l'intérieur. La multiplication de telles frontières de séparation peut causer une situation de sur-apprentissage, par perte de la capacité à généraliser.

Pour la classification d'images de télédétection, un arbre de décision est considéré comme trop faible, car trop sensible au sur-apprentissage et sans capacité satisfaisante de généralisation, pour appréhender la richesse apportée par de telles données. Il est cependant très rapide à construire et très lisible puisque l'information sur la séparation au niveau de chaque feuille est conservée, ce qui peut améliorer la compréhension des résultats.

Construire une forêt aléatoire

Le RF a été proposé [Breiman, 2001] comme un algorithme tirant parti des avantages des arbres de décision en atténuant leurs défauts. Le RF est un comité d'arbres de décision différents de ceux présentés précédemment. En effet, trois critères sont ajoutés lors de la création des N arbres au sein d'un RF :

1. Chaque arbre est construit à partir d'un sous-ensemble aléatoire issu du *bagging*
2. Pour chaque noeud, on considère un seul un sous-ensemble aléatoire de primitives lors de la recherche de la fonction de séparation
3. Les arbres sont systématiquement construits jusqu'à leur profondeur maximale

L'utilisation de sous-ensembles aléatoires pour les échantillons d'apprentissage, ainsi que pour les primitives considérées lors de la séparation d'un noeud, permettent d'introduire une diversité entre les arbres. Mais cela nécessite que le nombre d'échantillons soit suffisamment grand, afin d'éviter une trop forte corrélation entre deux sous-ensembles puisqu'il s'agit d'un tirage avec remise. Sous réserve de cette contrainte, les arbres ainsi générés sont tous différents et faiblement corrélés entre-eux. Deux arbres sont définis faiblement corrélés, si les résultats sont identiques pour la majorité des échantillons, mais que les erreurs commises sont différentes.

Le *bagging* permet d'utiliser un sous-ensemble différent pour chaque arbre. Les échantillons non sélectionnés pour un arbre sont nommés échantillons **Out OfBag (OOB)**. Ceux-ci ne participent pas à la construction de l'arbre, mais peuvent être utilisés pour calculer l'erreur **OOB** qui comptabilise le nombre d'erreurs de classification commise lors de la classification. On utilise cette valeur comme indicateur de la capacité de l'arbre à généraliser ses règles de décision ; en d'autres termes elle permet de vérifier qu'il n'y a pas de sur-apprentissage.

L'analyse des primitives utilisées par le RF s'avère plus complexe que pour un arbre de décision simple, car il faut prendre en compte le facteur aléatoire de sélection des primitives pour un noeud considéré. Ainsi, avec beaucoup de malchance une primitive peut n'être jamais vue par un noeud au cours de la construction des arbres.

Elle permet néanmoins un aperçu correct des primitives les plus utilisées lorsque l'intégralité de la forêt est considérée au lieu d'un seul arbre.

Une fois l'apprentissage terminé, la classification se réalise échantillon par échantillon, chacun parcourant l'ensemble des N arbres de la forêt. Chaque arbre choisit une classe. Les différentes réponses sont recueillies sous forme de voix. La décision finale est prise en choisissant la classe ayant obtenu le plus de voix, sur le principe du vote majoritaire. Des fonctions de pondération peuvent être exploitées pour influencer sur l'importance d'une voix au détriment d'une autre mais ces cas d'utilisation restent rares [Bonissone et al., 2010].

La réponse d'un RF peut être obtenue sous de multiple format : l'étiquette unique décidée par vote majoritaire, mais également la réponse de chaque arbre individuellement. En balayant les différents arbres et en associant un poids (en général unitaire) à une étiquette, nous générons un vecteur de voix. Ainsi, celui-ci est aisément transformable en probabilité, en divisant chaque élément du vecteur de voix par le nombre d'arbres. Ce vecteur devient alors une information de probabilité par classe très utile pour déterminer le comportement du RF face à un échantillon. La table 2.3 montre un exemple de vecteurs de probabilités factices pour un problème à 5 classes. Dans cette table, chaque ligne représente un échantillon différent, et les valeurs des colonnes représentent le ratio du nombre de voix pour la classe associée. La dernière colonne

représente la classe choisie par le RF après le vote majoritaire. À partir de cet exemple, plusieurs cas sont observés :

- Une classification franche : échantillons 0 et 2 où la majorité est très forte
- Une classification mitigée : échantillon 3, ce cas est assez fréquent lors de la classification d'un échantillon altéré (interpolation) ou mixte (bordure) ou issu d'une classe minoritaire.
- Une classification incertaine : échantillon 1 où l'écart entre les différentes probabilités est très faible

Ces valeurs de probabilités sont peu utilisées. En règle générale, les utilisateurs se contentent de la notion de confiance. Pour le RF la confiance correspond à la valeur de probabilité associée à la classe choisie. Dans notre exemple la confiance associée à la classification de l'échantillon 0 en tant que classe 5 a une valeur de 0.9 ce qui est une très bonne confiance. À l'inverse, pour l'échantillon 1 la confiance ne vaut que 0.25, incitant l'utilisateur à n'accorder que peu de crédit à cette décision. Il est également possible d'exploiter la notion de marge qui représente l'écart entre les deux plus grandes valeurs de probabilité. Plus la marge est importante plus le classifieur est confiant dans sa décision.

TABLE 2.3 : Exemple de vecteurs de probabilités obtenu par un RF pour 4 échantillons distincts. Les colonnes représentent la classe avec la valeur de probabilité associée. La dernière colonne représente le choix réalisé par le RF. Les lignes représentent les échantillons avec un identifiant différent.

Échantillons	Classes					Choix
	1	2	3	4	5	
0	0.1	0	0	0	0.9	5
1	0.2	0.2	0.2	0.15	0.25	5
2	1	0	0	0	0	1
3	0	0.22	0.78	0	0	3

Si la mise en oeuvre informatique d'un RF est relativement simple, différents types d'optimisations peuvent être utilisés pour accélérer les calculs ou réduire l'espace mémoire. Il existe de nombreuses implémentations libres des RF, parmi lesquelles les plus populaires sont celle de Scikit-Learn [Pedregosa et al., 2011] ou celle d'OpenCV [Bradski, 2000]. Le choix d'une mise en oeuvre existante a été privilégié pour réduire les risques d'erreurs et bénéficier de codes testés par ailleurs. Le langage de programmation constitue un autre critère de choix, pour permettre une intégration simple avec le reste de la chaîne de traitement. L'accès facile aux différents paramètres de l'algorithme constitue également un critère de sélection.

Choix techniques

Pour cette thèse nous avons choisi d'exploiter la bibliothèque d'AA Shark [Igel et al., 2008]. Cette librairie entièrement développée en C++, présente plusieurs avantages pour la production opérationnelle.

L'apprentissage est optimisé pour exploiter le parallélisme : chaque arbre de la forêt peut être construit en même temps offrant un gain de temps considérable, sous réserve d'avoir le matériel adéquat. De même la classification exploite la notion de *batch* caractérisé par un petit ensemble d'échantillons. Cela implique qu'au lieu de ne traiter qu'un seul échantillon arbre par arbre, tous les échantillons du *batch* sont traités

simultanément avant de récupérer les résultats de chaque arbre. Associé à la classification en parallèle, l'approche du découpage en *batch* permet un gain d'efficacité.

Nous utilisons également la librairie [Orfeo ToolBox \(OTB\)](#) [Inglada and Christophe, 2009], qui regroupe de nombreux algorithmes de traitement d'image, notamment les images de télédétection. Elle intègre de nombreux algorithmes d'AA, entre autres, ainsi que de nombreux outils exploitables pour implémenter des traitements exploitant des images de taille quelconques. L'OTB présente l'avantage d'exploiter au mieux les ressources disponibles : nombre de processeur, mémoire disponible pour optimiser l'utilisation des algorithmes. Elle gère les aspects de streaming (parcours de l'image) et de multi-threading permettant de réaliser les traitements des pixels en parallèle.

2.2.2 Chaîne de production d'une carte d'occupation des sols

Le processus de classification, malgré son rôle clef, n'est qu'une étape dans la production d'une carte d'OCS. La figure 2.6 offre une représentation de l'intégralité de la chaîne de traitement et de ses différentes étapes dans le cadre d'une classification supervisée.

À ce stade, nous considérons également que les images de télédétection sont prêtes à être utilisées, et que les différents masques sont disponibles. Ces éléments constituent la base de données Images (images pré-traitées par [MACCS](#) ou autre). Les données de référence sont également prêtes à l'emploi.

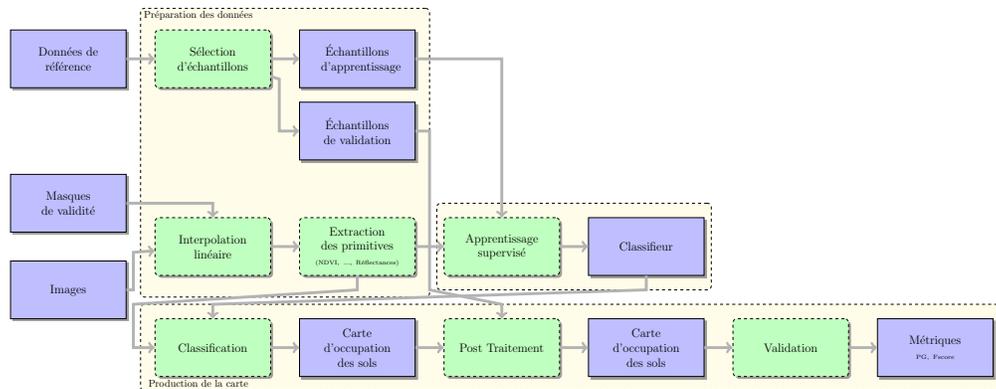


FIGURE 2.6 : Chaîne de traitement pour la production opérationnelle d'une carte d'occupation des sols (illustration d'origine : [Inglada et al., 2015]).

La première étape consiste à réaliser la séparation des données de référence en deux jeux de données représentatifs pour l'apprentissage et la validation. Les profils temporels sont reconstruits à partir des images, par interpolation linéaire. Au terme de cette étape, toutes les altérations détectées, comme les nuages, la saturation, ou les variations des bords de l'images sont interpolés. Le contenu des images présente exclusivement des valeurs de réflectances. Nous pouvons alors calculer les indices spectraux.

Une fois cette préparation des données terminée, nous réalisons l'extraction des échantillons d'apprentissage. Au cours de cette étape chaque échantillon se voit attribuer l'étiquette correspondante issue de la donnée de référence. Cette étape permet de réduire le volume de données à manipuler lors de l'apprentissage, en ne traitant que les échantillons étiquetés. Après l'extraction des échantillons, la phase d'apprentissage permet d'entraîner le classifieur supervisé. Lors de la phase de classification, le classifieur entraîné est appliqué

à la série temporelle interpolée, ce qui permet de produire une première version de la carte d'OCS. L'étape suivante, optionnelle, regroupe des opérations de post-traitement, par exemple un lissage, une application d'un masque de bordure ou encore une étape de fusion de classifications, amenant à la production de la version définitive de la carte.

Cette carte est alors validée par l'exploitation des données de référence dédiées. Puis les différentes statistiques sont alors estimées (cf. 2.3.1). Dans le cadre de l'utilisation des RF, les produits en sortie sont au nombre de trois : la classification (la carte d'OCS), la carte de confiance et la carte de probabilité (une image avec autant de dimensions que de classes).

2.3 Valider une carte d'occupation des sols

En classification, la validation désigne l'évaluation quantitative des décisions prises par un classifieur. La validation passe généralement par l'utilisation de données de référence qui sont comparées aux décisions du classifieur, permettant ainsi le calcul de métriques.

Les métriques calculées doivent toujours être positionnées dans un contexte d'utilisation pour être correctement interprétées. Dans le cadre de l'OCS, les statistiques sont des indicateurs satisfaisants de la qualité de la carte obtenue. Néanmoins, il est d'usage que l'opérateur évalue globalement le rendu visuel de la carte afin d'estimer la pertinence de la classification. En effet, les métriques ne permettent pas de mettre en valeur des erreurs qui pèsent peu en termes de pixels, mais qui sont très visibles à l'oeil, car nous faisons un produit cartographique [Derksen et al., 2019].

La validation visuelle permet également d'identifier les zones complexes à classer de l'image, comme les bords de parcelles où l'on recense généralement une forte variabilité des classes choisies.

Pour évaluer les performances des différents algorithmes, nous utiliserons les métriques standards de la validation, qui utilisent les données de référence.

En effet, nous disposons d'un jeu de données complet permettant d'évaluer ainsi les méthodes proposées. Afin de mesurer la stabilité des méthodes, mais aussi la différence significative entre deux approches, et ainsi définir une marge de performances dans laquelle une méthode donnée doit se situer, nous utilisons les différents tirages aléatoires de données de référence et les différentes combinaisons possibles entre les années, pour calculer des intervalles de confiance à 95 %. Un intervalle de confiance est centré sur la moyenne des valeurs mesurées, et indique les limites maximales entre lesquelles 95 % des mesures sont représentées.

2.3.1 Mesures de qualité d'une classification

Il est d'usage de présenter les résultats d'une validation sous la forme d'une matrice de confusion [Congalton and Mead, 1983]. Une matrice de confusion confronte les prédictions du classifieurs (en colonne) avec la réalité connue sur les échantillons testés (en ligne), ce qui donne la représentation telle que présentée dans la table 2.4 pour un problème de classification à N classes. À partir de cette matrice plusieurs métriques peuvent être calculées.

Dans cet exemple, les quantités n_{ij} représentent le nombre d'échantillons (dans notre cas de pixels) présents dans la donnée de référence de validation sous l'étiquette i et prédite comme appartenant à la classe j . Ainsi les valeurs sur la diagonale représentent les échantillons correctement classés, tandis que le reste de la matrice constitue les erreurs de classification. La matrice de confusion s'avère rarement symétrique, et l'objec-

TABLE 2.4 : Exemple de matrice de confusion

Étiquette	1	...	N
1	n_{11}	...	n_{1N}
...
N	n_{N1}	...	n_{NN}

tif est d'obtenir une matrice où seule la diagonale est non nulle. Dans un cas réel, une telle matrice implique très probablement un sur-apprentissage, ou une utilisation des données de validation lors de l'apprentissage. Cela explique l'extrême importance donnée à la distinction entre jeu d'apprentissage et de validation. De même, il est aussi fortement recommandé que la donnée de référence utilisée soit représentative de la zone d'étude afin que l'information qu'elle apporte soit pertinente [Congalton and Green, 2002]. Si cette condition est respectée, alors les métriques calculées sur la base de cette matrice de confusion seront représentatives de la qualité de la classification.

Un des indices les plus utilisés est la **Précision Globale** (*Overall Accuracy* en anglais) (PG), puisque il représente le nombre d'échantillons correctement prédits (somme des valeurs de la diagonale) divisé par le nombre total d'échantillons de validation ($total_{val} = \sum_{i=1}^N \sum_{j=1}^N n_{ij}$) :

$$PG = \frac{1}{total_{val}} \sum_{i=1}^N n_{ii}.$$

La PG permet de donner une vue globale des performances de la classification. Cependant, même un classifieur aléatoire possède une PG supérieure à 0.

Ainsi, de nombreux utilisateurs lui préfèrent le coefficient Kappa [Cohen, 1960]. Ce coefficient cherche à minimiser l'impact de la chance sur le taux de bonnes classification :

$$Kappa = \frac{PG - p}{1 - p}$$

, où p représente le pourcentage de bonnes classifications attribués au hasard :

$$p = \frac{1}{total_{val}^2} \sum_{i=1}^N \left(\sum_{j=1}^N n_{ij} \right) \left(\sum_{j=1}^N n_{ji} \right).$$

Une convention définit qu'un Kappa supérieur à 81% indique une excellente classification [Landis and Koch, 1977]. On critique cependant le Kappa sur plusieurs aspects [Powers, 2012]. Un premier litige concerne la formulation de p qui ne serait pas représentatif du facteur chance [Foody, 1992]. Une seconde critique concerne la complexité requise pour comparer les Kappa obtenus par des méthodes de classification différentes [Pontius Jr and Millones, 2011], ainsi que les nombreux cas d'interprétations faussés de celui-ci [Maclure and Willett, 1987]. Enfin un autre constat est qu'il reste très corrélé à la PG [Liu et al., 2007], c'est pourquoi nous allons essentiellement exploiter la PG comme indicateur global de performance [Stehman, 1997].

Cependant, les indices globaux sont sensibles aux déséquilibres des classes dans la matrice de confusion. En effet, les classes très représentées ont plus de poids dans la PG, masquant ainsi des variations pour les classes moins bien représentées. Or il est souvent appréciés d'évaluer les comportements de ces classes.

Dans cette optique deux métriques par classes sont exploitées : le rappel et la précision.

Le rappel représente le ratio d'échantillons appartenant à la classe i correctement prédit. On le mesure en divisant le nombre de bien classés par la somme des échantillons de la ligne concernée :

$$\text{rappel}_i = \frac{n_{ii}}{\sum_{j=1}^N n_{ij}}.$$

La précision d'une classe représente le pourcentage d'échantillons bien classés par rapport à l'ensemble des prédictions réalisées pour cette classe :

$$\text{précision}_i = \frac{n_{ii}}{\sum_{j=1}^N n_{ji}}.$$

Cependant, il est plus courant d'utiliser un compromis entre ces deux indices par l'exploitation de la mesure de F_1 Score (FScore) par classe. On calcule le FScore comme la moyenne harmonique de ces deux indices :

$$FScore_i = 2 \times \frac{\text{rappel} \times \text{précision}}{\text{rappel} + \text{précision}}.$$

Grâce à l'utilisation conjointe de la PG et du FScore, nous sommes en mesure de réaliser une analyse statistique précise des performances de classification. Dans le guide des bonnes pratiques pour la classification de cartes d'OCS [Olofsson et al., 2014], on recommande d'appliquer les proportions des classes lors du calcul des métriques : c'est la correction de Card [Card, 1982]. Cependant dans ce cas précis, les données de référence sont conçues pour être représentatives de la zone d'étude, ce qui n'affecte que très peu les performances corrigées avec cette méthode. Par conséquent, nous n'appliquons pas ce type de corrections.

2.3.2 Validation des méthodes de vote

Le vote peut produire trois types de décisions : une bonne décision, une mauvaise décision ou une non-décision. Cette dernière survient en cas d'égalité entre au moins deux classes. Ce genre de situation s'avère plus fréquente dans le cas où il y a deux votants.

En AA, la validation des résultats d'un vote suit des règles particulières. On ne comptabilise pas les échantillons non décidés lors du calcul des métriques standards [Lam and Suen, 1997]. En effet, dans l'AA une non-décision ne désigne pas une erreur, car elle représente un état dans lequel l'algorithme ne prend pas de décision plutôt qu'une potentiellement mauvaise.

Cependant cette justification n'est pas forcément valable dans notre contexte d'étude. En effet, une non-décision correspond à un pixel non exploitable par un utilisateur final. Les deux calculs représentent un intérêt différent. En effet, le calcul ne comptabilisant les non-décisions permet d'évaluer la qualité des décisions de la méthode. Inversement, les comptabiliser permet une évaluation de la méthode dans l'optique d'une utilisation de la carte.

C'est pourquoi, dans ce travail les non-décisions seront comptabilisées comme erreurs lors du calcul des métriques. Néanmoins, nous présenterons la différence entre ces deux calculs dans le chapitre 8.1.6

La classification Naïve

Ce chapitre propose une solution simple de production de cartes d'OCS, sans utiliser les données de référence de la période correspondante. Après avoir détaillé les conditions d'exploitation requises, nous proposons d'exploiter un classifieur déjà entraînés sur des données passées. Nous proposons également une différence de traitement selon le comportement annuel de la classe.

3.1 Principe de l'approche

Une manière simple de vérifier que le problème exposé dans la section 1.3.2 posé est réel, consiste à utiliser un classifieur précédemment entraîné sur une année disponible au sein de l'historique des données pour produire la carte de l'année en cours. Cette simple opération a été baptisée classification naïve ou cas naïf. Elle peut être comprise comme le cas opposé à la classification supervisée. La figure 3.1 représente la chaîne de traitement lors d'une classification naïve, la différence avec le cas supervisé 1.3 vient de l'année d'origine des échantillons d'apprentissage (en jaune dans la figure) par rapport à la série temporelle traitée (en magenta).

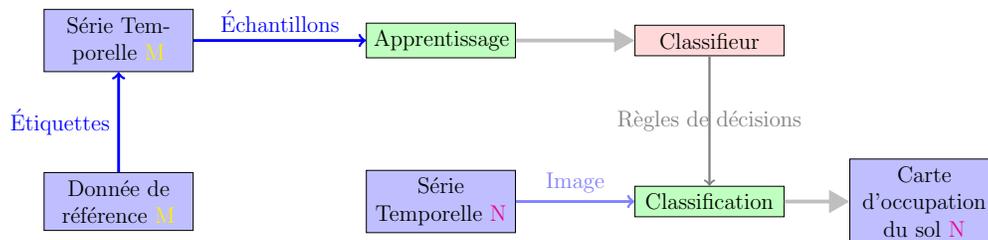


FIGURE 3.1 : Chaîne de traitement utilisée dans le cadre de la classification naïve.

Cette approche porte bien son nom, en effet aucune variation possible n'est prise en compte lors de ce processus. Cependant un certain nombre de conditions doivent être remplies afin de réaliser une classification

naïve.

La première, est de s’assurer de la cohérence des séries temporelles utilisées. Pour cela nous réalisons une interpolation linéaire temporelle sur une grille régulière. Cette opération permet d’obtenir une série temporelle complète (sans absence de données dues aux nuages) et également de contraindre les dates qui composent la série temporelle [Inglada et al., 2015]. Nous fixons une grille temporelle, avec un pas de temps de 28 jours, qui débute au 1^{er} octobre de l’année $N - 1$ et termine au 31 décembre de l’année courante N . On obtient ainsi pour chaque année de l’historique les mêmes dates d’acquisitions virtuelles. En d’autres termes, quelque soit l’année qu’il cherche à classer, le classifieur traitera toujours les mêmes primitives, ce qui assure la production d’une classification cohérente.

La seconde condition porte sur la cohérence entre les domaines, ce point largement décrit pour l’adaptation de domaine est valable pour la classification naïve. Le principal problème à gérer avec le jeu de données utilisé, est l’absence de certaines classes pour une année dans la campagne terrain. Logiquement, ces classes sont impossibles à prédire et donc représentent une source d’erreur à ajouter aux erreurs classiques de classification.

Nous avons déjà abordé les distorsions pouvant impacter les séries temporelles comme le climat, et aucune n’est corrigée par cette approche. En effet, pour la classification naïve les seules corrections réalisées sont les corrections atmosphériques comme indiqué précédemment 2.1, et l’interpolation temporelle. Intuitivement, on peut deviner que la disponibilité des images et leur couverture nuageuse vont impacter directement la qualité de l’interpolation. Le degré d’interpolation peut lisser fortement le signal impactant la dynamique du cycle annuel d’une classe, comme la levée pour les classes de culture, qui va mener à une dégradation de la qualité de séparation de cette classe par rapport à celles moins impactées. Le ré-échantillonnage temporel ne dégrade pas le signal acquis dans la mesure où il intervient pour palier à l’absence de données.

La qualité de la donnée d’apprentissage du domaine Source est un autre facteur potentiel qui impacte les performances. Une donnée de référence éparse, c’est-à-dire qui couvre une grande étendue de la surface à cartographier, avec une grande quantité d’échantillons permet l’apprentissage d’un classifieur avec une bonne capacité de généralisation. Pour un classifieur comme le RF, une grande diversité dans les échantillons d’une même classe, amène à trouver plusieurs feuilles correspondant au même identifiant de classes même si les échantillons auront un parcours dans l’arbre totalement différent.

Un autre facteur susceptible d’impacter les performances de la classification naïve est la similarité entre deux années. En effet, deux années similaires sur le plan climatique (température, ensoleillement, précipitation) atténuent une grande partie des distorsions entre elles.

Cette approche étant la plus simple envisagée, elle constitue une méthode de référence pour toutes les études menées au cours de cette thèse. Elle représente une application à très faible coût, similaire à la classification supervisée.

3.2 Résultats

3.2.1 Performances globales

La table 3.1 présente la PG moyenne et les intervalles de confiances à 95 % obtenus pour toutes les combinaisons possibles à partir des sept ans de données Formosat. L’année indiquée sur les colonnes représente le domaine Source, c’est-à-dire l’année d’apprentissage du modèle. Sur chaque ligne l’année indique alors le domaine Cible, la série temporelle que l’on cherche à classer. Les valeurs sur la diagonale (en bleu) représentent

TABLE 3.1 : PG et intervalle de confiances à 95 % obtenus pour la classification supervisée (diagonale en bleu) et la classification naïve. Les valeurs magenta et verte représentent les extrêmes obtenus.

Cible	Source							δ_{ST}
	2007	2008	2009	2010	2011	2012	2013	
2007	0.898 ± 0.024	0.719 ± 0.031	0.599 ± 0.025	0.751 ± 0.019	0.784 ± 0.016	0.752 ± 0.016	0.622 ± 0.028	0.185
2008	0.531 ± 0.026	0.859 ± 0.040	0.671 ± 0.017	0.690 ± 0.021	0.668 ± 0.020	0.588 ± 0.032	0.658 ± 0.028	0.159
2009	0.639 ± 0.021	0.731 ± 0.025	0.854 ± 0.026	0.760 ± 0.013	0.735 ± 0.020	0.671 ± 0.031	0.703 ± 0.018	0.121
2010	0.706 ± 0.024	0.740 ± 0.037	0.705 ± 0.025	0.892 ± 0.013	0.817 ± 0.018	0.789 ± 0.015	0.675 ± 0.031	0.142
2011	0.658 ± 0.013	0.682 ± 0.033	0.599 ± 0.028	0.768 ± 0.008	0.894 ± 0.007	0.716 ± 0.019	0.561 ± 0.041	0.207
2012	0.699 ± 0.014	0.737 ± 0.018	0.684 ± 0.033	0.813 ± 0.012	0.820 ± 0.014	0.884 ± 0.011	0.685 ± 0.022	0.136
2013	0.671 ± 0.022	0.674 ± 0.029	0.691 ± 0.031	0.724 ± 0.024	0.709 ± 0.023	0.694 ± 0.022	0.892 ± 0.012	0.053
Δ_{RF}	0.175	0.093	0.106	0.123	0.226	0.201	0.142	

la PG de la classification supervisée. Le reste de la table représente la classification naïve. Les extrêmes sont représentés en magenta pour le minimum avec 0.53 obtenus avec la classification de la série temporelle de 2008 et le classifieur entraîné sur 2007, et en vert pour le maximum qui atteint 0.81 pour la classification de 2010 avec le classifieur de 2011. Ces extrêmes définissent une plage de variation d'environ 30 % au sein de laquelle la classification naïve oscille en fonction des paires de domaine considérés avec une PG moyenne de 0.69. Ces résultats sont très positifs car cela traduit la forte capacité du RF à généraliser ces règles de décisions et à s'adapter aux distorsions entre les années. Dans le pire des cas rencontrés, plus de la moitié des décisions prises sont correctes, ce qui conforte cette conclusion. Un second résultat intéressant est la faible ampleur des intervalles de confiance, montrant la stabilité du RF face aux distorsions.

La table 3.1 est séparable en deux selon la diagonale, la partie inférieure représentant l'utilisation de données passées pour produire la carte d'OCS d'une année future par rapport au domaine Source. La partie supérieure de la table représente le procédé inverse, c'est-à-dire pour une carte d'une année antérieure à l'historique. Cette représentation s'avère d'autant plus intéressante que les valeurs obtenues ne sont pas symétriques vis-à-vis de la diagonale. Ainsi classer la série temporelle de 2008 avec le classifieur de 2007 et inversement laisse un écart de 18 % en PG entre les deux cartes produites. Nous pouvons en déduire que les classifieurs entraînés sur des signaux peu interpolés sont en mesure de s'adapter plus aisément à une absence de données. Inversement, une absence de données dans les profils temporels lors de l'apprentissage entraîne une baisse de la capacité de généralisation du classifieur.

En comparant les valeurs entre lignes et colonnes de la table 3.1, on extrait deux comportements. Le premier, visible sur les colonnes montre une bonne stabilité moyenne des performances obtenues par un classifieur.

Le second comportement est visible au travers de la variation des performances, plus marquée sur les lignes. La comparaison entre ces résultats et l'analyse de la qualité des données (2.1) met en évidence le lien entre couverture nuageuse, régularisation temporelle et stabilité lors de la classification de la série temporelle correspondante. En effet, 2007 et 2013 sont les années pour lesquelles le plus d'images sont disponibles (respectivement 24 et 20). Cependant, les objectifs liés à l'achat des images ayant évolué, les images de 2007 (et surtout 2006) sont fortement nuageuses. En effet, celles-ci sont consacrées aux tests de la chaîne MACCS (détection des nuages) alors que les images de 2013 sont dédiées à la production de cartes d'OCS.

Les années 2010, 2011 et 2012 sont très proches en termes de nombre d'images disponibles et de couverture nuageuse. Les séries temporelles correspondantes subissent la même régularisation temporelle, et offre des

performances similaires lorsque elles sont classées entre elles. Cette similarité est facilement visible via la figure B.1, un carré coloré apparaît clairement à l’intersection de ces années. Cette figure est une matrice, les axes indiquant les années correspondant aux domaines Source (abscisses) et Cible (ordonnées), l’échelle de couleur représente la valeur de PG. C’est une représentation graphique de la matrice présentée dans la table 3.1. Dans la suite de ce mémoire, cette représentation graphique sera privilégiée pour présenter de manière synthétique les performances.

La marge de progression est significative entre nos deux méthodes de référence (supervisée et naïve) avec un écart aux alentours de 20 %. Cet écart se montre irrégulier en fonction du domaine Source, les différentes méthodes proposées doivent chercher à corriger les impacts évoqués ci-dessus afin de permettre une amélioration de la qualité de l’image produite.

3.2.2 Performances par classe

Dans la mesure où les classes ne sont pas équilibrées dans les données de référence, les métriques globales ne sont suffisantes pour juger la qualité de la classification produite. Nous réalisons l’analyse par classe via le FScore. La figure 3.2 est une représentation du FScore obtenu pour chaque paire d’années possible et pour chaque classe. Cette figure se décompose en 17 matrices. Sur l’axe des abscisses se situe l’année correspondant au domaine Source et en ordonnées le domaine Cible. L’ordre des classes est celui utilisé dans la nomenclature 2.1, ordonné par l’identifiant de la classe. Dans ces matrices, la diagonale descendante de gauche à droite représente le cas supervisé pour chaque année. La partie inférieure de la matrice représente la classification naïve chronologique et la partie supérieure l’ordre anti chronologique. L’échelle de couleur représente les valeurs de FScore qui évoluent entre 0 (en rouge) et 1 (en bleu).

De manière générale, trois tendances sont extraites de cette analyse par classe :

- Une première catégorie de classes qui sont majoritaires dans la données de référence et dans la zone d’étude, pour ces classes, la classification naïve apporte déjà un bon score de classification avec un FScore supérieur à 75%. C’est le cas des classes Feuillus, Conifères, Blé, Maïs, Tournesol, et Gravières.
- Une seconde catégorie constituée par des classes assez bien représentées mais complexes à classer convenablement. Le FScore obtenu varie entre 50 % et 75 %. C’est le cas des classes de Colza, Surfaces Artificielles, Friches, Rivières, Lacs et Surfaces En Herbe (SEH).
- Enfin la troisième catégorie est constituée de classes très difficiles à classer, en grande partie en raison de leur mauvaise représentation dans la zone d’étude. Ces classes sont l’Orge, le Soja, le Sorgho, les Jachères et le Chanvre.

La figure 3.2 illustre un contraste édifiant avec une majorité de valeurs de FScore proche des bornes extrêmes et relativement peu de valeurs intermédiaires. Notons l’exceptionnelle prédiction des classes de Feuillus, Maïs et Gravières à la fois pour la classification supervisée mais également pour la naïve. Ces performances peuvent être justifiées par les particularités de ces classes. En effet, ces classes sont géographiquement très stable, dans la région couverte par la zone d’étude. Les Feuillus sont très bien représentés dans cette région riches en étendues forestières. Les Gravières sont une classe un peu particulière étant un mélange d’eau et de gravier, celles-ci peuvent être déplacées d’une année sur l’autre mais ces caractéristiques en termes de signal acquis en font une classe “facile” à classer. Le Maïs est une culture ayant de nombreuses contraintes matérielles et logistiques liées à sa production. En effet cette culture requiert une irrigation abondante pour

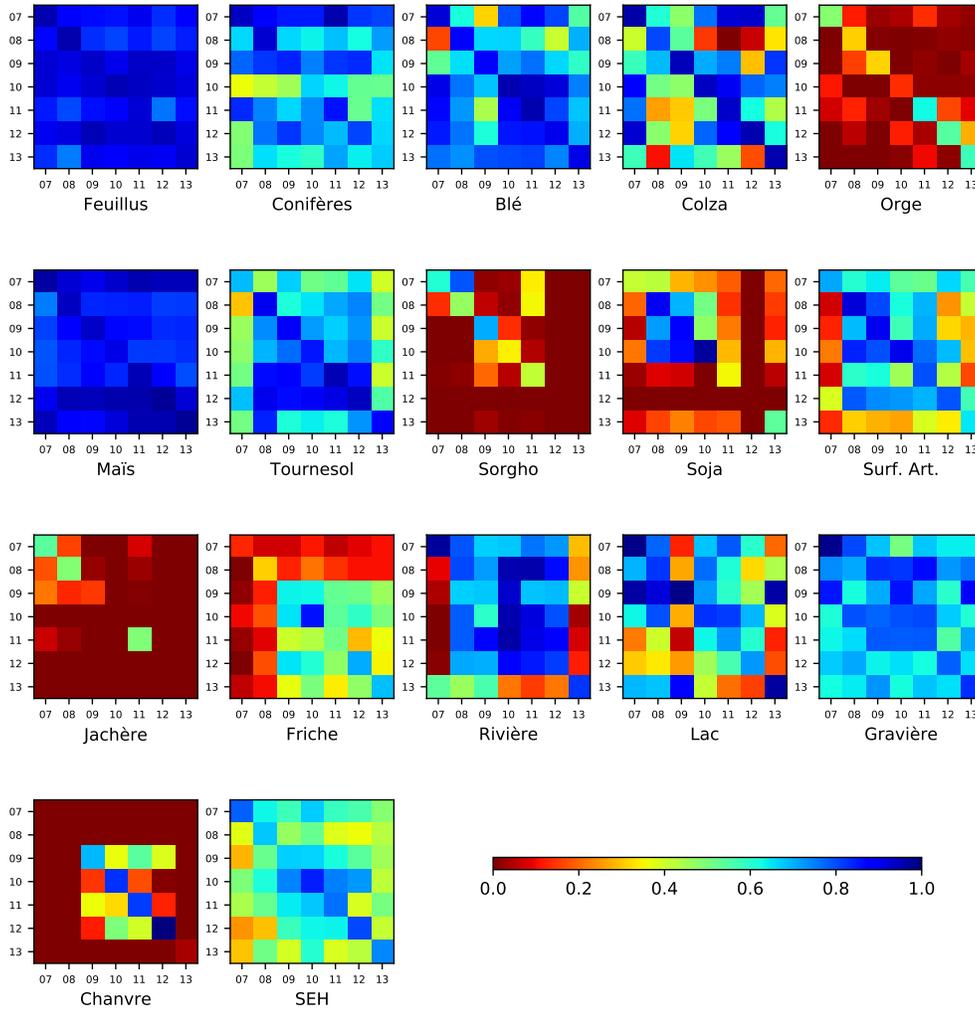


FIGURE 3.2 : Matrices de FScore obtenues par classification naïve. L'échelle de couleur représente le FScore par classe, l'axe des abscisses représente le domaine Source et l'axe des ordonnées le domaine Cible. Chaque classe est représentée par une matrice de gauche à droite dans l'ordre de l'identifiant de classe dans la nomenclature.

assurer la rentabilité. De manière générale, une fois le système d'irrigation implanté sur une parcelle, celle-ci devient dédiée à la culture du Maïs pour de nombreuses années. C'est une culture fortement représentée dans la zone d'étude, sa forme phénologique ainsi que son mode de plantation provoquent rapidement l'apparition d'une surface homogène sur l'intégralité de la parcelle. Le fait que ce soit une culture irriguée atténue également les distorsions liées aux conditions climatiques telle que la sécheresse.

Les autres cultures majoritaires dans la zone d'étude sont le Blé et le Tournesol et sont également bien classées de manière générale. Ces classes sont très représentées dans les données de référence. Le parallèle entre le FScore et le nombre d'échantillons par classe est aisément réalisable en considérant la table 2.1. La comparaison du FScore du Tournesol pour 2007 et 2009 en domaine Source appuie ce constat, avec un facteur

3 entre le nombre d'échantillons de Tournesol de 2007 et de 2009.

La seconde catégorie regroupent des classes ayant des caractéristiques particulières mais qui sont très proches d'autres classes. L'exemple typique sont les classes de Rivières et Lacs. Ces classes ont des particularités visuellement évidentes, puisque le lac est un plan d'eau avec une forme plutôt compacte, alors qu'une rivière est plus linéaire. La différenciation entre Rivières et Lacs opère sur les variations des Rivières au cours de l'année, avec les crues et décrues. Cependant, les pixels représentant le milieu du cours d'eau sont régulièrement confondu avec les Lacs et inversement pour les berges des Lacs. Ce qui rend particulièrement complexe une bonne séparation de ces classes.

Dans le même contexte, les classes de SEH et Friches représentent des étendues d'herbes, une friche incluant des arbustes et autres fourrés difficilement identifiables à l'échelle du pixel. La classe de Surfaces Artificielles est ici un mélange entre différentes classes de surfaces urbaines : Bâti dense, Bâti diffus et Zone industrielle (cf section 2.1.1). D'ordinaire ces classes sont considérées indépendamment. Ici, devant le manque de données pour ces classes et l'intérêt se portant plus sur les classes annuelles et surtout agricole, ces classes ont été regroupées sous une étiquette unique. Cependant, la grande diversité au sein de cette classe explique les performances variables obtenues pour cette classe.

Nous pouvons expliquer les performances de la troisième catégorie par la nature de ces classes, mais également par leur manque de représentation dans les données de référence. L'Orge et le Sorgho sont connues dans cette zone d'étude comme étant des classes complexes à identifier. Ces classes ont un développement phénologique similaire à celui du Blé et du Maïs respectivement, rendant leur identification difficile lors des campagnes terrains. Le Chanvre a été implanté dans la zone d'étude en 2009 et ne représente que quelques parcelles en bordure de la zone d'étude, ce qui explique sa mauvaise représentation. Les classes de Jachères et Friches sont très difficiles à définir (y compris au niveau thématique). Ce sont deux classes relativement proches ce qui explique les confusions que l'on peut observer.

Ces trois catégories de FScore sont visibles dans la figure 3.2, en comparant les valeurs de la diagonales au reste de la matrice. Une matrice homogène bleue indique que la classe appartient à la première catégorie alors qu'une matrice homogène rouge indique la troisième catégorie. Les matrices très colorés indiquent alors essentiellement les classes de la seconde catégorie.

La notion de confusion entre classes joue un rôle déterminant dans l'analyse des résultats. La figure 3.3 illustre une matrice de confusion obtenue par la classification naïve de 2011 (domaine cible) par le classifieur de 2009. Cette classification se situe dans la limite basse en termes de performances avec une PG affichée à 57 %. Le FScore est indiqué sur le côté droit de la matrice, le rappel et la précision sont également indiqués pour information. Cette classification n'est pas la meilleure obtenue, elle est inférieure à la moyenne de tous les tirages (59 % cf table 3.1), cependant elle illustre bien les tendances de confusions possibles réalisées par le classifieur. Nous pouvons voir grâce à cette figure qu'il y a très peu de confusion aberrantes, les classes d'eau par exemple sont essentiellement confondues entre elles. À l'inverse, de nombreuses classes agricoles, notamment les céréales à paille (Orge, Blé) peuvent être confondues entre elles, mais aussi avec la classe SEH. Les confusions entre cultures d'été et d'hiver sont également restreintes. Dans le cas particulier de 2008 qui ne possède aucune image pour la période hivernale, les cultures d'hiver sont classées comme Jachères (cf. B.2). Ces confusions logiques¹ mettent en évidence la bonne capacité de généralisation du RF, et la pertinence de la classification naïve. À titre de comparaison la figure 3.4 présente la matrice de confusion du cas supervisé 2011 pour les mêmes échantillons. Nous pouvons noter que ces confusions ce retrouvent

¹dans le sens où un intervenant humain peut réaliser les mêmes erreurs

dans une moindre mesure dans les performances de la classification supervisée. Le cas naïf conserve donc les propriétés fondamentales du **RF** mettant en évidence la cohérence des résultats.

KAPPA : 0.514 PG : 0.575

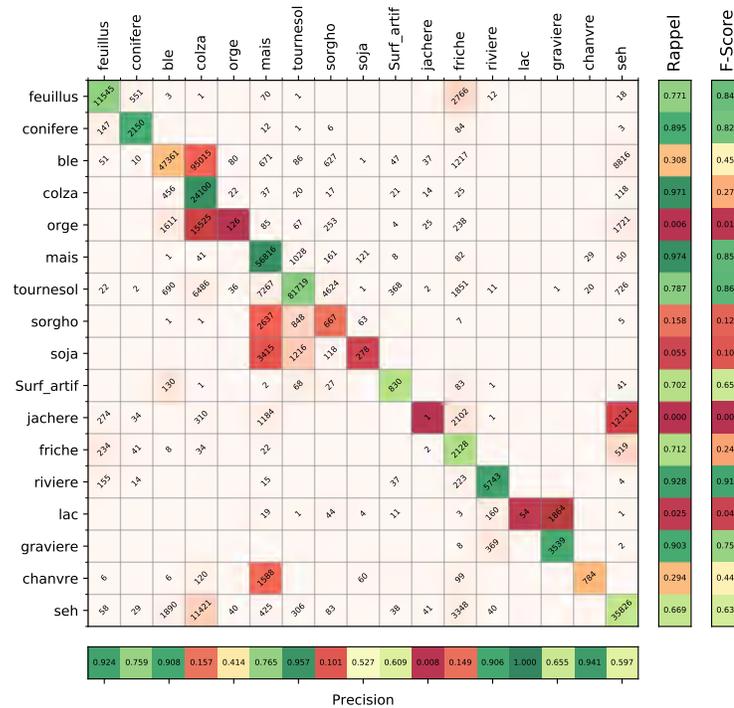


FIGURE 3.3 : Exemple de matrice de confusion pour la classification Naïve du domaine Cible 2011 par le domaine Source 2009. Les cellules en rouges indiquent un mélange important entre les classes. Le Fscore et Rappel sont illustrés sur l’axe des ordonnées et la précision sur l’axe des abscisses. Les classes sont indiqués sur chaque axe. La diagonale représente le nombre d’échantillons correctement classés.

Au travers de la figure 3.2, nous pouvons voir que le FScore est toujours supérieur lors de la classification supervisée pour toutes les classes, et il demeure inégalé par la classification naïve. Cette analyse par classe met bien en avant les tendances extraites depuis les performances globales mais permet d’identifier où doivent se concentrer les efforts d’améliorations, à savoir sur les classes annuelles.

Ces résultats montrent également l’intérêt de conserver une nomenclature complexe qui va permettre d’analyser les différents comportements des classes en fonction des méthodes. L’analyse par classe permet également de visualiser où se situe le gain (ou la perte) réalisé par une méthode de manière plus pertinente que les métriques globales. En effet, une amélioration d’une classe majoritairement représentée dans la donnée de référence va avoir un effet positif sur la PG en augmentant fortement sa valeur et inversement.

Les confusions observées dans la matrice de confusion de la figure 3.3 se retrouvent dans la carte produite. La figure 3.5 illustre la classification supervisée de 2011 et la classification naïve de 2011 à partir du classifieur 2009. La forte confusion entre Blé et Colza est nettement visible avec la forte coloration rose indiquant la classe Colza en lieu et place du Blé (jaune pale).

Une autre piste, délaissée ici, serait de réaliser la classification au degré le plus précis de la nomenclature

KAPPA : 0.861 PG : 0.889

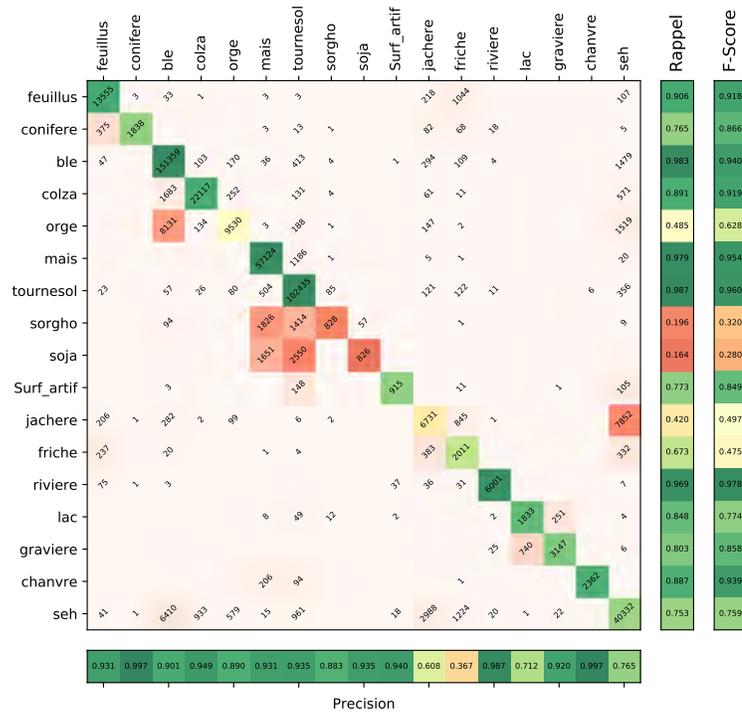


FIGURE 3.4 : Exemple de matrice de confusion pour la classification supervisée de 2011. Les cellules en rouges indiquent un mélange important entre les classes. Le Fscore et Rappel sont illustrés sur l'axe des ordonnées et la précision sur l'axe des abscisses. Les classes sont indiqués sur chaque axe. La diagonale représente le nombre d'échantillons correctement classé.

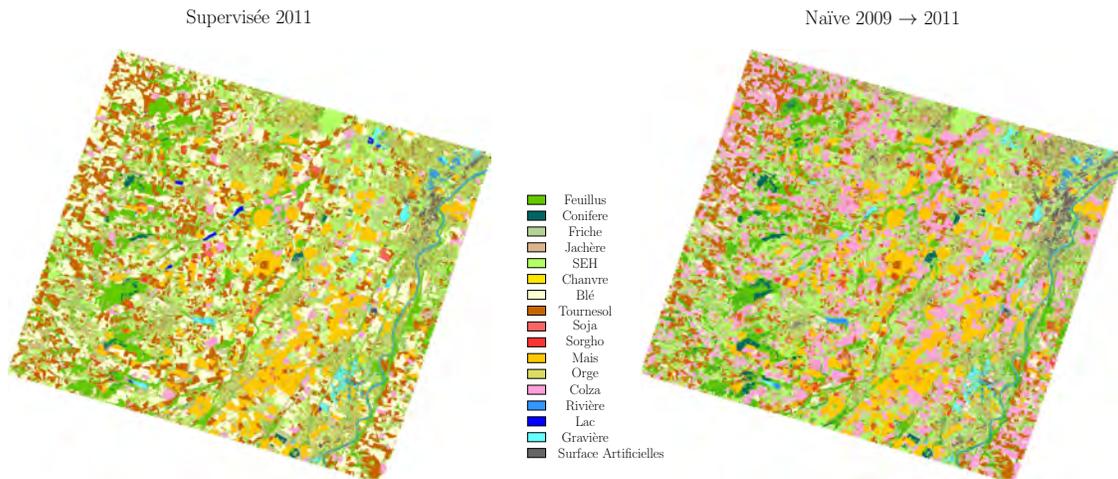


FIGURE 3.5 : Classifications supervisée et naïve pour le millésime 2011.

hiérarchique, puis de fusionner les classes sur un ou plusieurs niveaux (cf 1.4). Par exemple, les classes de Rivières, Lacs et Gravières sont regroupées sous l’étiquette “Eaux”. La nature cohérente des confusions réalisées par le classifieur assure alors une qualité de la carte tout à fait acceptable. La table 3.2 représente la matrice de la classification naïve présentée dans la figure 3.3, après fusion des classes selon la nomenclature 1.4. En termes de métriques, cette fusion permet un gain, par rapport à l’exemple précédent, de 37 % de PG et de très bons FScore ($> 75\%$), mais en contrepartie nous perdons la finesse de la carte. La table 3.3 présente les résultats de ce même regroupement pour le cas supervisé (3.4). Le FScore obtenu est bien meilleur, avec une valeur minimale de 0.849 contre 0.65 dans le cas naïf. Nous pouvons constater que la classe de Bâti bénéficie le moins de ce regroupement, en effet si sa valeur de précision grimpe en flèche (de 0.61 à 0.94), sa valeur de rappel reste plus stable (de 0.7 à 0.77). Ce qui se traduit par une réduction de la confusion des échantillons des autres classes avec le bâti mais que des échantillons de bâti sont très complexes à traiter même pour la classification supervisée. Sa forte confusion avec des classes de végétation conforte l’hypothèse que la classe de bâti diffus initialement fusionnée avec les autres classes de surfaces artificielles constitue une source potentielle d’erreur.

Cette amélioration conforte la notion de confusion logique abordée précédemment. Cet exemple permet d’illustrer le lien fort existant entre besoins de l’utilisateur (c’est-à-dire la nomenclature cible), la nomenclature la plus détaillée disponible et les performances obtenues. Pour de nombreuses applications une telle nomenclature s’avère insuffisante car elle est trop générale. Dans la suite de ce manuscrit nous conservons la nomenclature à 17 classes.

TABLE 3.2 : Matrice de confusion obtenue lors de la fusion post classification naïve de 2011 (cible) par 2009 (source). La fusion est réalisée en réduisant la finesse de la nomenclature.

Classe	Eaux	Bois	Prairie	Culture d’hiver	Culture d’été	Bâti	Rappel	FScore
Eau	11698	160	291	1	78	42	0.953	0.974
Bois	2	14443	2858	6	61	0	0.831	0.881
Prairie	35	731	57830	11736	2180	48	0.797	0.767
Culture d’hiver	0	58	14208	181996	2187	55	0.917	0.909
Culture d’été	11	26	2831	7962	162878	374	0.936	0.954
Bâti	1	0	126	159	63	834	0.705	0.658
Précision	0.996	0.937	0.740	0.902	0.973	0.616		PG : 0.90

TABLE 3.3 : Matrice de confusion obtenue lors de la fusion post classification supervisée de 2011. La fusion est réalisée en réduisant la finesse de la nomenclature.

Classe	Eaux	Bois	Prairie	Culture d’hiver	Culture d’été	Bâti	Rappel	FScore
Eau	11999	76	84	3	69	39	0.978	0.986
Bois	18	15771	1524	34	23	0	0.908	0.934
Prairie	44	486	62698	8325	989	18	0.864	0.884
Culture d’hiver	4	47	4193	193479	780	1	0.975	0.966
Culture d’été	11	23	636	257	173155	0	0.995	0.992
Bâti	1	0	116	3	148	915	0.773	0.849
Précision	0.994	0.961	0.905	0.957	0.989	0.940		PG : 0.96

3.3 Séparation entre classes annuelles et pérennes

Nous avons évoqué lors de la mise en contexte des présents travaux, que la chaîne de traitement [Infrastructure pour l'Occupation des sols par Traitement Automatique Incorporant les Orfeo Toolbox Applications \(iota²\)](#) a été développée en parallèle des travaux menés au cours de ceux-ci. Se fondant sur l'idée de la classification naïve, une procédure de [Séparation des classes Annuelles et Pérennes \(SAP\)](#) a été mise en place. Le principe est très simple. En effet, une classe pérenne est, par définition, stationnaire géographiquement dans le temps : les cours d'eau et surfaces artificielles, et à échelle de temps plus variables les forêts et certaines cultures, comme le Maïs. Pour ces classes, la donnée de référence d'une année précédente s'avère donc valide. Les positions de ces échantillons sont donc extraites depuis le domaine Source et les échantillons sont extraits dans le domaine Cible directement, comme illustré dans la figure 3.6.

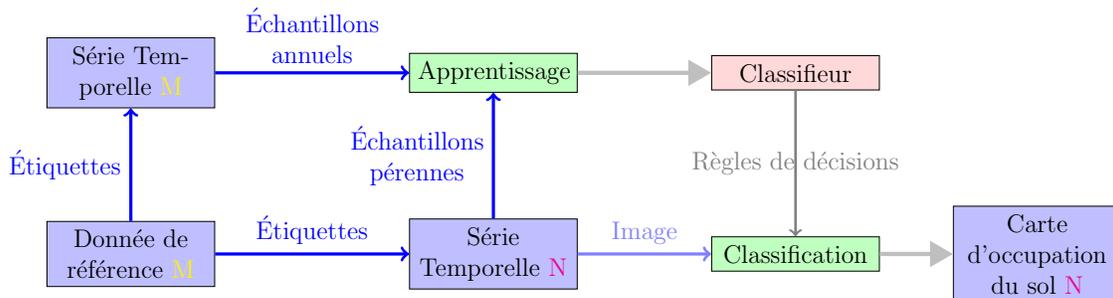


FIGURE 3.6 : Chaîne de traitement utilisée dans le cadre de la classification SAP.

Pour les classes annuelles, cette hypothèse n'est pas valable, les échantillons sont donc extraits dans le domaine Source exclusivement (comme pour la classification Naïve). Par rapport à la nomenclature définie dans la table 2.1, seules les classes de Feuillus, Conifères, Surfaces Artificielles, Rivières et Lacs sont considérées comme pérennes. Toutes les autres sont donc des classes annuelles.

Cette approche a été validée par le [Centre d'Expertise Scientifique Occupation des Sols \(CES OSO\)](#) avec des produits composites Spot5-Landsat8 et plus tard Sentinel-2. Elle a été testée dans le cadre des travaux de thèse pour comparaison avec la classification Naïve. Le résultat obtenu reste aléatoire et moins tranché comme l'illustre la figure 3.7. Cette figure présente trois matrices, l'échelle de couleurs représente les valeurs de PG. L'axe des abscisses et des ordonnées représentent l'année du domaine Source et Cible respectivement. La matrice de gauche représente la valeur de PG obtenue par la classification SAP, l'échelle de couleur représente la valeur obtenue le rouge indiquant une PG faible, les teintes d'orange indiquant une PG proche de 50 % et bleues une bonne PG. La matrice du milieu représente la différence de PG entre l'approche SAP et naïve ($PG_{SAP} - PG_{naïf}$). Une valeur positive, représentée par une couleur bleue met en évidence une amélioration par rapport au cas naïf. Une valeur négative (en rouge) représente une dégradation de la PG. Plus la teinte s'éclaircit moins la différence avec le cas naïf est importante. Enfin la matrice de droite représente les intervalles de confiance à 95 % obtenus par la classification SAP, une valeur en bleu indique un intervalle de confiance très petit, et une valeur en rouge un intervalle très grand ($> 15\%$).

Grâce à la figure 3.7, nous pouvons constater de la forte similarité de l'approche SAP avec la classification naïve, d'autant plus que les intervalles de confiance sont très petits, annulant l'écart mesuré dans la matrice. Ce qui ne coïncide pas avec le résultat attendu dans la mesure où le classifieur accède à des échantillons du

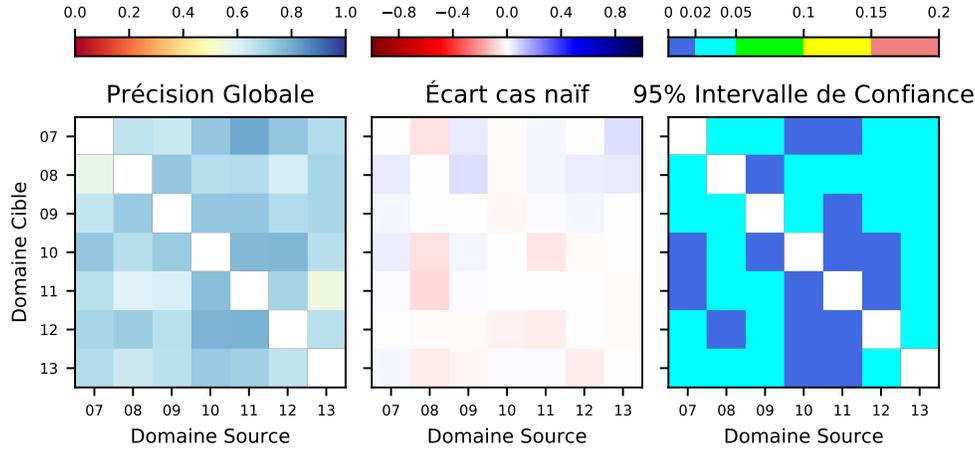


FIGURE 3.7 : Matrices représentant la PG obtenue par la classification SAP (gauche), son écart vis-à-vis du cas naïf (milieu) et la borne maximale des intervalles de confiance à 95 % obtenus (droite). L'axe des abscisses indique l'année du domaine Source, et celui des ordonnées le domaine Cible.

domaine Cible directement.

L'analyse par classe s'avère plus révélatrice de l'impact de la procédure SAP. La figure 3.8 s'appuie sur la même représentation que celle de la figure 3.2, cependant les valeurs affichées correspondent à l'écart de FScore entre la classification SAP et naïve. Comme pour la figure précédente, l'échelle de couleur commence par le rouge indiquant une dégradation du Fscore par rapport au cas naïf et termine par le bleu synonyme d'amélioration. L'intensité de la teinte témoignant de l'importance de l'écart. Sur les cinq classes pérennes, trois d'entre elles bénéficient de cette procédure avec un gain de plus de 60 % pour la classe Rivières avec 2007 en domaine source. Les Conifères bénéficient aussi dans une moindre mesure de la SAP. L'impact sur la classe des Feuillus déjà très bien classée par la classification naïve est négligeable. La classe de Surfaces Artificielles est celle qui bénéficie le moins de cette approche, probablement pour les mêmes raisons que précédemment. Sur les classes annuelles, les effets sont totalement négligeables.

En résumé, en termes de PG et même de Fscore, on note aucune amélioration significative par rapport à la classification naïve, avec une tendance à la dégradation. Cependant, en analysant les quantités d'échantillons dans les matrices de confusion, nous constatons une amélioration des bonnes décisions pour les classes pérennes ; mais une dégradation pour les 12 autres classes. On obtient par conséquent un résultat contradictoire avec le gain observé par le [CES OSO](#).

L'origine du problème vient de l'opposition des 108 bandes spectrales et dates de Formosat-2 aux quelque 730 de Sentinel-2. Ce qui indépendamment du facteur d'interpolation, permet au signal Sentinel-2 d'être beaucoup plus fourni et similaire d'une année sur l'autre. En contre exemple, dans le jeu de données Formosat-2 une absence d'image pendant une longue période est observée notamment pour 2008 (aucune image d'hiver) et 2010 (pas d'image en août). La différence de comportement de la SAP est clairement due à la qualité du signal. Reprenons l'exemple de 2008, illustré dans la figure 2.3. Durant les trois mois sans images, le signal est donc interpolé, linéairement, ce qui génère un artefact dans le signal. Avec une image absente, tous les pixels à cette date sont concernés. Ce qui n'est pas le cas lors de la SAP, où vont se mêler des échantillons

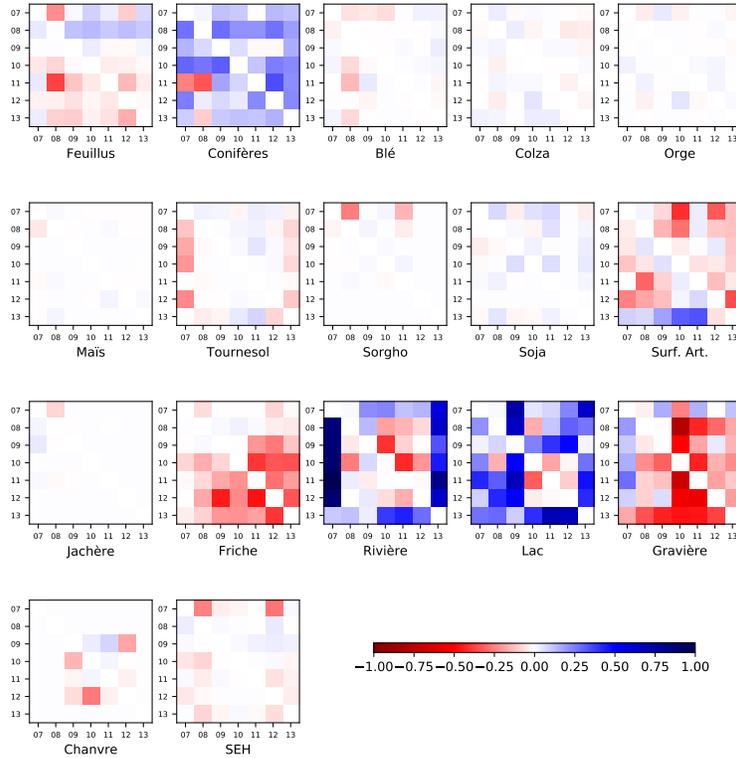


FIGURE 3.8 : Matrices représentant la différence de FScore entre la classification SAP et naïve. L'échelle de couleur est centrée en 0. Le rouge indique une dégradation et le bleu une amélioration. Plus les couleurs sont claires moins la différence est significative.

avec un signal cohérent et ceux fortement interpolés.

3.4 Synthèse

La classification naïve et SAP permettent de mettre en lumière un problème concret. Les distorsions entre les années sont trop importantes pour qu'un classifieur, même aussi robuste qu'un **RF**, puisse passer outre de lui même.

L'apport du ré-échantillonnage temporel est indéniable, puisqu'il permet l'utilisation de ces approches de classification.

Ces deux approches nous ont permis de mettre l'accent sur plusieurs points :

1. Les distorsions n'ont pas le même impact en fonction des classes lors de la classification.
2. Le ré-échantillonnage temporel peut avoir un effet négatif lorsque le signal est fortement dégradé par la présence de nuages dans la procédure PAS.
3. Le **RF** reste très performant malgré les distorsions et les mélanges d'échantillons entre les années.
4. L'écart entre le cas naïf et supervisé reste tout de même conséquent.

Dans le reste de ce manuscrit nous allons mettre à l'épreuve diverses méthodes existantes dans la littérature issues notamment de l'adaptation de domaine, et des approches de fusion. L'accent sera mis, dans la mesure du possible, sur la comparaison des performances de chaque méthodes vis à vis de la classification naïve. Ces comparaisons porteront à la fois sur les métriques standards de validation de classification mais également sur l'effort requis à l'exploitation d'une méthode et son temps de calcul.

Troisième partie

**Propositions de méthodes exploitant
un unique domaine Source**

Précédemment, nous avons mis en lumière l'apport de la télédétection et des nouvelles techniques automatiques pour la production de cartes d'OCS. Nous avons également mis en évidence un certain nombre de contraintes liées à leur utilisation, notamment la dépendance aux données de référence, obligatoires pour réaliser une classification supervisée. Cette dépendance induit des délais importants dans le cadre d'une production opérationnelle, mais interdit également toute tentative de classification d'une période antérieure, pour laquelle les données de référence n'auraient pas été acquises. De plus, des distorsions complexifient une exploitation des données de référence d'une période différente à celle de son acquisition. Ce condensé du sujet de thèse correspond à une des définitions d'un des plus grands volets de l'AA : le [Transfert d'Apprentissage \(*Transfer Learning*\) \(TA\)](#) et notamment l'Adaptation de Domaine. Ces techniques, très utilisées en vision par ordinateur [Patel et al., 2015] proposent des solutions permettant la classification d'un Domaine Cible D_C pauvre en données de référence, à partir d'informations issues du Domaine Source D_S , pour lequel les données de référence sont disponibles.

4.1 Transfert d'apprentissage

Le [Transfert d'Apprentissage \(*Transfer Learning*\) \(TA\)](#) trouve son origine dans un processus simple, que nous pourrions qualifier de naturel. En effet, généralement, on préfère exploiter des connaissances acquises au préalable et les adapter à un problème similaire, plutôt que chercher des solutions relevant d'un nouvel apprentissage, ou de la recherche empirique.

Les domaines d'application du TA constituent un ensemble très varié. Il prédomine en traitement automatique du langage et la reconnaissance d'image par exemple. Ces domaines d'apprentissage, qui exploitent des algorithmes ayant une forte complexité calculatoire, constituent une situation favorable au TA. En effet, il permet d'optimiser ces traitements en exploitant au maximum des classifieurs déjà entraînés.

La figure 4.1 illustre les différents éléments constituant le TA [Pan and Yang, 2010]. Dans cette figure, nous distinguons différentes catégories d'algorithmes (cadres bleus) et les conditions requises à leur exploitation

(cadres verts). En TA, l'objectif consiste à produire une classification du D_C . Un premier critère d'exploitation concerne la disponibilité des étiquettes (et donc des données de référence), amenant aux trois catégories de TA :

1. Le TA inductif : on cherche à améliorer les performances dans le D_C grâce à la connaissance des étiquettes dans D_S
2. Le TA transductif : on cherche à classer les échantillons appartenant au D_C à partir des connaissances sur les étiquettes du D_S
3. Le TA non supervisé : on ne possède aucune étiquettes sur D_S et D_C

Le TA inductif considère que nous possédons les étiquettes d'échantillons sur le domaine Cible mais que nous souhaitons améliorer les performances et la généralisation du classifieur en introduisant les connaissances issues du domaine Source. Deux cas de figures se présentent alors en fonction de l'étiquetage du D_S .

En cas d'absence d'étiquettes pour le D_S , nous pouvons exploiter les algorithmes d'auto apprentissage [Raina et al., 2007]. En résumé, les auteurs exploitent les échantillons non étiquetés afin d'extraire une représentation plus abstraite (haut-niveau) des échantillons étiquetés amenant à l'apprentissage d'un classifieur supervisé plus performant. Cette approche exploite des domaines pouvant s'avérer totalement divergents (images d'animaux en cible et images de paysage en source). Déterminer des caractéristiques dans la structure des images afin de révéler des zones fortement corrélées, demeure l'attrait principal de cet algorithme.

À l'inverse lorsque les deux domaines possèdent des étiquettes, le TA réalise un apprentissage multi-tâches, consistant à procéder à l'apprentissage des deux domaines simultanément. L'apprentissage multi-tâche est très utilisés dans des problèmes d'optimisation de coût. Dans le contexte du TA, le but du classifieur se rapproche de la sélection de primitives pertinentes pour la classification du domaine Cible. Les contraintes sur les primitives sont exprimées sous la forme d'équations de minimisation, résolues par diverses approches [Rückert and Kramer, 2008, Jebara, 2004, Argyriou et al., 2007].

Le TA transductif semble le plus proche des conditions de notre cas d'étude. En effet, cette approche exploite seulement les étiquettes du domaine Source. Deux hypothèses sont envisagées :

- deux domaines différents avec une seule tâche de classification.
- un domaine unique associé à une tâche de classification ;

La première hypothèse semblerait la branche la plus en adéquation avec nos travaux : on considère deux domaines différents reliés (une même empreinte avec des distorsions), dans lequel on exploitera les informations issues du domaine Source pour classer le domaine Cible. Cette branche amène aux méthodes d'adaptation de domaine, qui sont présentés dans la section suivante.

Dans la seconde hypothèse, la classification concerne une partie non étiquetée du domaine Source, par exemple une zone différente d'une même image. Ce cas d'étude est fréquent, lors d'une acquisition d'une image en plusieurs passages, dans des conditions d'illumination ou d'angle de vue variés par exemple. Dans ces conditions, les approches exploitées pour traiter le biais de sélection apportent des éléments de réponses à la résolution du TA transductif. Le terme "biais de sélection" désigne en statistique, une erreur systématique commise lors de la sélection des échantillons. Le jeu d'échantillons ainsi sélectionné ne constitue pas un ensemble représentatif du problème étudié. La classification supervisée est toujours soumise à ce dilemme

(cf compromis biais-variance) pour lequel, généralement, on recourt à une équation de minimisation des risques de sélection. On évalue généralement le risque comme une probabilité à posteriori déterminée par une fonction de coût. Le TA cherche le transfert d'un classifieur issu du domaine Source vers le domaine Cible par l'introduction des échantillons de ce dernier dans l'équation de minimisation. De nombreuses approches existent afin de résoudre ce problème de minimisation, exploitant des mesures issues de plusieurs classifieurs [Fan et al., 2005], des méthodes à noyaux cherchant à associer les moyennes des domaines [Huang et al., 2007], ou plus généralement une des nombreuses techniques utilisées pour la classification supervisée [Quionero-Candela et al., 2009].

Enfin la dernière catégorie de TA, exploite de manière non supervisée les deux domaines, et élabore des parallèles entre eux afin de faciliter les algorithmes comme le clustering [Dai et al., 2008] ou la réduction de dimension [Wang et al., 2008]. On utilise fréquemment cette dernière pour réduire le volume de données, en sélectionnant des primitives pertinentes, ou en regroupant des primitives non corrélées sous forme d'indices. Cette étape facilite alors l'exploitation des méthodes souffrant de difficultés de gestion de données de très grandes dimensions. L'approche non supervisée autorise les différences entre les deux domaines, tant qu'ils demeurent reliés, dans notre cas d'étude nous pouvons traduire cette contrainte comme une stabilité des classes d'OCS et du climat (zone écoclimatique).

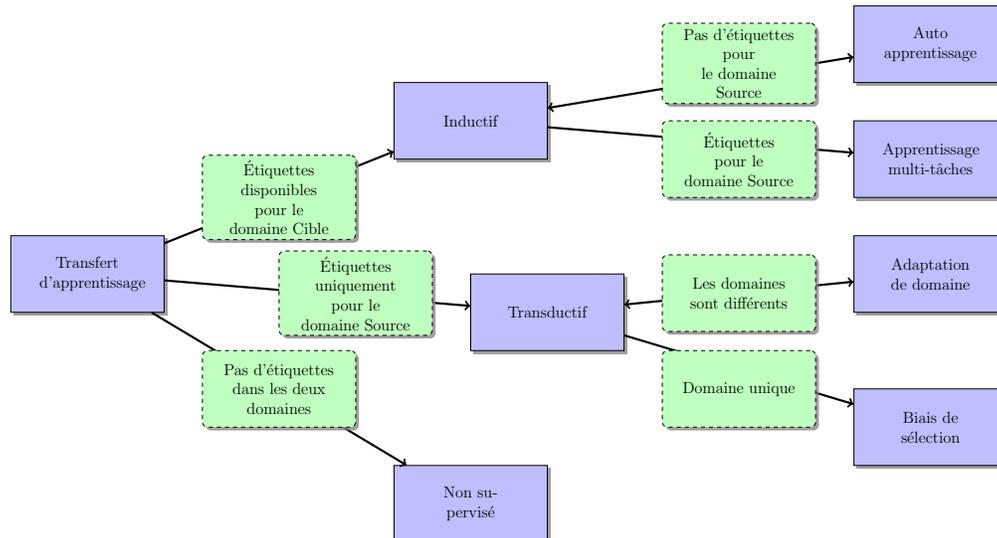


FIGURE 4.1 : Une vue d'ensemble des différentes configurations du transfert d'apprentissage.

4.2 L'adaptation de domaine

En 2016, Tuia *et al.* [Tuia et al., 2016] ont réalisé une revue de la bibliographie de l'adaptation de domaine appliquée à la télédétection. Dans cet article, ils catégorisent l'adaptation de domaine selon quatre grands axes :

1. Adaptation semi-supervisée de classifieur
2. Adaptation de classifieur par apprentissage actif

3. Extraction de primitives invariantes

4. Adaptation des données

Il est possible, pour adapter les classifieurs et/ou les données, de combiner les méthodes issues de différentes catégories. De manière générale, les méthodes d'adaptation de classifieur sont très appréciées lorsque le coût d'apprentissage sur le domaine Source est important ; il en est de même pour les distorsions entre les domaines. À l'inverse, on utilise les méthodes d'adaptation de données lorsque les distorsions sont moins importantes et les coûts d'apprentissage faibles. Nous avons décidé, pour cette revue de la littérature, de suivre la catégorisation initiale réalisée par les auteurs en y intégrant des travaux postérieurs.

4.2.1 Adaptation semi-supervisée de classifieur

En classification, une approche semi-supervisée est une méthode qui consiste à mélanger des échantillons étiquetés et non étiquetés, lors de l'apprentissage. Il s'agit exactement du même principe en adaptation de domaine. Le principe vise à entraîner un classifieur sur le domaine Source, en utilisant des informations issues du domaine Cible pour pondérer les règles de décision. En général, les étiquettes du domaine Cible ne sont jamais disponibles. On réalise l'adaptation par estimation de ressemblance, corrélation ou autres mesures statistiques entre les deux domaines. Pour ce type d'approches, on ne cherche plus à étiqueter le domaine Cible mais à identifier les échantillons les plus pertinents dans les deux domaines.

Les travaux présentés par [Bruzzone and Prieto, 2001] sont précurseurs des approches semi-supervisées pour l'adaptation de domaine. Cette méthode consiste à exploiter un classifieur par Maximum de Vraisemblance paramétrique entraîné sur le domaine Source. L'innovation consiste à calculer les distributions des classes dans une nouvelle image et de les utiliser pour mettre à jour ces paramètres. Cette approche a rapidement montré ces limites, d'une part en étant moins performante que la version supervisée d'un classifieur par Maximum de Vraisemblance, et d'autre part en n'étant pas exploitable pour des images provenant de capteurs ou de scènes différents. Ces travaux ont néanmoins mis en évidence que la complexité du problème lié à l'OCS offre des gains moins importants que ceux observés dans d'autres disciplines, comme la vision par ordinateur, par exemple. Rapidement, dans [Bruzzone and Cossu, 2002] les auteurs ont proposé une amélioration de cette approche, en exploitant une cascade de classifieurs. Une cascade de classifieurs consiste à l'entraînement de plusieurs classifieurs. L'enchaînement permet d'intégrer à chaque classifieur les sorties du classifieur précédent. La différence dans le cadre de l'adaptation de classifieur, réside dans l'ajout d'échantillons du domaine Cible à chaque nouvelle étape. Les résultats de chaque classifieur sont ensuite combinés pour mettre à jour la carte d'occupation du sol. Ces approches sont performantes. Néanmoins elles traitent essentiellement des zones d'études restreintes (étendue géographique), mais également en termes de classes d'OCS.

Ces deux approches ont ouvert la voie aux méthodes semi-supervisées. De nombreuses déclinaisons de ces approches existent en modifiant les classifieurs utilisés, exploitant les *Séparateurs à Vaste Marge (SVM)* [Bruzzone et al., 2006, Sun et al., 2013], les arbres de décision [Rajan et al., 2006], le principe étant toujours d'introduire des échantillons des deux domaines lors de l'apprentissage.

Des méthodes proposent d'estimer les distorsions entre les domaines et d'intégrer ces mesures lors de l'apprentissage de classifieurs. Les *SVM* sont sensibles à ce type d'informations. Les travaux de [Chi and Bruzzone, 2007] proposent de mettre à jour la fonction de coût utilisée dans le calcul de l'hyperplan de séparation du *SVM*, celle-ci étant estimée à partir d'échantillons des deux domaines simultanément. Dans

la même idée, les travaux présentés dans [Gomez-Chova et al., 2008] réalisent une régularisation du SVM à partir d'un graphe Laplacien non normalisé.

Ces travaux intéressants d'un point de vue méthodologique, présentent néanmoins plusieurs désavantages liés à la problématique abordée et sa composante opérationnelle. La plupart d'entre eux sont fondés sur des approches itératives. Chaque itération intègre de nouveaux échantillons du domaine Cible au jeu d'apprentissage. La phase d'apprentissage peut alors devenir extrêmement chronophage. La possibilité d'exploiter plusieurs domaines Source semble aussi complexe. Elle requiert le renouvellement de l'apprentissage du classifieur initial à chaque ajout d'un nouveau domaine Source.

L'adaptation d'un RF peut s'avérer aussi complexe à mettre en oeuvre : cela requiert d'agir sur chaque critère de séparation au sein de chaque feuille. Le coût d'apprentissage lié au RF n'est pas un critère déterminant qui contraint l'utilisation de ces techniques semi-supervisées.

4.2.2 Adaptation de classifieur par apprentissage actif

L'apprentissage actif (ou *Active Learning* en anglais) désigne les méthodes impliquant une interaction de l'utilisateur lors du processus. Le principe de cette approche permet à une quantité minimale d'étiquettes d'améliorer grandement les méthodes semi-supervisées. L'utilisateur intervient alors pour étiqueter les échantillons du domaine Cible [Matasci et al., 2012, Jun and Ghosh, 2008, Rajan et al., 2008], et peut également être responsable de la réalisation d'une nouvelle itération ou de l'arrêt de l'algorithme. On peut envisager une sélection d'échantillons sur de petites zones, mais la sélection manuelle représente une charge de travail conséquente pour la classification de grandes étendues. Dans cette optique, des travaux proposent des aides à la sélection fondées sur la détection de changements, la mesure de similarité ou d'entropie voire même du clustering [Alajlan et al., 2014, Persello et al., 2014, Stumpf et al., 2012]. Néanmoins, le clustering ne représente pas une méthode fiable pour traiter des jeux de données à grande dimension. En effet, les algorithmes exploitent en grande partie la distance euclidienne, qui perd en fiabilité quand le nombre de dimensions augmentent.

La notion de vastes étendues est également un critère important. Dans les travaux cités, elle désigne une zone couverte par une tuile Landsat qui au demeurant est moindre qu'une emprise Sentinel-2.

En suivant les notations introduites par [Li and Sethi, 2006], le quintuplet (C, L, S, U, Q) définit un apprentissage actif. Dans ce quintuplet, C est un classifieur, L le jeu d'échantillons étiquetés d'apprentissage, S est l'utilisateur en charge de déterminer les étiquettes des échantillons retenus dans l'ensemble des non étiquetés U . Les échantillons U sont triés selon le critère Q . C'est le critère Q qui différencie une simple stratégie d'échantillonnage et l'apprentissage actif. Les sorties du classifieur C impactent généralement l'estimation Q . Par exemple, les échantillons capables de devenir des vecteurs support dans le cadre d'un SVM constituent l'ensemble U , la distance à l'hyperplan séparateur servant de critère de tri [Mitra et al., 2004]. Un autre exemple consiste à exploiter la mesure de confiance fournie par le classifieur [Demir et al., 2010, Tuia et al., 2009]. Le premier échantillon de U est sélectionné puis étiqueté.

Cependant, inclure un seul échantillon par itération soulève un problème majeur d'efficacité car chaque ajout implique une nouvelle phase d'apprentissage. De plus, n'ajouter qu'un seul échantillon par itération n'exclut pas l'ajout d'information fortement corrélées par rapport à l'itération précédente. En fonction du classifieur la redondance d'information peut dégrader ses performances. Pour palier ce problème il est d'usage d'ordonner les échantillons U par groupes. Chaque groupe répond à un critère de diversité minimum, et celui présentant une diversité maximale est sélectionné. De nombreux travaux portent leur intérêt autour de

l'estimation du critère de diversité, exploitant les mesures fournies par les classifieurs [Demir et al., 2010] ou en mesurant la similarité entre échantillons d'un même groupe [Di and Crawford, 2011].

La finalité de nos travaux de thèse vise la classification de la France entière (+100 tuiles Sentinel-2), et donc une approche complètement automatique. De même, sur un volume de données important, il n'est pas envisageable d'itérer des apprentissages et des classifications très nombreux. Bien que ces approches semblent très prometteuses vis-à-vis des approches semi-supervisées, l'interaction avec l'utilisateur est totalement réhibitoire dans le contexte de ces travaux.

4.2.3 Sélection de primitives invariantes

La sélection de primitives invariantes est la troisième catégorie d'adaptation de domaine. Dans la littérature, elle se concentre sur un cas bien particulier : l'utilisation d'images issues de capteurs différents ou dépendantes de conditions d'acquisitions différentes. En effet, elle cherche à corriger les effets de variations liés à l'angle de vue du satellite, de l'illumination de la scène au moment de l'acquisition et des distorsions géométriques [Izquierdo-Verdiguier et al., 2013].

On peut définir l'invariance d'une primitive p comme la préservation de ces propriétés lors de l'application d'une transformation quelconque T . Ainsi on obtient toujours $T(p) = p$. L'identification des primitives invariantes porteuses des informations discriminantes pour l'OCS, ou les informations spatiales associées, devient l'objectif principal à atteindre [Bruzzone and Persello, 2009]. La principale difficulté résulte de l'identification des données discriminantes pour ces deux aspects à la fois.

Les variations induites par un changement d'angle ou d'illumination se cumulent aux distorsions déjà présentées dans la section 2.1.3, complexifiant l'utilisation de telles données. Dans la mesure où les images constituant les séries temporelles sont corrigées de ces effets et exprimées en réflectance de surface, ces problèmes ne nous concernent pas directement.

4.2.4 Adaptation de données

L'adaptation de données constitue une des approches les plus utilisées dans la littérature. Elle permet de corriger les distorsions entre les domaines afin d'extraire les informations du domaine Source et les appliquer au domaine Cible sans difficulté. Cette catégorie regroupe un ensemble d'algorithmes à complexité de calcul et d'utilisation variables. Des méthodes très simples existent, fondées sur l'adaptation d'histogrammes [Inamdar et al., 2008], à partir de l'histogramme des deux images, nous définissons une fonction permettant de transformer la distribution du domaine Cible à partir de l'histogramme du domaine Source. Cette approche reste très limitée pour des images de grande dimension (nombreux canaux et grande étendue), ou pour des zones d'études hétérogènes. Une autre approche consiste à projeter les deux domaines dans un seul en exploitant l'Analyse en Composantes Principales (ACP) ou l'Analyse Canonique des Corrélations (ACC) [Nielsen, 2007, Volpi et al., 2015]. Généralement, ces deux méthodes constituent la première alternative considérée lorsque nous cherchons à réaliser une réduction de dimension ou une projection de données.

Dans une autre optique, des travaux tentent de conserver l'information de géométrie lors de la projection [Gonzalez et al., 2015, Montoya-Zegarra et al., 2013, Yang and Crawford, 2016]. Cette information est importante, car représentative de la diversité de la classe au sein de la parcelle. Ces méthodes reposant sur une comparaison pixel à pixel, demeurent très sensibles au changement d'OCS, ce qui introduit une distorsion supplémentaire. Pour palier ce type de problèmes, il existe des méthodes de projection fondées sur les

distributions dans les séries temporelles [Tuia and Camps-Valls, 2016] mais elles exploitent les étiquettes dans les deux domaines.

Des approches d’alignement de séries temporelles ont été proposées comme la *Dynamic Time Warping* (ou *déformation temporelle dynamique en français*) (DTW) [Petitjean et al., 2011], ou exploitant les variétés¹ (*manifold*) [Tuia et al., 2014]. Ces méthodes sont associées à un coût de calcul proportionnel aux dimensions des séries temporelles.

Récemment le Transport Optimal [Courty et al., 2017] a été utilisé comme méthode d’adaptation de domaine avec succès. Le principe de cette approche est de transporter à moindre coût l’intégralité des échantillons du domaine Source vers le domaine Cible. Un classifieur entraîné sur les échantillons du domaine Source transportés est donc applicable aux échantillons du domaine Cible.

4.3 Autres approches d’exploitation de l’historique

Il existe des approches permettant d’exploiter un historique d’OCS afin de pallier le manque de données de référence. Nous en avons sélectionné deux : la détection de changements, et les protocoles de mise à jour des cartes d’occupation des sols.

4.3.1 Détection de changements

Les algorithmes de détection de changements cherchent à déterminer les éléments qui ont subi un changement entre deux scènes acquises à différentes dates. Un changement, dans le cadre des images de télédétection, correspond à toute altération de la surface entre deux moments d’observation. Par cette définition, toute évolution naturelle des classes d’OCS, à l’image du cycle de la végétation, constitue une source de changement.

Dans le cadre de cette thèse, une approche de détection de changements constitue un avantage considérable. L’idée étant de pouvoir exploiter les éléments qui ne changent pas. En effet, dans ce cas, il est possible d’exploiter les étiquettes déjà connues. Ce transfert d’étiquette nous permettrait alors d’étiqueter partiellement le domaine Cible, facilitant l’adaptation de domaine ou permettant l’apprentissage d’un classifieur supervisé. La détection de changements permettrait également de réduire le volume de données traitées en ne cherchant une étiquette que pour les échantillons ayant subi un changement.

Les travaux de [Coppin et al., 2004, Zhu, 2017] et plus récemment [Liu et al., 2019] recensent différentes approches de détection de changement, adaptées aux images multispectrales et hyperspectrales.

La détection de changements vise deux objectifs distincts : détecter l’élément changeant (l’altération) ou alors identifier la cause d’un changement lambda (l’agent de changement : feu de forêts, tornades, expansion urbaine, ...).

Sélectionner un algorithme de détection de changements résulte de la combinaison de trois facteurs distincts : l’application souhaitée, la disponibilité de données de référence et le degré d’automatisation. Le cadre de la thèse se focalise sur les deux derniers critères, seules les approches totalement automatiques et n’utilisant pas de données de référence peuvent être envisagées. L’application souhaitée représente le produit attendu en sortie de l’algorithme de détection :

- une carte de changement binaire : changement ou non

¹un espace topologique dans lequel chaque point ressemble à un espace Euclidien

- une carte comportant plusieurs classes indiquant des changements de différentes natures
- la détection d’agent de changement au sein d’une série temporelle

L’obtention d’une carte binaire se révèle assez simple, le moyen le plus courant [Margono et al., 2012, Hansen et al., 2013] consistant à réaliser une classification (supervisée) des deux images et de calculer les différences entre ces deux cartes via une simple fonction binaire (0 si les cartes sont identiques et 1 dans le cas contraire, par exemple). Le résultat obtenu indique alors si tel pixel a changé. On trouve également des méthodes exploitant des approches de segmentation [Franklin et al., 2015], des opérations de filtrage à partir de seuils sur des bandes spectrales ou indices [Chen et al., 2012b, Hilker et al., 2009] ou encore des analyses statistiques [Brooks et al., 2014, Dutrieux et al., 2015]. Ces travaux sont essentiellement orientés sur des zones forestières, sujettes à des changements minimes par rapport à des zones d’activités intenses comme les champs agricoles.

Les deux derniers cas d’application concernent essentiellement l’identification des agents de changements.

Les causes de changement constituent une variable d’un grand intérêt, surtout dans l’analyse de l’utilisation des sols. On distingue les causes naturelles telles que les grands incendies, ou des conséquences liées à l’activité humaine telle que l’agriculture. Autant d’agents qui contribuent aux besoins réguliers d’une mise à jour des cartes d’OCS. On dénombre une quantité non négligeable d’algorithmes différents de détection des agents de changements dédiés à la détection d’un type d’agent. Nous pouvons citer en exemple, le calcul de la différence entre deux cartes résultantes de classifications pour la détection de l’expansion urbaine [Li et al., 2015, Ramadan et al., 2004, Mihai et al., 2015], mais également le suivi des forêts exploitées pour le bois de chauffage [Wilson and Sader, 2002]. On retrouve également des mesures de phénomènes naturels spécifiques comme le passage de tornades [Pardo-Pascual et al., 2014], les ravages des incendies [Parker et al., 2015] ou l’effet d’une invasion d’insectes sur les forêts [Senf et al., 2015, Meigs et al., 2015]. Ces travaux identifient avec succès les agents de changement, mais uniquement un par un.

Cependant, ces travaux réalisent pour la plupart une différence entre cartes d’OCS. Dans notre contexte, ces approches sont inutilisables. En effet, nous ne possédons aucun moyen d’assurer la précision de la classification naïve, ni aucune autre version de la carte. Ainsi, les méthodes de détection de changements exploitant directement les images doivent être privilégiées.

Les travaux de [Neigh et al., 2014] exploitent plusieurs mesures de différences entre bandes spectrales et indices pour détecter tous les agents de changement sur une zone forestière puis les identifient grâce à une classification supervisée. Dans le même esprit, [Alaibakhsh et al., 2015] proposent d’utiliser le *Multivariate Alteration Detector (MAD)* [Nielsen and Conradsen, 1997] pour identifier six agents de changement différents à partir des six premières cartes de différences. Le MAD génère, à partir de deux images, des cartes de différences non corrélées entre elles. La première carte contient les changements à plus grande échelle, causés par une variation climatique par exemple, alors que les dernières cartes de différences contiennent des changements à plus petite échelle comme une trace de l’activité de l’homme. Le chapitre 5.4 détaille le MAD.

La détection de changements dans les séries temporelles constitue un centre d’intérêt relativement récent. Quelques travaux très spécifiques abondent dans ce sens. L’essentiel des travaux portant sur les séries temporelles se concentrent sur l’identification des facteurs de changement. Nous pouvons citer à titre d’exemple, la modélisation d’un signal à partir des informations spectrales, spatiales et temporelles de celui-ci [Hemissi et al., 2013]. Cette approche permet grâce à la qualité du signal reconstruit de détecter plus précisément les changements dûs aux cycles saisonniers. La définition de “série temporelle” peut varier, alors que certaines

approches demandent une fréquence d’acquisition journalière, d’autres se contentent d’une image par an pour obtenir une série temporelle couvrant 5 années [Zanotta et al., 2015]. Dans ces travaux les auteurs parviennent à détecter l’impact de la déforestation et des incendies sur une région d’étude fixe au cours de 5 ans continus.

4.3.2 Mise à jour des cartes d’occupation des sols

Les méthodes de détection de changements ouvrent une opportunité intéressante : la mise à jour des cartes antérieures pour produire la carte du millésime attendu. En effet, rappelons ici que tout pixel non associé à un changement détecté conserve sa classe d’OCS, réduisant ainsi l’effort déployé pour prédire la carte du domaine Cible.

La littérature offre de nombreuses approches, proposant des stratégies variées pour la partie détection, mais également pour prédire les classes des pixels soumis à un changement. Il n’existe, à notre connaissance, aucune méthode n’exploitant pas l’adaptation de domaine afin de réaliser une mise à jour d’une carte existante. En effet, les études présentées exploitent un ou plusieurs algorithmes de détection de changements voire d’adaptation de domaine, introduits préalablement. Dans [Bruzzone and Marconcini, 2009] les auteurs proposent un algorithme d’adaptation de domaine semi-supervisé qui incorpore et remplace au fil des itérations les échantillons du domaine Source par des échantillons du domaine Cible. Dans ces travaux, les auteurs reformulent la problématique de mise à jour pour réduire la problématique à un simple problème d’adaptation de domaine, ne réduisant pas l’effort de production. Dans le même esprit, dans [Bahirat et al., 2012] les auteurs exploitent des mesures de probabilités a posteriori pour émettre des hypothèses sur les différences entre les classes des deux domaines. Une fois formulées, ces hypothèses sont validées après corrections des distorsions, grâce à une adaptation de domaine, afin de les exploiter et prédire la carte du domaine cible.

Les travaux de [Chen et al., 2012a], proposent un algorithme itératif de mise à jour exploitant une détection de changements et plusieurs étapes de classification. L’intérêt de cette étude réside dans l’intégration du voisinage lors de la détection de changements, limitant la détection de pixels isolés (bordure ou autres).

Plus récemment, un protocole de mise à jour visant, non pas la carte, mais les données de référence a été proposé par [Paris et al., 2019]. Exploitant une carte existante, en l’occurrence CLC, les auteurs parviennent à corriger les étiquettes d’une donnée de référence obsolète, permettant ainsi une classification supervisée d’une série temporelle. Dans cette approche, ils exploitent la diversité au sein d’une parcelle identifiée par la donnée de référence pour choisir l’étiquette la plus pertinente. Néanmoins, on relève un coût important, appliquant un facteur 10 aux temps de traitement d’un simple RF qu’ils compensent par un gain équivalent en PG d’environ 13%.

4.4 Synthèse

Au travers de ces quatre catégories d’algorithmes constituant l’adaptation de domaine pour la télédétection, on relève une très grande diversité des solutions possibles. Dans leur revue [Tuia et al., 2016] les auteurs proposent également un guide pour la sélection des méthodes en fonction de la problématique retenue. Par exemple, l’apprentissage actif et les approches semi-supervisées sont fortement recommandés lorsque un classifieur est disponible pour le domaine Source et que son coût d’apprentissage s’avère important. Dans le cas où les distributions par classes dans les deux domaines peuvent être connues, on privilégie les méthodes fondées sur l’alignement des variétés (*manifold*). En d’autres termes, on conçoit les adaptations de classifieurs

afin de gérer les fortes disparités entre les domaines, comme la situation multi-capteurs ou dans le cas de coûts d'apprentissage élevés.

Pour les méthodes privilégiant l'adaptation de données et la sélection de primitives, le choix d'une méthode dépend essentiellement de la configuration du problème. Si les distorsions entre les domaines relèvent d'une moindre importance, les approches [ACP](#) et [ACC](#) seront performantes pour un coût acceptable. À l'inverse elles seront pénalisées dans le cas où les distorsions s'avèrent importantes ou trop complexes : non linéaires car la distance euclidienne ne constitue pas une mesure pertinente dans un cas à grande dimension (ce qui est le cas avec les séries temporelles). Cette revue donne un conseil utile, à savoir la possibilité d'exploiter des étiquettes dans le domaine Cible garantissant de meilleures performances, car cela permet souvent d'impacter l'adaptation réalisée.

Ces dernières années, les travaux publiés concernent essentiellement l'amélioration des méthodes existantes, en proposant des combinaisons, avec un intérêt prononcé pour des méthodes d'apprentissage actif et de l'intégration de l'apprentissage profond (*deep learning*) [[Postadjian et al., 2018](#)] dans ces approches. Néanmoins, il semble que les solutions aux problèmes à forte dimensionnalité (spectrale, temporelle ou spatiale) se révèlent assez rares dans la littérature.

Au cours de cette étude de la bibliographie, notre second constat vient contraster la variété des méthodes proposées. En effet, trois zones sont principalement exploitées dans ces travaux : l'université de Pavia, la zone *Indian Pine* et le delta de Okavango en Botswana. Ces jeux de données se révèlent très représentatifs des distorsions nécessitant de réaliser une adaptation de domaine. Pour les jeux de données Pavia et Indian Pine, les images présentent de nombreuses dimensions en raison de leur acquisition par deux capteurs aéroportés hyperspectraux ROSIS et AVIRIS. La différence avec notre zone d'étude ne réside plus alors dans le nombre de dimensions à traiter, mais par l'information qu'elles apportent. En effet, les images aéroportées représentent une scène pour une date d'acquisition unique. Alors que la série temporelle transcrit l'évolution de la zone sur une année complète, les distorsions observées couvrent alors l'intégralité des deux années.

De même, lors de l'évaluation des méthodes de détection de changements, nous avons constaté que les principales zones d'études focalisent les forêts et comparent des acquisitions monodates.

La plupart des méthodes de détection de changements considèrent deux images correspondantes à deux vues différentes de la même scène. Les méthodes traitant les séries temporelles sont rares, et requièrent en priorité leurs décompositions par saison pour un suivi annuel, ou se contentent d'images successives en grand nombre sur un pas de temps restreint. Ce découpage saisonnier peut représenter un obstacle à une production opérationnelle à grande échelle, car il multiplie alors les étapes de traitements, d'apprentissages et de classifications.

Les approches de mises à jour des cartes d'OCS souffrent malheureusement des mêmes contraintes, puisqu'elles reposent sur des méthodes d'adaptation de domaine ou de détection de changements. L'approche proposée par [[Paris et al., 2019](#)] semble prometteuse mais n'a pas pu être abordée au cours de cette thèse.

Une attention particulière doit être apporté quant aux choix des méthodes abordés, en terme de complexité de mise en oeuvre et de coûts de traitements. En effet, l'utilisation combinée de séries temporelles annuelles Formosat-2 et d'au moins 17 classes d'OCS ne représente qu'un aperçu de la complexité de la tâche lorsque les données Sentinel-2 seront exploitées sur une zone étendue.

Du point de vue méthodologique, le sujet de la thèse et les données utilisées nous amène à privilégier les méthodes d'adaptation de données potentiellement associées à une approche de détection de changements. Les coûts d'apprentissage du [RF](#) minimisent l'intérêt de l'adaptation de classifieurs. De même, l'interaction

avec l'utilisateur de l'apprentissage actif, nous amène à l'écarter des pistes envisagées.

Analyse Canonique des Corrélations

Face aux très nombreuses approches existantes dans la revue de littérature, nous avons choisi de focaliser nos travaux sur les méthodes d'adaptation de données, afin de réduire les distorsions entre les domaines et ainsi améliorer les performances de la classification naïve. Par conséquent, nous avons choisi de nous intéresser à l'[Analyse Canonique des Corrélations \(ACC\)](#) car cet algorithme permet d'étudier d'une part une solution de projection de données très simple à mettre en place, et d'autre part un algorithme de détection de changements.

5.1 Introduction

L'[Analyse Canonique des Corrélations \(ACC\)](#) publiée pour la première fois en 1936 par Hotelling [[Hotelling, 1936](#)] constitue une des approches d'analyse statistique les plus utilisées dans de nombreux domaines d'applications : économie, médecine, génétique, entre autres. Une rapide analyse bibliométrique met en avant pas moins de 10000 publications en rapport avec l'ACC lors des cinq dernières années.

L'ACC partage des points communs avec l'[ACP](#), notamment pour la construction d'un ensemble de nouvelles primitives descriptives, ainsi qu'avec la régression linéaire, pour la nature des données. L'objectif général de l'ACC consiste à déterminer les relations existantes entre deux groupes de variables quantitatives observées sur le même événement, en exploitant la corrélation entre celles-ci.

L'analyse des relations entre deux groupes de variables représente la principale caractéristique de l'ACC, la plaçant comme méthode centrale de la statistique descriptive multidimensionnelle. En effet, elle généralise plusieurs autres approches. Par exemple, la régression linéaire multiple qui permet de donner une explication d'une variable quantitative par un groupe d'autres variables. Elle devient un cas de l'ACC lorsque l'un des deux groupes considérés ne comporte qu'une seule variable. D'autres cas particuliers existent, comme l'analyse factorielle discriminante [[Benzecri, 1977](#)] ou l'analyse factorielle des correspondances [[Escofier-Cordier, 1969](#)]. L'[ACP](#) représente également un cas particulier de l'ACC, obtenu si chaque groupe ne contient qu'une seule variable quantitative. Dans ce chapitre nous ne nous intéressons qu'à la formulation classique de l'ACC qui

ne considère que deux groupes de variables quantitatives : le domaine Source et le domaine Cible représentés par leur série temporelle respective.

5.2 Principe de l'analyse canonique des corrélations

Considérons une variable aléatoire de dimension $(p + q)$ avec $(p \leq q)$, suivant idéalement une distribution gaussienne N , de moyenne μ nulle et de variance Σ , séparée en deux groupes de dimensions p et q respectivement :

$$\begin{bmatrix} X \\ Y \end{bmatrix} \in N(\mu, \Sigma) = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right).$$

\mathbf{X} et \mathbf{Y} représentent les échantillons :

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \text{ et } \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix}.$$

Et on suppose, sans perte de généralité que les données sont normalisées et centrées en 0, donc que les moyennes $\mu_X = \mu_Y = 0$. On note Σ_{11} la matrice de dispersion (appelée matrice de variance-covariance) associée à la variable \mathbf{X} , Σ_{22} associée à \mathbf{Y} , et Σ_{12} la covariance entre ces deux variables. On cherche des combinaisons linéaires $a^T \mathbf{X}$ et $b^T \mathbf{Y}$ permettant de maximiser la corrélation ρ entre les deux variables aléatoires :

$$\begin{aligned} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} a &= \rho^2 \Sigma_{11} a, \\ \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} b &= \rho^2 \Sigma_{22} b, \end{aligned}$$

ou sous forme de quotients de Rayleigh :

$$\begin{aligned} \rho^2 &= \frac{a^T \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} a}{a^T \Sigma_{11} a}, \\ &= \frac{b^T \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} b}{b^T \Sigma_{22} b}. \end{aligned}$$

On obtient les projections voulues pour la variable \mathbf{X} en considérant les vecteurs propres conjugués (mutuellement orthogonaux) a_1, \dots, a_p correspondant aux valeurs propres $\rho_1^2 \geq \dots \geq \rho_p^2 \geq 0$ de la matrice $\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$ par rapport à Σ_{11} . Similairement, on obtient les projections de \mathbf{Y} en considérant b_1, \dots, b_q , les vecteurs propres de la matrice $\Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ par rapport à Σ_{22} associés aux mêmes valeurs propres ρ_1^2 .

On définit les variables canoniques par :

$$V_{1i} = a_i^T \mathbf{X}, i = 1, \dots, p,$$

$$V_{2i} = b_i^T \mathbf{Y}, i = 1, \dots, p,$$

que nous noterons :

$$V_1 = A^T \mathbf{X},$$

$$V_2 = B^T \mathbf{Y},$$

avec :

$$A = [a_1, \dots, a_p] \text{ de taille } p \times p,$$

$$B = [b_1, \dots, b_p] \text{ de taille } q \times p.$$

Dans notre cas, \mathbf{X} et \mathbf{Y} correspondent, respectivement, aux réflectances de chaque date de chacune des deux séries temporelles, l'ACC en est la projection dans un espace commun où leur corrélation est maximale.

5.3 Une approche d'adaptation de domaine

Dans le cadre de nos travaux, l'ACC offre des perspectives attractives. Elle permet de traiter, en théorie, des données de diverses origines, par exemple des satellites différents, dans la mesure où l'ACC est invariante par rapport aux facteurs constants. De tels facteurs sont fréquemment associés à l'angle de vue et l'illumination de la scène lors de l'acquisition.

La figure 5.1 illustre le principe de l'ACC en tant que méthode pour l'adaptation de données. Le domaine Source, représenté par une série temporelle, constitue une des variables aléatoires, et le domaine Cible, également une série temporelle, constitue la seconde variable aléatoire. Les matrices V_1 et V_2 contiennent les combinaisons linéaires, associées à chaque variable aléatoire ainsi qu'à une valeur de corrélation maximale unique. Elles permettent ainsi la projection des deux séries temporelles dans l'espace maximisant la corrélation entre elles. Dans notre cas d'application, les données de référence ne peuvent pas étiqueter directement les données provenant du domaine Cible, car cette approche ne résout pas l'obsolescence de l'information géographique. Cependant, un classifieur entraîné sur les données étiquetées du domaine Source après l'étape de projection pourra classer les données provenant du domaine Cible.

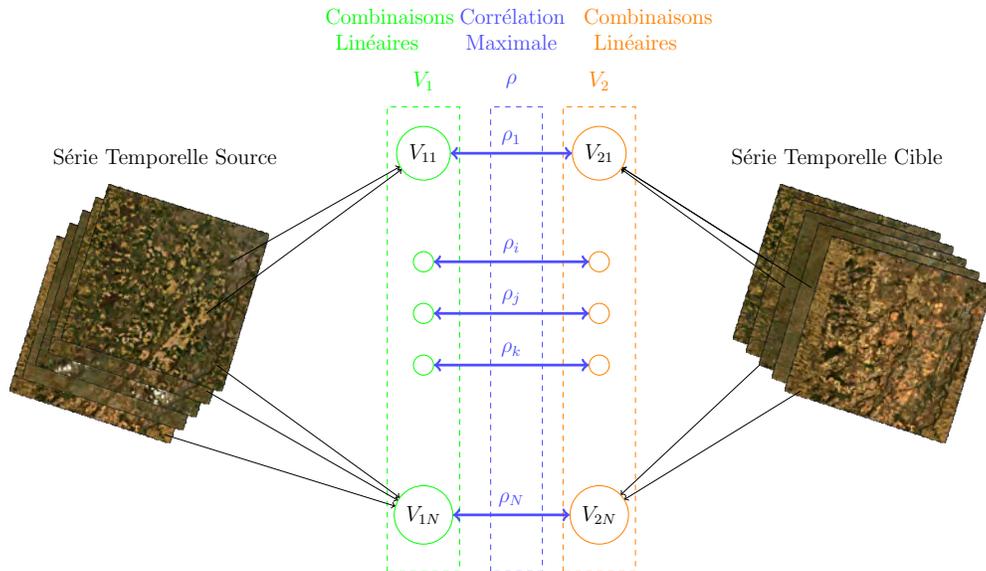


FIGURE 5.1 : Principe de l'analyse canonique des corrélations exploitant des séries temporelles.

L'ACC traite des données de dimensions différentes, permettant d'éviter le ré-échantillonnage temporel, réduisant ainsi les délais de pré-traitement. Cet aspect permettrait également de produire une carte au fil de

l'arrivée des images de l'année courante, en déterminant les primitives disponibles les plus proches avec celles de l'historique.

La linéarité de l'estimation des matrices de projection constitue un avantage important car elle autorise le recours aux paradigmes de programmation parallèle, et par conséquent des temps de traitement optimisés via une structure [Calcul à Haute Performance \(High Performance Computer\) \(HPC\)](#).

L'ACC permet de déterminer les relations entre deux jeux d'observations de la même scène. Néanmoins, le contexte de ces travaux ne remplit pas entièrement ces conditions. En effet, nous cherchons des relations entre deux mêmes scènes acquises, mais à des temps différents. De plus, ici la notion du temps d'observation de la scène porte sur une année complète et non pas sur une date unique. Dès lors, les changements entre deux années, constituent alors une source potentielle d'erreur dans l'estimation de l'ACC. Par exemple, les nuages présents dans les acquisitions représentent un cas typique d'artefact pouvant interférer avec une approche comme l'ACC [Nielsen et al., 1998] (et par extension à la régression linéaire [de Carvalho et al., 2013], l'ACP, etc). Nielsen identifie deux autres facteurs représentant les plus gros risques de dégradation de l'ACC : les erreurs de recalage (mise en correspondance des pixels des différentes images) et les effets de capteurs. Néanmoins, les erreurs de recalage représentent un facteur rare, dans la mesure où les images sont pré-traitées et corrigées en amont de la mise à disposition. Cependant, un changement d'OCS peut être assimilé à une erreur ponctuelle de recalage, et constitue donc un facteur potentiel de dégradation de l'ACC. De même, les effets de capteurs ne constituent pas un problème dans la mesure où nous utilisons des images d'un satellite héliosynchrone¹, ce qui limite les effets de changement d'angle ou d'illumination, l'altitude, etc.

5.3.1 Supprimer toute source de perturbations potentielles

Principe

L'ACC exploite un calcul de corrélation au niveau du pixel, impliquant que chaque source de bruit potentielle impacte l'estimation de la corrélation entre deux paires de variables canoniques. Par exemple, une forte présence nuageuse sur plusieurs dates disponibles va créer une zone de valeurs constantes sur plusieurs primitives, constituant des artefacts influant la corrélation. La présence d'un tel artefact dans une série temporelle constitue un problème majeur, essentiellement en raison de l'absence de son correspondant dans l'autre série temporelle.

Partant de ce constat, les changements d'OCS représentent une autre source de bruits potentiels pour le calcul des corrélations au niveau du pixel, dans la mesure où ces classes possèdent des caractéristiques distinctes.

Dans l'optique de réaliser une ACC la plus performante possible, nous avons proposé d'appliquer un masque aux deux séries temporelles, afin de minimiser leurs différences. Nous réalisons ce masque commun à partir des deux séries temporelles. Ce masque binaire associe une valeur non nulle si le pixel remplit une des conditions suivantes :

- Ce n'est pas un bord de l'image (absence de données).
- Un taux de validité supérieur à un tiers du nombre total d'images de la série temporelle.
- Une classe pérenne, identifiée à partir de la classification supervisée du domaine Source. Les classes de

¹en orbite polaire avec des acquisition à angles constants

Feuillus, Conifères, Surfaces Artificielles, Rivières, Lacs, Gravières constituent les six classes pérennes dans notre cas.

Les éléments non nuls sont utilisés lors de l'estimation des matrices de projection. Ce masque est appelé "masque de classes annuelles" dans la suite.

Nous évaluons les deux premières conditions dans les deux séries temporelles, la troisième ne concerne que le domaine Source.

Dès lors, nous réalisons la projection des données en deux temps, tout d'abord lors de l'estimation des matrices de projection V_1 et V_2 nous appliquons le masque aux deux séries temporelles, interdisant l'utilisation de ces pixels. Puis nous projetons les séries temporelles non masquées, avant de réaliser l'étape de classification présentée dans la section 3.1.

Performances obtenues

La figure 5.2 présente la PG obtenue par cette approche. Cette figure comporte trois matrices, la première présente les valeurs de PG, celle au centre l'écart entre la PG obtenue et celle du cas naïf et la troisième les intervalles de confiance à 95 %. L'échelle de couleur diffère en fonction de la matrice, celle de gauche utilise le rouge pour une PG proche de 0 et le bleu pour 1, celle du milieu utilise une palette de couleur divergente, une teinte claire signifiant un écart faible. La variable bleue indique un écart positif, alors que le rouge signale un écart négatif. L'échelle de couleurs des intervalles de confiance est irrégulière, ce qui permet d'observer la forte dynamique au sein des valeurs très faibles (cette représentation est commune à toutes les approches d'adaptation de domaine présentées dans cette partie).

La matrice de gauche présente des valeurs de PG globalement supérieures à 50 %, avec une moyenne tout cas confondus de 61 %, contre une moyenne de 70 % pour le cas naïf. Néanmoins la matrice de différence nous informe que, même si majoritairement les performances sont moins élevés, dans 12 cas l'ACC produit une carte d'au moins aussi bonne précision que le cas naïf. Les intervalles de confiance, très petits (< 5 %) mettent en évidence la cohérence de la projection des données à l'échelle de la zone d'étude.

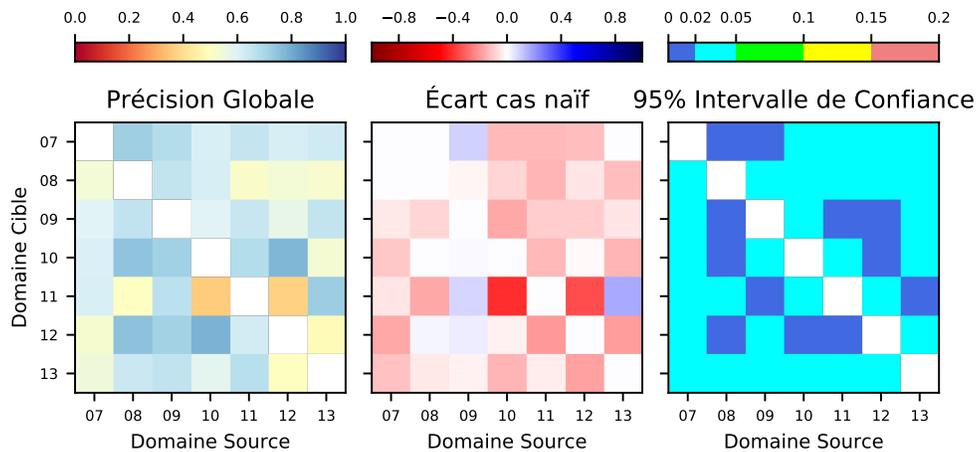


FIGURE 5.2 : PG obtenue lors de la projection ACC minimisant toute source de bruit potentielle (nuages et classes annuelles).

Une fois encore, la PG ne constitue pas une source d'information suffisante pour comprendre l'origine des différences entre les méthodes. Pour tenter de mieux appréhender ce problème, nous présentons le FScore dans la figure 5.3. Cette figure recense 17 matrices (1 par classe) d'écart entre le FScore de la méthode et le cas naïf.

Nous pouvons noter que parmi les six classes pérennes utilisées pour estimer la projection, cinq bénéficient d'une amélioration vis-à-vis du cas naïf. Ainsi, les classes de Rivières et Lacs profitent le plus de cette projection améliorant leur FScore de plus de 50 % (pour rappel la figure 3.2 présente le FScore du cas naïf). De même, la classe de Feuillus enregistre une amélioration mais moins importante. Néanmoins, le cas naïf offrait déjà une très bonne performance avec un FScore supérieur à 90 % pour cette classe.

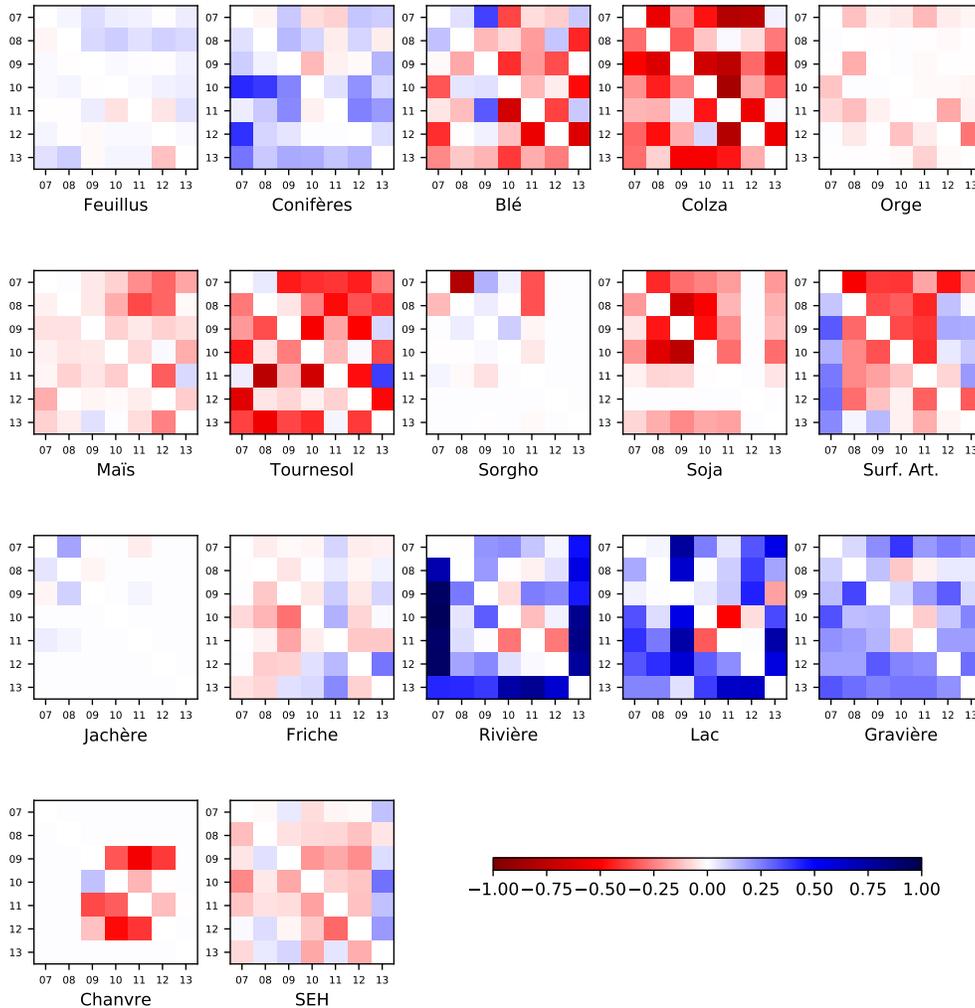


FIGURE 5.3 : FScore obtenus lors de la projection ACC minimisant toute source de bruit potentielle (nuages et classes annuelles).

Par contre, toutes les autres classes témoignent d'un l'effet inverse avec des chutes de FScore de plus de 50 %. Ces résultats indiquent que l'omission des classes annuelles dans l'estimation des matrices de projection

nuit probablement à la bonne performance de cette approche.

5.3.2 Conservation du maximum d'informations

Principe

Les résultats obtenus par la précédente approche ne permettent pas d'offrir une solution viable d'adaptation de domaine. Nous avons avancé l'hypothèse que la prise en compte des classes annuelles lors de l'estimation des matrices de projection pouvait améliorer les performances obtenues. Par conséquent, pour réaliser cette expérimentation nous conservons le protocole précédent mais nous modifions le calcul du masque commun. Nous ne plus considérons que deux conditions pour qu'un pixel soit utilisé dans l'estimation de l'ACC :

- Ce n'est pas un bord d'image.
- Un taux de validité supérieur à un tiers du nombre total d'images de la série temporelle.

Ce masque est appelé "masque de pixel altérés" dans la suite. Ces deux conditions sont évaluées sur les deux séries temporelles et nous appliquons ce masque commun lors de l'estimation des matrices de projection.

Performances obtenues

La figure 5.4 présente, sous le même format que la figure 5.2, les performances globales obtenues par cette approche. Visuellement, les performances sont améliorées, mais faiblement puisque la moyenne, tous cas confondus, atteint 62 % de PG (contre 61 % pour le cas précédent). Cependant, si les cas d'amélioration sont plus nombreux, l'amplitude des cas de dégradation se révèle plus importante (pour les paires 2010 et 2011 par exemple).

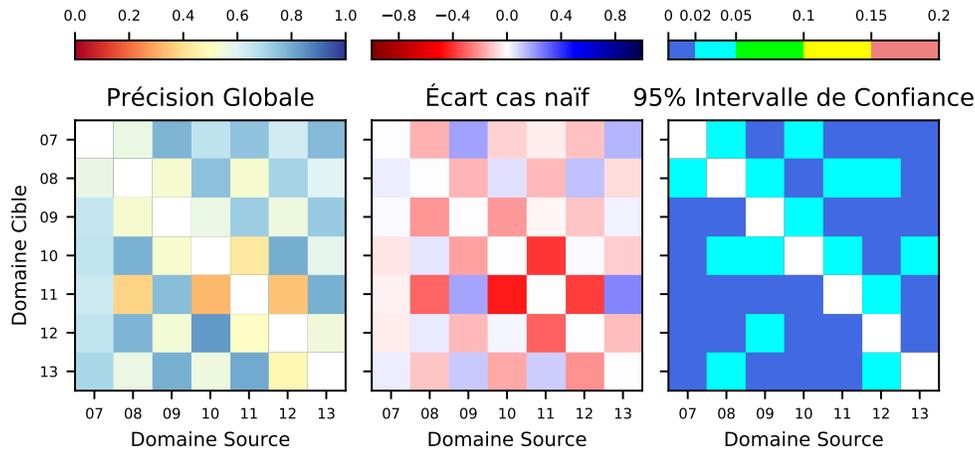


FIGURE 5.4 : PG obtenue lors de la projection ACC masquant les pixels altérés (nuages et bord).

De nouveau, le FScore permet de comprendre l'origine de tels écarts. Ainsi, la figure 5.5 présente les 17 matrices de différence de FScore entre cette approche et le cas naïf. Si nous pouvons retrouver dans la même gamme d'amplitude les gains observés sur les cinq classes pérennes (Feuillus, Conifères, Rivières,

Lacs et Gravières), nous pouvons constater quatre nouvelles classes bénéficiant de la projection ACC : le Maïs, les Jachères, les Friches, et les SEH. Ces quatre classes ont la particularité de rester stationnaires (géographiquement) au fil des ans, le Maïs pour des raisons de systèmes d'irrigation nécessitant une mise en oeuvre lourde, alors que les Jachères, Friches et SEH sont considérées comme des classes pluri-annuelles, favorables à la reconstitution des sols avant de nouvelles mises en culture.

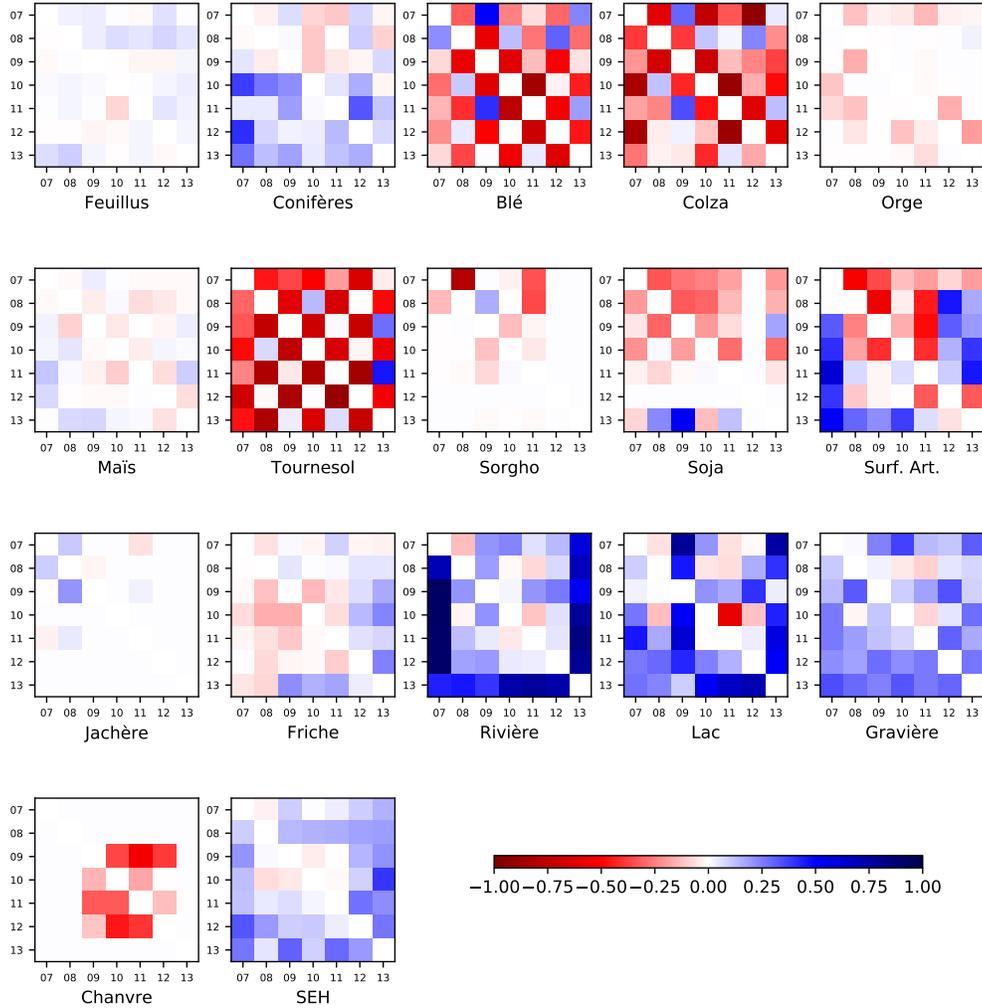


FIGURE 5.5 : FScore obtenus lors de la projection ACC masquant les pixels altérés (nuages et bord).

Les figures 5.2, 5.3, 5.4 et 5.5 mettent en évidence des écarts importants, atteignant au maximum 19 % de PG entre l'estimation de l'ACC avec deux masques différents. Nous avons calculé ces différences entre ces deux expérimentations ($PG_{\text{Altérés}} - PG_{\text{Annuelles}}$) pour détecter les écarts les plus significatifs. La table C.1 (en annexe) regroupe ces valeurs. Une valeur positive indique que l'approche du masque des pixels altérés produit de meilleures performances. Ainsi, parmi les 42 combinaisons possibles, nous enregistrons 27 écarts positifs, dont 16 présentent un écart supérieur à 5 %, et se situent donc au-delà des intervalles de confiance. Nous recensons 15 cas d'amélioration ou de dégradation sur 43, témoignant d'un écart supérieur à 10 %

dont 5 dépassent les 15 %. Pour comprendre ces écarts très importants nous proposons de synthétiser les performances par classe dans la figure 5.6, pour ces cinq cas représentatifs.

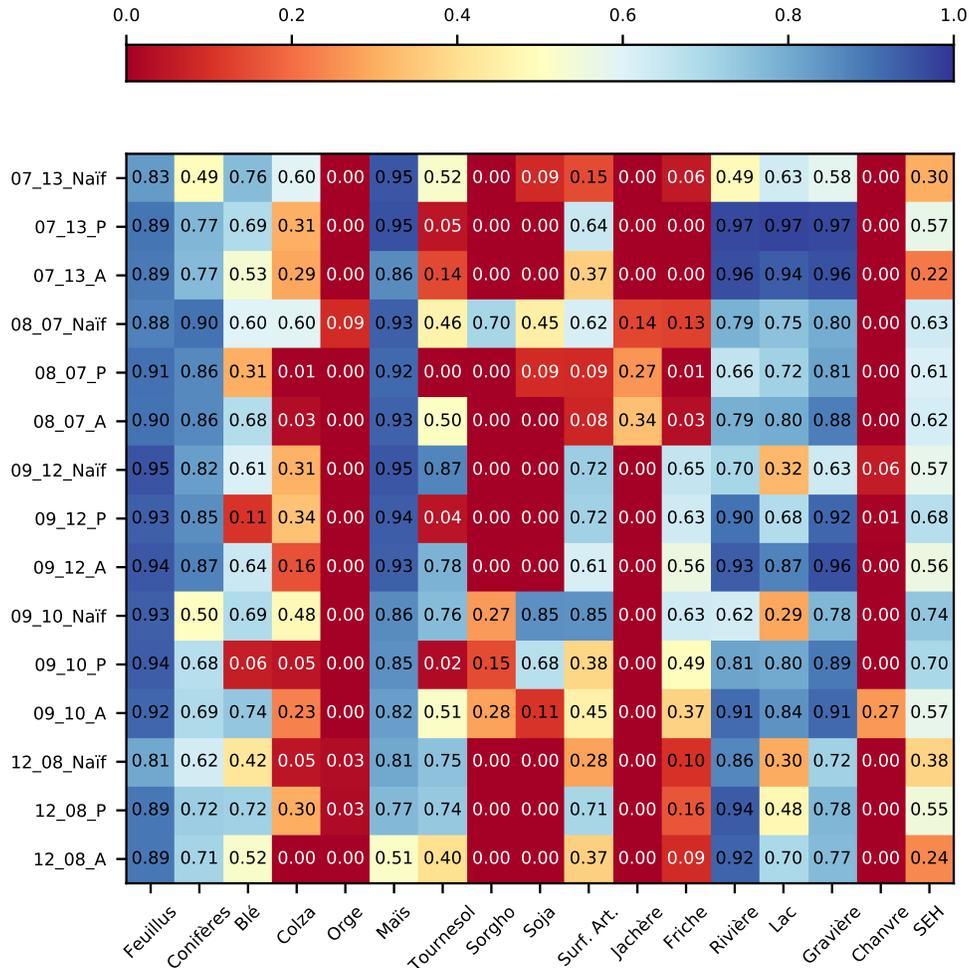


FIGURE 5.6 : FScore obtenus pour cinq cas particuliers par les deux approches de projection des données ACC utilisant le masque de classes annuelles (A) ou de pixels altérés (P). Ces deux approches sont comparées au cas naïf (Naïf).

Cette figure représente une matrice, l'échelle de couleur représente les valeurs de FScore entre le cas naïf ($S_C_Naïf$), l'utilisation de l'ACC avec le masque de pixels altérés (S_C_A) et le masque de classes pérennes (S_C_P), où S et C représentent le domaine Source et Cible pour les cinq cas retenus. Nous constatons que ces cas correspondent tous à des dégradations vis-à-vis des performances du cas naïf indépendamment du masque exploité. Ces cinq cas concernent les paires de domaines suivantes : (2007, 2013), (2008, 2007), (2009, 2010), (2009, 2012) et (2012, 2008). Cette figure nous permet de distinguer plusieurs tendances récurrentes.

La première tendance concerne le FScore sensiblement supérieur obtenu lors de l'utilisation du masque

de classes annuelles sur les cinq classes bénéficiant de cette approche (Feuillus, Conifères, Rivières, Lacs et Gravières). Cette différence systématique implique une variation sur la PG, mais ne justifie pas l'écart observé.

Ainsi, la majeure différence de performances s'observe sur les autres classes notamment les classes de Blé, Colza, Tournesol et [SEH](#) largement représentées dans les données de référence, et dans une moindre mesure les classes de Surfaces Artificielles et les Friches. Enfin, la faible et variable représentation des autres classes, sur l'ensemble des années, minimisent leur impact lors du calcul des performances.

La classe de Blé illustre parfaitement le grand impact de la proportion des classes sur les métriques, ce qui explique les écarts importants de performance observés. On observe une différence minimale de 16 % pour (2007, 2013) de FScore et maximale de 58 % pour le cas (2009, 2010).

L'analyse de ces mêmes classes révèle une seconde tendance : si le pas de temps entre l'année du domaine Source et Cible est pair ($2013-2007 = 6$), alors l'utilisation du masque de pixels altérés offre de meilleures performances sur les classes de culture. À l'inverse, pour ce même masque, un pas de temps impair implique une dégradation des performances pour ces mêmes classes.

Pour expliquer cet effet, nous pouvons nous référer aux particularités de l'OCS et notamment aux rotations de cultures. L'étude des pratiques agricoles dans le Sud-Ouest met en évidence une rotation de culture à deux classes [[Jouy and Wissocq, 2011](#)]. La délimitation des parcelles agricoles représente un facteur stable au fil des ans : seule la nature des cultures change. En additionnant ces deux éléments, nous pouvons supposer que la pertinence de la projection [ACC](#) implique que les pixels représentent la même classe d'OCS. En d'autres termes, les changements des cultures d'une année sur l'autre induisent en erreur l'algorithme.

On peut supposer que cet effet soit spécifique au sud ouest, les rotations étant différentes dans les autres régions. Dès lors, le cas idéal pour valider cette hypothèse serait de pouvoir exploiter ces algorithmes sur un jeu de données couvrant une autre région, mais nous n'avons pas cette possibilité. Nous proposons alors une expérimentation afin d'analyser l'impact des changements entre les séries temporelles sur la projection [ACC](#).

5.3.3 Mesurer l'impact des changements sur la projection ACC

Afin d'évaluer l'impact des changements d'OCS sur la qualité de la projection ACC, nous proposons de supprimer les changements annuels en utilisant des demi-séries temporelles.

Nous répartissons les images en deux groupes en entrelaçant les images originales. La première demi-série temporelle regroupe toutes les images impaires de l'année (dans l'ordre d'acquisition), et la seconde toutes les images paires dans leur ordre d'acquisition. Ainsi, nous obtenons, pour chaque année, deux séries temporelles représentant la même OCS annuelle.

Nous appliquons deux traitements indépendants à ces demi-séries :

1. L'interpolation linéaire simple, afin de corriger les absences de valeurs causées par les nuages.
2. La même interpolation mais également le ré-échantillonnage temporel.

Ces deux traitements nous permettent alors de comparer la classification naïve et l'utilisation de la projection ACC. Dans cette expérimentation, nous utilisons uniquement un masque de pixels altérés puisque l'OCS ne change pas.

Durant cette expérimentation nous simulons deux conditions différentes d'utilisation de la régularisation temporelle. La première simule une donnée fournie amenant à un faible taux d'interpolation : les dates

souhaitées pour la série temporelle résultante correspondent à celles disponibles dans la première demi-série temporelle (numéro d’acquisition impair). La seconde demi-série, subit une interpolation plus forte car aucune des dates la composant ne correspond à celles demandées. Les effets varient en fonction des années puisque chacune possède sa propre répartition d’images sur l’année étudiée (cf 2.2).

À partir de ce jeu de données, nous ne considérons plus toutes les combinaisons possibles entre paires d’années, mais uniquement les demi-séries d’une même année, chacune représentant à tour de rôle le domaine Source ou Cible. Grâce à ce jeu de données, nous analysons plusieurs aspects :

- l’apport de l’ACC dans le problème d’adaptation de domaine
- la cohérence de cette projection, en réalisant une classification supervisée du domaine Source projeté
- l’impact du ré-échantillonnage temporel sur l’estimation de l’ACC

La table 5.1 présente les performances globales obtenues dans les cinq cas d’utilisations possibles des 14 demi-séries temporelles. Dans cette table, chaque ligne correspond à une demi-série temporelle (1 pour les dates impaires et 2 pour les dates paires). Les six colonnes associées correspondent à la PG obtenue par le classifieur utilisant les échantillons de différentes origines. Le mot clef “RT” correspond à l’utilisation du Ré-échantillonnage Temporel (notre cas typique d’utilisation des séries temporelles). Les mots clefs “sup” et “naïf” indiquent que l’apprentissage exploite les échantillons provenant de la série temporelle correspondante ou de sa complémentaire. L’acronyme ACC indique l’utilisation de la projection des données avant l’apprentissage.

La première colonne regroupe donc les performances obtenues en classification supervisée pour chaque demi série. Nous pouvons noter ici, l’impact de la séparation des primitives. Bien que nous ne puissions pas comparer directement ces performances aux résultats précédents, cette colonne met en évidence une différence minime de PG (< 0.01) dans les cas d’un équilibre relatif lors de la répartition des dates (voir 2.2). Cette différence est légèrement plus importante pour les années souffrant d’absence de données étendues, néanmoins ces écarts restent minimes. L’utilisation de la classification naïve sur les demi-séries après ré-échantillonnage temporel met en avant une légère baisse de performances, explicable par les taux d’interpolations différents, ainsi que la différence des dates composant les deux séries temporelles originales.

L’intérêt majeur de cette expérimentation demeure l’apport de l’ACC lorsque les changements d’OCS sont limités. Les quatre colonnes restantes de la table 5.1 représentent les performances obtenues. Les colonnes $(RT+)ACC + sup$ (3^{ème} et 5^{ème}) assurent que la projection ne dégrade pas l’information contenue dans les séries obtenues. Les bonnes performances, proches de celles obtenues par les séries après ré-échantillonnage temporel (1^{ère} colonne) mettent en évidence la bonne conservation des caractéristiques de chaque classe. L’effet du ré-échantillonnage temporel s’avère minime sur les performances de la classification supervisée après projection. Les colonnes de classification naïve (4^{ème} et 6^{ème}) mettent clairement en évidence l’excellent apport de la projection ACC pour pallier les différences entre les demi-séries temporelles. L’écart moyen mesuré entre la classification supervisée et naïve reste inférieur à 2 % pour la projection des séries sans ré-échantillonnage temporel. Celui-ci s’avère plus important lors de l’utilisation du ré-échantillonnage temporel avec un écart moyen d’environ 3 %. L’analyse du FScore n’apporte aucune information complémentaire, mais elle étaye ces constats. L’essentiel des variations entre les différentes approches se situant au niveau des classes peu représentées explique le faible taux de variation obtenu.

Cette expérimentation ne nous permet pas de nous prononcer sur l’impact du ré-échantillonnage temporel sur la qualité de la projection ACC. Elle permet cependant de mettre en évidence les bonnes performances

TABLE 5.1 : PG obtenues à partir des demi-séries temporelles. “RT” indique l’utilisation du ré-échantillonnage temporel, “sup” une classification supervisée (échantillons issus de la série temporelle classée) et “naïf” une classification naïve.

ST \ Méthode	RT + sup	RT + naïf	ACC + sup	ACC + naïf	ACC + RT + sup	ACC + RT + naïf
2007 ₁	0.932 ± 0.030	0.786 ± 0.025	0.935 ± 0.020	0.915 ± 0.015	0.922 ± 0.030	0.832 ± 0.018
2007 ₂	0.927 ± 0.033	0.812 ± 0.016	0.935 ± 0.020	0.904 ± 0.017	0.918 ± 0.030	0.885 ± 0.018
2008 ₁	0.918 ± 0.036	0.810 ± 0.030	0.906 ± 0.043	0.850 ± 0.037	0.896 ± 0.045	0.836 ± 0.022
2008 ₂	0.897 ± 0.046	0.758 ± 0.043	0.900 ± 0.044	0.833 ± 0.048	0.908 ± 0.038	0.806 ± 0.050
2009 ₁	0.865 ± 0.044	0.808 ± 0.033	0.899 ± 0.029	0.871 ± 0.027	0.893 ± 0.029	0.868 ± 0.028
2009 ₂	0.883 ± 0.037	0.813 ± 0.047	0.889 ± 0.031	0.874 ± 0.027	0.886 ± 0.036	0.871 ± 0.029
2010 ₁	0.898 ± 0.032	0.747 ± 0.032	0.900 ± 0.036	0.872 ± 0.029	0.893 ± 0.041	0.859 ± 0.031
2010 ₂	0.890 ± 0.037	0.830 ± 0.021	0.918 ± 0.026	0.897 ± 0.021	0.903 ± 0.038	0.891 ± 0.024
2011 ₁	0.928 ± 0.027	0.771 ± 0.007	0.918 ± 0.027	0.880 ± 0.015	0.920 ± 0.026	0.872 ± 0.013
2011 ₂	0.919 ± 0.026	0.784 ± 0.020	0.917 ± 0.026	0.888 ± 0.018	0.918 ± 0.026	0.875 ± 0.015
2012 ₁	0.916 ± 0.033	0.823 ± 0.016	0.928 ± 0.028	0.908 ± 0.023	0.926 ± 0.026	0.903 ± 0.015
2012 ₂	0.924 ± 0.028	0.796 ± 0.016	0.930 ± 0.026	0.899 ± 0.015	0.936 ± 0.024	0.878 ± 0.011
2013 ₁	0.931 ± 0.022	0.900 ± 0.008	0.945 ± 0.016	0.933 ± 0.012	0.932 ± 0.019	0.914 ± 0.014
2013 ₂	0.935 ± 0.020	0.884 ± 0.016	0.941 ± 0.018	0.929 ± 0.013	0.930 ± 0.022	0.907 ± 0.014

d’une adaptation de données via l’ACC. Cependant, ces excellentes performances requièrent l’identification et l’élimination de tout changement d’OCS susceptible de perturber l’estimation de la corrélation. Il faut noter néanmoins que cette expérimentation conforte l’hypothèse selon laquelle la pertinence de l’ACC dépend du taux de changement entre les domaines considérés. Cette contrainte ne nous permet pas d’exploiter l’ACC comme approche d’adaptation de domaine dans notre contexte d’utilisation. Néanmoins il existe des approches de détection de changements, exploitant les fondements de l’ACC.

5.4 Une approche de détection de changements

Nous venons de tester l’ACC comme approche d’adaptation de domaine. Cependant, les résultats ne nous permettent pas d’offrir une solution fiable et stable, hormis dans des conditions très favorables. Cependant, l’ACC constitue une première étape d’un algorithme de détection de changements le [Multivariate Alteration Detector \(MAD\)](#) proposé par [\[Nielsen and Conradsen, 1997\]](#).

Le principe de cet algorithme repose sur un constat simple. Lorsque nous disposons de deux images multi-variées \mathbf{X} et \mathbf{Y} , de même dimension et normalisées, une détection de changements consiste à calculer le vecteur de changement :

$$\mathbf{X} - \mathbf{Y} = [X_1 - Y_1, \dots, X_d - Y_d],$$

où d représente le nombre de primitives dans chaque image. Dans le vecteur de changement, une valeur proche de 0 signifie une absence ou un faible changement et une valeur de 1 un changement brutal. Pour assurer la cohérence dimension à dimension d’un tel vecteur, les images doivent posséder les mêmes canaux. Généralement, le vecteur de changement est rapidement dépassé par le nombre de dimension, car il devient trop complexe à analyser. C’est pourquoi, dans ce cas, on utilise généralement un algorithme de réduction

de dimension (par exemple l'ACP) qui présente le risque de perdre une partie de l'information.

C'est dans cette optique que [Nielsen and Conradsen, 1997] proposent l'utilisation de l'ACC pour améliorer les performances du vecteur de changement et condenser l'information des changements sur les différentes variables canoniques.

5.4.1 Principe du MAD

Replaçons-nous dans les conditions présentées dans la section 5.2. Soient \mathbf{X} et \mathbf{Y} deux variables multi-variées :

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \text{ et } \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix}.$$

Où \mathbf{X} et \mathbf{Y} sont normalisées afin que la moyenne soit égale à 0.

On définit alors la transformation MAD par :

$$MAD := \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \rightarrow \begin{bmatrix} a_p^T \mathbf{X} - b_p^T \mathbf{Y} \\ \vdots \\ a_1^T \mathbf{X} - b_1^T \mathbf{Y} \end{bmatrix}.$$

Où a_i et b_i représentent les coefficients obtenus par l'ACC. La transformation MAD est donnée par les variables obtenues en soustrayant les variables canoniques, en inversant leur ordre [Nielsen et al., 1998]. La première variable du vecteur MAD obtenu correspond donc aux changements maximums, puisque la corrélation est minimale et la variance maximale. Chaque dimension du vecteur de MAD permet d'extraire une carte de changements différente des i précédentes grâce aux contraintes sur la corrélation strictement positive et l'orthogonalité mutuelle des vecteurs canoniques.

Ainsi le MAD permet d'extraire une carte de changement par dimension du vecteur $\mathbf{X} - \mathbf{Y}$ dans laquelle l'information de changements détectée diffère de toutes les autres.

Cependant, l'analyse d'un tel vecteur ne permet pas, de manière simple, de déterminer les changements détectés. Pour faciliter l'analyse on calcule une probabilité de changement. On obtient une telle probabilité, pour un pixel j , grâce à une approximation au moyen d'une distribution χ^2 à p degrés de liberté :

$$T_j = \sum_{i=1}^p \left(\frac{MAD_{ij}}{\sigma_{MAD_i}} \right)^2 \in \chi^2(p).$$

Cette approximation permet de calculer un contraste. En effet, on calcule le niveau de différence normalisé par l'écart-type, associé à une norme L_2 (somme des carrés). Comme ces variables sont indépendantes et orthogonales, elles suivent une loi du χ^2 . Cependant, interpréter cette approximation comme des changements, suppose que le MAD est bien estimé et donc que les changements existants n'ont pas influencé l'estimation de l'ACC. Ainsi, pour satisfaire cette contrainte, ces changements doivent être minoritaires.

5.4.2 Expérimentations

La validation d'une carte de changements représente un défi, à défaut de posséder des données de référence exhaustives. Dès lors, pour obtenir une mesure suffisante, la seule option disponible consiste à réaliser des

classifications supervisées des deux années évaluées, afin de réaliser une référence de changements. Les résultats obtenus dans ces expérimentations étant très proches pour toutes les paires d'années, nous choisissons de présenter les résultats des années 2009, 2010 et 2011.

Détection de changements entre séries temporelles pour la prédiction de cartes d'OCS

Le MAD reposant en grande partie sur l'ACC, nous exploitons par conséquent certains protocoles utilisés lors de l'étude de l'ACC afin de mener ces expérimentations. L'utilisation d'un masque lors de l'estimation de l'ACC impacte la projection obtenue, nous utilisons donc les masques de classes annuelles (A) et de pixels altérés (P) défini précédemment.

Pour simplifier la présentation des résultats, nous fixons 2009 en tant que domaine Source pour produire les cartes de 2010 et 2011. Nous nous plaçons dans une perspective de report d'étiquettes du domaine Source vers le domaine Cible lorsque le MAD indique un non-changement.

La procédure se déroule en 4 étapes :

1. Classification supervisée du domaine Source à partir des séries temporelles (avec ré-échantillonnage temporel)
2. Classification naïve du domaine Cible à partir du classifieur de l'étape 1
3. MAD à partir des séries temporelles sans ré-échantillonnage temporel
4. Pour chaque pixel on sélectionne l'étiquette provenant de la classification :
 - (a) Naïve si le MAD indique un changement
 - (b) Supervisée dans le cas contraire

Une fois la carte produite, nous la validons à partir des données de référence, qui sont disponibles pour cette étape. Cependant, cela ne nous permet pas d'évaluer le pourcentage de bonnes décisions dues aux MAD de celles de la classification naïve.

Nous évaluons la qualité du report d'étiquettes issu de la classification supervisée en appliquant une valeur hors nomenclature aux échantillons détectés comme changements. Cette astuce permet d'écarter ces pixels lors de la validation et ainsi mesurer précisément l'impact du MAD sur la classification.

La figure 5.7 présente les performances obtenues lors de cette expérimentation. Dans cette matrice, la première colonne représente la PG obtenue. Les 17 colonnes restantes représentent le FScore des classes de la nomenclature. Nous différencions l'usage du masque de pixels altérés (P) et de classes annuelles (A) afin d'évaluer son effet en comparaison avec les observations réalisées lors de la projection ACC. Les lignes impaires (en commençant du haut de la matrice) indiquent les performances du report seul des étiquettes du domaine Source en cas de non-changements détectés par le MAD (MAD en ordonnées), les lignes paires montrent les performances obtenues lors de l'utilisation combinée du MAD et de la classification naïve.

Nous pouvons observer que l'utilisation du masque de classes annuelles lors de l'estimation du MAD constitue un réel problème, puisque à la différence de l'ACC, il empêche systématiquement la détection de changements. Nous ne retrouvons pas l'effet alternatif observé lors de l'analyse des performances de l'ACC lorsque le pas de temps entre les années est pair. Cet effet met en évidence les différences entre l'ACC et le MAD, un effet bénéfique sur l'ACC ne se répercutant pas sur les performances de la détection de changements,

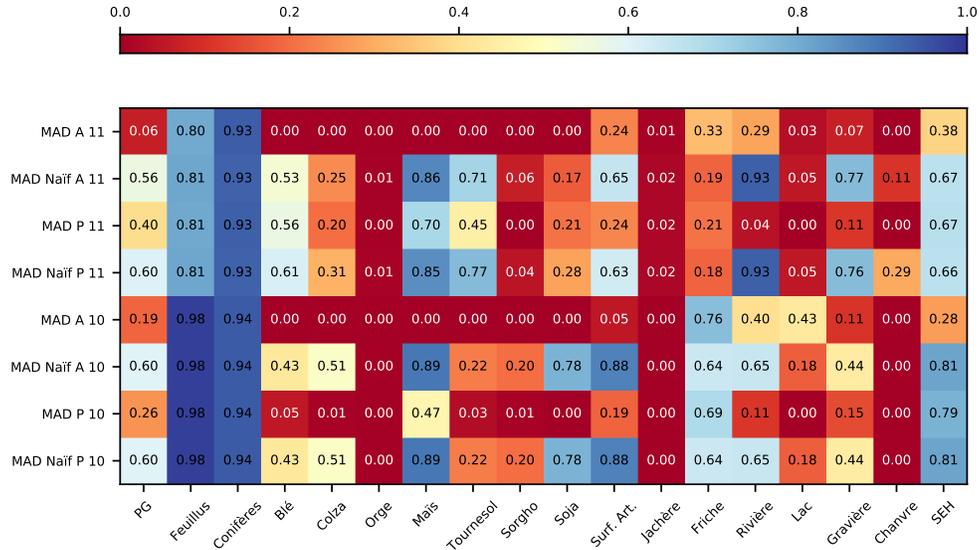


FIGURE 5.7 : PG et FScore obtenus par la détection de changements seule (MAD) ainsi que par son utilisation combinée avec la classification naïve (MAD Naïf), avec l’utilisation du masque de classes annuelles (A) et de pixels altérés (P), pour les domaines Cible 2010 et 2011 à partir du domaine Source 2009.

il s’agit par conséquent de deux approches distinctes qui nécessitent une amélioration indépendante pour une utilisation opérationnelle.

Ce constat nous amène à concentrer l’analyse sur les seules performances du masque de pixels altérés. La PG ne fournit qu’un indicateur relativement faible. En effet, dans le cas du seul report, nous comptabilisons l’absence de label existant comme étant une erreur. Ce choix permet de quantifier, via les métriques standards la validité de la détection de changements réalisée. Nous pouvons voir que pour les entrées “MAD P 11” et “MAD P 10”, les performances demeurent faibles avec 40 % et 26 % de PG respectivement. Cependant, le FScore met en évidence des performances par classe très variées notamment sur les classes de cultures. Le cas “MAD P 10” occulte totalement les classes de cultures, à l’inverse du cas “MAD P 11” montrant ici un comportement similaire à l’ACC, appuyant l’hypothèse selon laquelle ces algorithmes s’avèrent plus efficaces lorsque l’OCS est la même dans les deux domaines.

Nous pouvons voir que le MAD considère les trois classes d’eau en constant changement. En effet ces trois classes n’apparaissent qu’en très faibles quantités lors de l’utilisation du MAD seul. Pour les autres classes géographiquement stables, le MAD offre une prédiction satisfaisante.

L’approche de classification mêlant classification naïve et MAD présente de moins bonnes performances que la classification naïve seule (cf table 3.1), avec une PG constante à 60 %. L’usage de la classification naïve permet de corriger efficacement les problèmes du MAD. En effet, certaines classes toujours étiquetées “changements” par le MAD, sont en contrepartie bien classées par la classification naïve (Lacs, Rivières, Maïs).

Les performances moindres de cette approche de classification s’expliquent par les erreurs de non-détection commises par le MAD. En effet, dans ce cas l’étiquette sélectionnée provient du domaine Source, et peut représenter une source d’erreur sur laquelle nous ne pouvons pas intervenir. Une perspective intéressante

serait d'évaluer une mesure permettant de pondérer la décision du MAD et de la classification naïve afin de réaliser une décision moins catégorique.

Le manque de fiabilité du MAD ne nous permet pas de le considérer comme une approche viable (du moins seul) permettant d'étiqueter le domaine Cible en vue d'entraîner un classifieur. Cependant, les performances obtenues dépendent là encore du choix de la nomenclature, car le MAD offre une bonne reconnaissance de certaines classes. En effet, il permet de détecter les changements opérant au niveau des classes géographiquement stables. Ces conclusions nous amènent à prospecter vers des méthodes plus complexes d'adaptation de domaine en réponse à notre problème. Cependant, nous réalisons deux études complémentaires afin de mieux comprendre les limitations du MAD : l'utilisation des demi-séries et la détection de changements artificiels au sein d'une série temporelle.

Détection de changements entre demi-séries

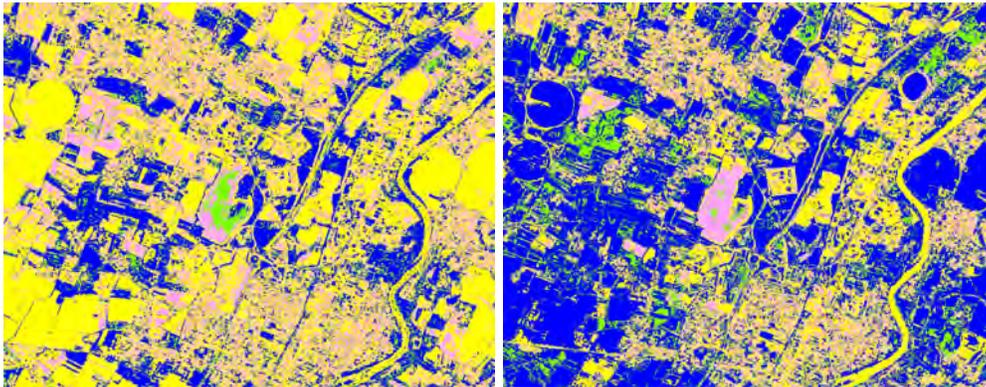
Similairement à l'approche introduite lors de l'ACC, l'idée est d'éliminer les changements d'OCS et d'évaluer la quantité de changements détectés entre les deux demi-séries. Nous continuons à comparer les deux masques présentés lors de l'estimation de l'ACC : le masque de pixels altérés et le masque de classes agricoles, qui correspondent aux classes annuelles de la nomenclature. L'idée générale consiste à montrer l'impact du masque sur l'estimation du MAD.

Les demi-séries temporelles appartiennent à la même année d'acquisition, et donc représentent la même OCS annuelle. Ainsi nous attendons que la détection de changements ne retourne aucun (ou très peu de) cas de changements. La figure 5.8 présente une illustration de la carte de changements obtenue dans les deux cas d'utilisation du masque pour une détection de changements entre les deux demi-séries de l'année 2009. Nous introduisons un code couleurs afin de comprendre les différences entre le résultat du MAD et la référence. La référence est calculée par simple différence entre la classification supervisée de chaque demi-série. Un accord de changements entre le MAD et la référence apparaît en rose, un changement détecté par le MAD uniquement apparaît en jaune, un changement non détecté apparaît en rouge et un accord de non-changement apparaît en bleu. Pour des raisons de visualisation nous concentrons cette figure sur une zone d'intérêt couvrant les différents milieux de la zone d'étude : zones agricole (champs de Maïs circulaire), Rivières (enfin un fleuve en l'occurrence : la Garonne) et des zones urbaines et naturelles (Feuillus).

Alors que l'utilisation du masque de classes annuelles peut impacter fortement les performances de l'ACC, nous notons qu'il perturbe la prédiction du MAD en signalant les classes de cultures comme changements (en jaune à gauche et en bleu à droite). Notons également le cas typique des classes d'eau que le MAD détecte systématiquement comme changements.

La table 5.2 quantifie les proportions de changements détectés par le MAD et par la référence. Cette table regroupe les performances obtenues pour l'année 2009, car toutes les combinaisons d'années offrent des résultats similaires pour ce cas d'étude. Dans cette table, nous retrouvons un taux de changements détectés dans la référence directement corrélé avec sa PG (80 %). Le résultat intéressant concerne l'utilisation du masque de classes annuelles qui inverse littéralement le taux de changements et de non-changements entre les deux cartes produites. Le rendu visuel montre néanmoins qu'il ne s'agit pas d'une simple inversion d'étiquettes entre les deux méthodes (pixels en jaune dans les deux images de la figure 5.8), mais bien d'une détection différente.

Si ces résultats nous renseignent sur le taux de pixels faussement détectés, la table 5.3 permet une meilleure compréhension du comportement du MAD, en indiquant les proportions entre les quatre cas que



■ Accord Changements
 ■ Changement MAD
 ■ Changement Ref
 ■ Accord Non Changements

FIGURE 5.8 : Détection de changements entre demi-série de l'année 2009

TABLE 5.2 : Proportions de changements détectés par le MAD entre les demi-séries de 2009

	Changements	Non-Changements
Référence	0.23	0.77
Pixels Altérés	0.31	0.69
Classes annuelles	0.69	0.31

nous rencontrons :

1. La référence et le MAD sont d'accord sur le changement (Accord chgts)
2. Le MAD détecte un changement mais pas la référence (MAD chgts)
3. La référence indique un changement mais pas le MAD (Ref chgts)
4. La référence et le MAD sont d'accord sur un non-changement (Accord Non Chgts)

Ces proportions représentent un cas d'analyse intéressant, car si nous pouvons supposer que les changements détectés dans la référence proviennent essentiellement des pixels de bordure, ou fortement interpolés par le ré-échantillonnage temporel, nous attendons que ces mêmes pixels soient problématiques pour le MAD. Or, le MAD détecte plus de changements qu'il ne devrait, et se positionne souvent en désaccord franc avec la référence.

TABLE 5.3 : Cohérences des changements détectés par le MAD et la référence pour 2009

	Accord chgts	MAD chgts	Ref chgts	Accord Non Chgts
Pixels Altérés	0.10	0.22	0.13	0.56
Classes annuelles	0.17	0.52	0.05	0.26

Similairement à l'approche de projection ACC, les changements d'OCS trop importants semblent perturber les performances du MAD. Or, il n'y a pas de changements, puisqu'il s'agit de la même année. Ainsi, nous

pouvons affirmer que les différents stades de l'évolution des classes au cours de l'année sont détectés comme changements. Nous allons donc essayer de déterminer si ces performances sont dues aux différences entre les séries temporelles, ou si le MAD ne constitue pas un algorithme performant pour traiter les changements d'OCS.

Détection de changements artificiels

Pour évaluer la résistance du MAD face aux changements d'OCS au cours d'une année, nous essayons de réduire au maximum les facteurs pouvant impacter ces performances. Pour cela, nous introduisons du bruit dans une des deux demi-séries temporelles. Nous cherchons alors à détecter ces changements entre la série bruitée (ST_b) et la série originale (ST_1).

Afin de ne pas faciliter la tâche du MAD, nous introduisons un bruit similaire à un changement d'OCS, dans différentes proportions de 1 %, 5 %, 10 % jusqu'à 50 % (par pas de 10 %) de pixels bruités. Par exemple, nous sélectionnons aléatoirement 40 % des échantillons présents dans l'emprise de l'image (hors bords), 20 % de ces échantillons remplacent les valeurs des 20 % restant. Ainsi les profils de bruits introduits sont des classes existantes, appartenant au même domaine. Nous utilisons les positions des pixels bruités afin de créer une carte de validation indiquant si le pixel a été modifié.

Nous utilisons le MAD afin de détecter les changements et exploitons directement la carte de probabilité dérivée de la distribution χ^2 pour réaliser une carte binaire dans laquelle les étiquettes sont changements (1) ou non changements (0). Nous comparons cette carte binaire à la carte de validation puis nous validons cette comparaison en comptabilisant l'intégralité des valeurs non nulles. Dans cette expérience le taux de détection de changements plafonne à 100 % de détections valides, l'algorithme détecte toutes les permutations que nous réalisons. Ces résultats montrent que le MAD peut, dans des conditions idéales détecter un changement de classe d'OCS dans un profil temporel.

Ce cas d'utilisation constitue un cas très facile de détection de changements dans lequel une simple différence entre les images donne un résultat similaire. Pour complexifier cette situation, nous calculons le MAD entre les deux demi séries originale ST_1 et ST_2 , puis le MAD entre la série bruitée ST_b et ST_2 . Cette expérimentation vise à détecter les changements introduits artificiellement dans la ST_1 . L'objectif consiste alors à déterminer si l'algorithme détecte uniquement les changements artificiels, ou si ceux-ci ne seront pas détectés au détriment des changements analysés dans la section précédente.

TABLE 5.4 : Proportion de changements (chgts) artificiels (art) détectés par le MAD.

Proportion de bruit (%)	Non Chgts	Chgts	Chgts Art.	ratio Art. (%)
1	6 673 636	2 665 953	92 663	99.2
5	6 880 764	2 458 825	457 262	97.9
10	6 828 743	2 510 846	895 212	95.9
20	6 515 395	2 824 194	1 673 824	89.6
30	6 220 585	3 119 004	2 263 489	80.8
40	6 038 450	3 301 139	2 655 952	71.1
50	5 958 229	3 381 360	2 898 918	62.1

La table 5.4 recense pour les différents niveaux de bruits artificiels introduits, le nombre total de pixels étiquetés non-changements par le MAD (Non Chgts), ceux détectés comme changements et parmi ceux-ci le nombre de pixels changés artificiellement (Chgts Art.). La dernière colonne indique le ratio de pixels

artificiellement changés correctement détectés par le MAD. Alors que lors de la détection entre ST_b et ST_1 les seuls changements détectés correspondaient à ceux ajoutés artificiellement, nous retrouvons un effet de sur-détection systématique de la part du MAD. Lorsque le bruit augmente, le taux de détection du bruit artificiel diminue pour un nombre de changements détectés en constante augmentation.

En conclusion, le MAD ne fait pas la différence entre une distorsion entre deux séries temporelles annuelles représentant la même scène et des changements d'OCS. L'augmentation des permutations impacte le nombre de changements détectés, ce qui diminue l'intérêt du MAD pour la détection de changements entre séries temporelles.

Dans ce chapitre, nous abordons une technique avancée d'adaptation de données : le **Transport Optimal (TO)**. Cette approche d'adaptation de données, a intéressé les mathématiciens et industriels pendant de longues années. Après une introduction portant sur l'évolution du TO et son fonctionnement, nous présenterons les aspects permettant de l'exploiter comme méthode d'adaptation de domaine. Nous mettrons en avant, l'importance de la stratégie de sélection d'échantillons pour l'estimation du TO, à travers différentes propositions.

6.1 Introduction

6.1.1 Le Transport Optimal à travers les siècles

Le **Transport Optimal (TO)** ou la “brouette de Monge” est une théorie élaborée par le mathématicien du même nom au cours du XVIII^e siècle. Cette innovation, portée par des applications militaires, cherche une solution au transport de volumes de terre (ou de sable) d'un point de départ à un d'arrivée, selon une méthode rigoureuse ; en d'autre terme : optimale. Au cours de son “Mémoire sur la Théorie des Déblais et des Remblais”¹, Gaspard Monge livre un des premiers essais historiques de traitement systématique et efficace des problèmes combinatoires.

Dans son mémoire, Monge définit la motivation aux fondements du TO comme suit :

“Lorsqu'on doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de Déblai au volume des terres que l'on doit transporter, et le nom de Remblai à l'espace qu'elles doivent occuper après le transport. Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids et à l'espace qu'on lui fait parcourir, et par conséquent le produit du transport total devant être proportionnel à la somme des produits des molécules multipliées par l'espace parcouru², il s'ensuit que le

¹<https://gallica.bnf.fr/ark:/12148/bpt6k35800/f796.image>

²Le produit du poids des molécules et de l'espace parcouru.

déblai et le remblai étant donnés de figure et de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine distribution à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits sera la moindre possible, et le prix du transport total sera un minimum."

Il achève sa préface par une affirmation qu'il contredira lui-même au terme de son mémoire : "C'est la solution de cette question que je me propose de donner ici" puisqu'il admettra que le problème formulé et les contraintes associées sont trop complexes pour offrir une solution. Il n'en reste pas moins le précurseur d'une théorie toujours d'actualité.

La formulation de Monge fut reprise par un autre mathématicien : Leonid Kantorovich dans les années 1940 comme un problème d'allocation optimale des ressources. Là encore, la visée militaire conduit l'avancée scientifique. En effet, ses travaux l'ont occupé avant et après la seconde guerre mondiale, durant laquelle il a pu tester en conditions réelles certains points de sa théorie, en assurant une distribution optimale des ressources en particulier durant le siège de Leningrad. Ses travaux ont également contribué au développement de l'optimisation moderne, et donc eu un impact énorme dans de très nombreux domaines appliqués.

De très nombreux scientifiques de tous horizons ont participé au développement des théories actuelles du TO, qui touche de nombreux domaines tels que l'économie, l'analyse des langages, la sociologie, la mécanique des fluides [Benamou and Brenier, 2000] la vision par ordinateur, etc. Son domaine d'application principal fut pendant longtemps l'industrie : dès que nous cherchons à mettre en relation des quantités de marchandises produites avec des quantités attendues, on cherche à résoudre un problème de TO. Dans ce manuscrit nous nous intéressons exclusivement à la formulation dite de **Monge-Kantorovitch**.

6.1.2 Formulation de Monge-Kantorovitch

Prenons un problème simple afin d'expliquer avec une illustration la formulation de Monge, puis les améliorations apportées par Kantorovitch permettant une résolution de nombreux problèmes de la théorie du TO. Le crédit pour l'exemple utilisé, ainsi que les illustrations revient à Gabriel Peyré et sont issus de deux articles [Peyré, 2019a, Peyré, 2019b]. L'objectif de cette section consiste à introduire des concepts peu exploités en télédétection.

Un problème de transport de marchandises : la brouette de Monge

Au fil de cet exemple, nous allons nous intéresser au problème d'approvisionnement de n cafés (population y , en bleu dans la figure 6.1) parisiens en croissants par n boulangeries (population x , en rouge). La formulation de Monge impose qu'il y ait autant de points de départ que d'arrivée. En vue de simplifier la formulation du problème et sa représentation, seuls 6 boulangeries et 6 cafés sont considérés. Nous supposons également que chaque boulangerie produit la quantité m de croissants attendue par les cafés. Ainsi chaque boulangerie ne peut desservir qu'un seul café, et celui-ci ne peut être approvisionné que par une boulangerie. Posons $i \in \{1, \dots, 6\}$ associé au nombre de boulangeries et $j \in \{1, \dots, 6\}$ associé aux cafés. On peut alors exprimer l'ensemble des permutations S_n comme étant une association entre un élément x_i et y_j :

$$s : i \in \{1, \dots, 6\} \mapsto j \in \{1, \dots, 6\},$$

où s représente une permutation. Avec ces hypothèses nous cherchons donc les paires de cafés et boulangeries pour lesquelles l'effort de livraison (le coût) est minimal et respectant S_n . Dans cet exemple, le coût est

représenté par le temps de trajet entre une boulangerie et un café. La première étape consiste alors à recenser tous les coûts entre les différentes possibilités. Notons $C_{i,j}$ le temps de trajet entre la boulangerie x_i et le café x_j . L'ensemble des associations possibles est représenté sous forme matricielle dans la table 6.1.

TABLE 6.1 : Exemple de temps de trajet entre les boulangeries x_i (lignes) et les cafés y_j . Les valeurs vertes indiquent un exemple de choix associés à une permutation s .

$C_{i,j}$	y_1	y_2	y_3	y_4	y_5	y_6
x_1	12	10	31	27	10	30
x_2	22	7	25	15	11	14
x_3	19	7	19	10	15	15
x_4	10	6	21	19	14	24
x_5	15	23	14	24	31	34
x_6	35	26	16	10	34	15

Chaque ligne de cette matrice représente le trajet entre la boulangerie i et tous les cafés. Les valeurs en vert représentent un choix possible de permutation s respectant la contrainte d'approvisionnement.

Un choix de permutation est illustré par le biais de la figure 6.1. Pour ce choix, la permutation s vaut :

$$s(1) = 5, s(2) = 2, \dots, s(5) = 3, s(6) = 4.$$

À partir d'une permutation et de la matrice de coût nous pouvons calculer le coût total associé à ce choix :

$$Coût(s) = \sum_{i=1}^n C_{i,s(i)}.$$

Pour notre exemple et le choix représenté dans la table 6.1, le coût est alors obtenu par :

$$Coût(s) = C_{1,5} + C_{2,2} + C_{3,6} + C_{4,1} + C_{5,3} + C_{6,4} = 10 + 7 + 15 + 10 + 14 + 9 = 65.$$

Il faut alors résoudre le problème d'optimisation suivant :

$$\min_{s \in S_n} Coût(s).$$

Cette formulation est celle proposée par Monge. On obtient la solution optimale de ce problème en évaluant toutes les permutations s . Nous sommes donc face à un problème où le nombre de solutions possibles s'élève à $n!$. En effet, pour le premier élément x_1 il y a n choix d'affectation possible, pour x_2 il reste alors $n - 1$ permutations possibles, et ainsi de suite. Pour l'exemple ci-dessus, nous obtenons 720 solutions possibles. Cette quantité de solutions possibles présente la contrainte majeure de la formulation de Monge, car elle est alors limitée aux seuls cas où n est très petit.

La formulation continue de Kantorovich

Il aura fallu attendre 1940 et les travaux de Kantorovich pour assister à une nette amélioration de la généralisation du TO. L'innovation consiste à modifier le problème de Monge en remplaçant l'ensemble des permutations S_n par une matrice binaire de permutation P de taille $n \times n$. L'ensemble des $n!$ permutations

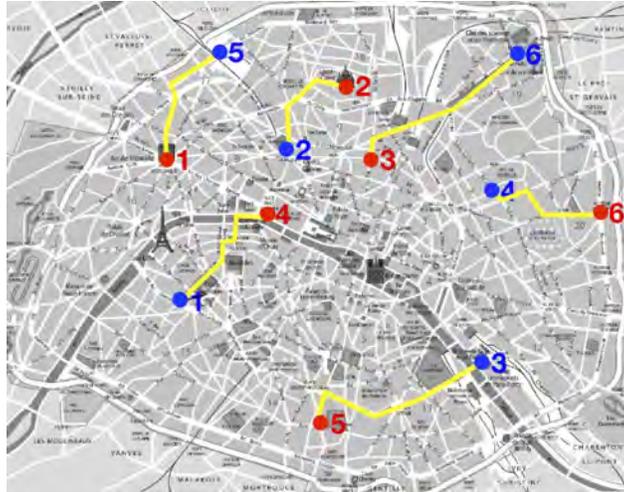


FIGURE 6.1 : Exemple de permutation respectant la contrainte d'approvisionnement.

Source : <https://images.math.cnrs.fr/Le-transport-optimal-numerique-et-ses-applications-Partie-1.html>

[//images.math.cnrs.fr/Le-transport-optimal-numerique-et-ses-applications-Partie-1.html](https://images.math.cnrs.fr/Le-transport-optimal-numerique-et-ses-applications-Partie-1.html)

S_n devient alors l'ensemble de $n!$ matrices de permutation P_n . Dans ces matrices, un 1 équivaut à une permutation et 0 à une absence de permutation. Pour le choix de permutation réalisée dans la table 6.1 la matrice P équivalente vaut :

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Dans chaque matrice, seuls n éléments sont non-nuls et égaux à 1, nous pouvons donc réécrire le problème de Monge comme étant une double somme sur les n éléments :

$$\min_{P \in P_n} \sum_{i=1}^n \sum_{j=1}^n P_{i,j} C_{i,j}.$$

À ce stade, il s'agit d'une simple réécriture du problème de Monge. L'apport de Kantorovich consiste à introduire la possibilité de passer d'une formulation discrète (n est fini et très grand) à une formulation continue donc potentiellement infinie. Les propriétés des matrices bistochastiques rendent possibles le passage à la formulation continue. Une matrice est dite bistochastique si la somme des valeurs sur chaque ligne et sur chaque colonne est égale à 1. (Elle est dite stochastique si seulement une des deux conditions est remplie.) Ce qui est obligatoirement le cas d'une matrice binaire de permutation P . On peut alors définir l'ensemble des matrices bistochastiques β_n dans lequel la contrainte $P \in \{0, 1\}^{n \times n}$ est relaxée; au lieu d'imposer des valeurs 0 ou 1 dans P , elles sont comprises entre 0 et 1. β_n est alors défini comme :

$$\beta_n := \left\{ P \in [0, 1]^{n \times n} : \forall i \sum_j P_{i,j} = 1, \forall j \sum_i P_{i,j} = 1 \right\}.$$

Cet ensemble β_n est plus grand que l'ensemble P_n et contient au moins toutes les solutions possibles de P_n , comme le prouve l'inégalité de Monge-Kantorovich [Rachev, 1985], que nous n'aborderons pas ici.

La résolution du TO selon la formulation de Kantorovich devient alors

$$\min_{P \in \beta_n} \sum_{i=1}^n \sum_{j=1}^n P_{i,j} C_{i,j}.$$

La différence par rapport à la formulation initiale consiste à l'appartenance de P à l'ensemble β_n et non plus P_n . La particularité de cette formulation par rapport à celle de Monge est que l'ensemble β_n est convexe : la moyenne de deux matrices bistochastiques est elle-même bistochastique. Ce qui n'est pas le cas pour les matrices de permutation qui possèdent des valeurs binaires. En effet la somme de deux matrices binaires n'est pas binaire, la seule exception étant que ces deux matrices soient égales. L'intérêt de la formulation de Kantorovich bénéficie du développement au même moment de l'algorithme du simplexe [Dantzig et al., 1955] qui représente une référence pour la résolution de problèmes d'optimisation convexe linéaire. Dans le cadre de la formulation de Kantorovich, l'algorithme du simplexe correspond à une complexité de l'ordre n^3 , beaucoup plus performant que le $n!$ de la formulation de Monge.

Le transport optimal pondéré

La formulation initiale du problème du TO s'avère contraignante par le postulat de départ, avec le même nombre de points de départs et d'arrivées. Si ce nombre diffère, alors le problème demeure insoluble. En effet, la permutation exclusive (bijection) entre deux espaces de tailles différentes reste impossible. On lève cette contrainte en ne considérant pas les objets (terre, cafés ou boulangeries) en tant que masses à transporter, mais plutôt les distributions (a_1, \dots, a_n) de production (associées aux boulangeries dans notre exemple) et les distributions (b_1, \dots, b_m) de consommation (les cafés), où n et m sont le nombre d'instances de production et de consommation respectivement. La principale contrainte imposée dans la formulation initiale, où $n = m$ les quantités a_i et b_j sont obligatoirement égales, est levée. Néanmoins, ces quantités doivent être positives et respecter l'égalité totale des masses :

$$a_1 + a_2 + \dots + a_{n-1} + a_n = b_1 + b_2 + \dots + b_{n-1} + b_n.$$

On généralise la formulation de Kantorovich en utilisant des matrices de couplage satisfaisant la contrainte de conservation des masses, plutôt que des matrices bistochastiques :

$$\beta(a, b) := \left\{ P \in \mathbb{R}_+^{n \times m} : \forall i \sum_j P_{i,j} = a_i, \forall j \sum_i P_{i,j} = b_j \right\}.$$

Dans la matrice de couplage P , chaque valeur non nulle implique un transfert de masse. Contrairement au cas des matrices bistochastiques, pour lesquelles l'unique solution consiste en une permutation (l'intégralité de la masse a_i associé à x_i est transporté au point y_j), la solution d'une matrice de couplage peut être multiple. Ainsi, une unité de production peut approvisionner une ou plusieurs unités de consommation.

La formulation généralisée du problème de Kantorovich devient alors :

$$\min_{P \in \beta(a,b)} \sum_{i=1}^n \sum_{j=1}^m P_{i,j} C_{i,j}.$$

La différence est que nous associons un coût à chaque transfert d’une unité de masse entre i et j . Cette formulation convexe permet donc une exploitation efficace de l’algorithme du simplexe.

L’écriture de la formulation de Kantorovitch a ouvert la voie à l’application du TO dans de nombreux domaines applicatifs. Dès que nous pouvons représenter les données par des distributions de probabilités (et que le problème le requiert) le TO devient exploitable. Il a fallu patienter 150 ans pour que le problème de Monge soit résolu par Kantorovitch. Puis le problème a de nouveau été délaissé par les mathématiciens avant de revenir en force dans les années 2000 [Villani, 2008]. La théorie présentée ici illustre les prémices de nombreuses améliorations et variations du problème de Monge-Kantorovitch. Les innovations majeures de ces dernières années portent sur des méthodes de résolutions efficaces des problèmes convexes et des formulations dédiées à des applications spécifiques notamment sur la notion de coût. La prochaine section présentera les formulations dédiées à l’adaptation de domaine et exploitées dans cette thèse.

6.2 Transport Optimal et adaptation de domaine

Nous venons d’introduire les bases de la théorie du TO. Dans cette section nous présenterons la formulation du TO pour l’adaptation de domaine tout en réalisant un parallèle entre les travaux existants et notre cas d’étude.

Pour l’adaptation de domaine, les objets à transporter n’ont pas de connotations matérielles, comme une usine de production ou un bien manufacturé. Il s’agit, dans notre cas, de données représentant une observation des sols par télédétection. Le TO s’inscrit dans les méthodes d’adaptation de données : notre objectif consiste à réduire au maximum l’effet des distorsions entre les domaines D_S et D_C tout en conservant les spécificités des différentes classes observées. Pour rappel, D_S est l’ensemble des échantillons d’une année pour laquelle les données de référence sont disponibles alors que D_C ne contient que les échantillons.

6.2.1 La formulation discrète

Dans le cadre de cette thèse nous avons utilisé la formulation discrète du TO pour l’adaptation de domaine proposée par Courty et al. [Courty et al., 2014]. Une librairie python *Python Optimal Transport (POT)* [Flamary and Courty, 2017] permet d’accéder à cette variante.

Posons les notations utilisées dans cette formulation. Soit $D_S = \{x_i^s, y_i^s\}_{i=1}^{n_s}$ où $x_i^s \in \mathbb{R}^d$ est un échantillon de la série temporelle du domaine Source et $y_i \in \mathcal{C}$ son étiquette associée. \mathcal{C} représente les valeurs possibles d’étiquettes. Notons $D_C = \{x_i^c\}_{i=1}^{n_c}$ le domaine Cible où seuls les échantillons sont disponibles. Les variables n_s et n_c correspondent respectivement au nombre d’échantillons dans le domaine Source et Cible.

La première étape consiste à passer d’un jeu de données multi-variées à des probabilités a priori de chaque classe. Cette étape pose un obstacle majeur dans la plupart des cas d’applications réels, du fait de la grande complexité à estimer cette distribution pour le domaine Source. Il a été montré que le choix de la distribution impacte fortement le résultat du TO [Tuia et al., 2015]. En conclusion, nous estimons préférable de recourir à un équilibrage en amont des classes impliquées dans le transport afin de pouvoir exploiter une distribution uniforme pour tous les échantillons. Ainsi ceux-ci possèdent une masse de probabilité équivalente. Notons μ la distribution empirique associée à chaque domaine, calculée par :

$$\mu_s = \sum_{i=1}^{n_s} p_i^s \delta_{x_i^s}, \mu_c = \sum_{i=1}^{n_c} p_i^c \delta_{x_i^c}.$$

Avec :

- p_i est la masse de probabilités associée au $i^{\text{ème}}$ échantillon.
- δ_{x_i} la distribution de Dirac à la position $x_i \in \mathbb{R}^d$, qui permet ici de discrétiser la fonction de masse.

Les masses de probabilités respectent la loi d'égalité des masses de Monge : $\sum_{i=1}^{n_s} p_i^s = \sum_{i=1}^{n_c} p_i^c = 1$. Dans la mesure où nous utilisons une distribution uniforme on obtient $p_i^s = \frac{1}{n_s}$ et $p_i^c = \frac{1}{n_c}$.

La figure 6.2 illustre un problème factice de TO. Trois classes, générées aléatoirement se composent de 10 échantillons chacune. Les deux domaines sont constitués à partir de deux distributions gaussiennes disposant de paramètres légèrement différents. Le plan de représentation correspond aux deux composantes, les trois couleurs représentent les classes. Les distorsions entre les deux domaines apparaissent, nettement visibles. Dans cet exemple, nous connaissons les étiquettes de tous les échantillons, mais elles ne sont utilisées qu'à des fins de visualisation.

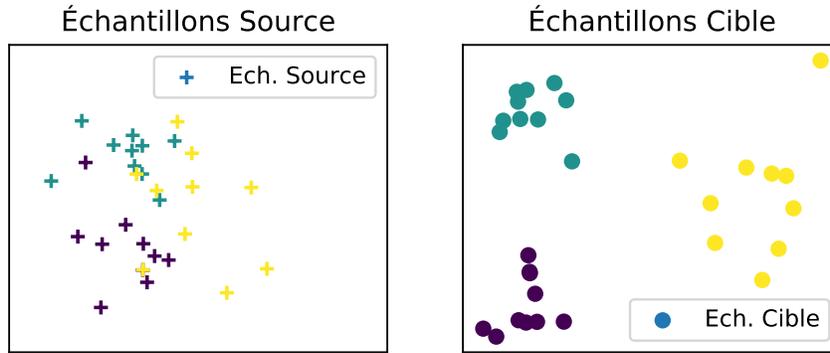


FIGURE 6.2 : Échantillons, à deux dimensions, du domaine Source et Cible générés aléatoirement selon des distributions gaussiennes. Trois classes sont représentées par des couleurs différentes.

Chaque échantillon possède donc une masse équiprobable associée pour des raisons de simplicité. Il convient alors de calculer la matrice de coût \mathbf{C} entre les échantillons des deux domaines.

Au cours de nos expérimentations, nous utilisons la distance euclidienne au carré comme mesure de coût. Ce coût représente la norme de l'écart primitive par primitive entre deux échantillons de la série temporelle :

$$C(i, j) = \|x_i^s - x_j^c\|_2^2.$$

La figure 6.3 présente la matrice de coût correspondant à notre exemple. Dans cette figure, chaque carré représente un coût variant entre une valeur faible, représentée en bleu, ou élevée en jaune. À partir de cette représentation, le couplage optimal entre les échantillons Source (en ligne) et Cible (en colonne) ne peut pas être déterminé.

Le problème de TO que nous souhaitons résoudre est alors de trouver une transformation $\mathbf{T} : D_S \rightarrow D_C$ qui met en correspondance les deux domaines, en réduisant les distorsions possibles. Il est attendu que la transformation \mathbf{T} conserve les informations associées aux étiquettes, en d'autres termes il doit préserver la distribution conditionnelle $P_c(y|x^c) = P_s(y|T(x^s))$. Ainsi, nous ne conserverons que les solutions satisfaisant

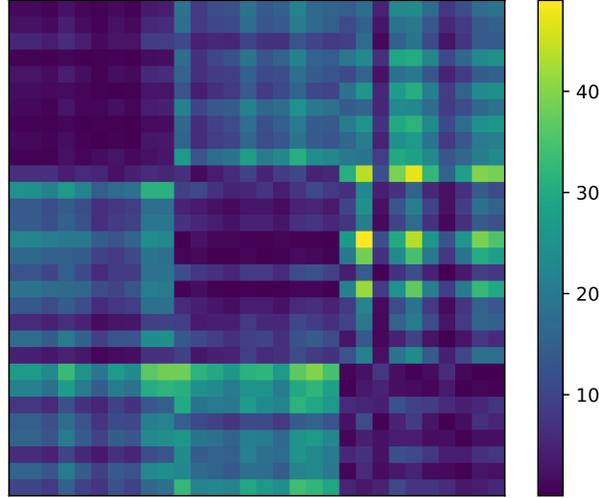


FIGURE 6.3 : Matrice de coût associée au jeu de données test.

cette contrainte. L'ensemble des solutions à ce problème est composé de fonctions de couplage γ entre les densités de probabilités des deux domaines :

$$\beta = \{\gamma \in (\mathbb{R}^+)^{n_s \times n_c} \mid \gamma \mathbf{1}_{n_c} = \mu_s, \gamma^T \mathbf{1}_{n_s} = \mu_c\},$$

où $\mathbf{1}_d$ est un vecteur unitaire de dimension d . Parmi toutes les solutions $\gamma \in \beta$ le transport optimal sélectionnera celle qui minimise le coût total.

La formulation de Kantorovitch devient alors :

$$\gamma_0 = \underset{\gamma \in \beta}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F.$$

où :

- \mathbf{C} est la matrice de coût
- $\langle \cdot, \cdot \rangle_F$ est le produit scalaire de Frobenius

Il s'agit de la formulation discrète de l'*Earth Mover's Distance* (EMD), que nous désignerons comme la méthode [EMD](#) dans la suite du manuscrit.

Nous avons également exploité une version régularisée de l'EMD. Cette variante permet d'éviter que des échantillons proches se dispersent durant le calcul du transport en régularisant l'entropie de la matrice de couplage $h(\gamma)$:

$$h(\gamma) = - \sum_{i,j} \gamma_{i,j} \log(\gamma_{i,j}).$$

Le calcul de l'entropie négative est également connue comme la mesure de distance Sinkhorn. Une telle mesure va augmenter le nombre de permutations possibles dans la matrice de couplage et va privilégier la solution possédant la plus haute entropie : celle proposant le moins d'éléments nuls. La formulation du TO devient alors :

$$\gamma_0^\lambda = \underset{\gamma \in \beta}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F - \frac{1}{\lambda} h(\gamma),$$

où λ est un terme de régularisation. Pour λ très grand, le terme de régularisation tend vers 0 et nous retrouvons le TO [EMD](#). Nous définissons la matrice résultante γ_0 par une matrice de permutation entre les échantillons des deux domaines. La figure [6.4](#) représente la solution optimale trouvée dans le cas de notre exemple. La colonne de gauche présente la matrice de couplage γ_0 (en haut) et les permutations associées (en bas) pour la résolution [EMD](#), la colonne de droite représente la résolution Sinkhorn. Dans ce problème simple, il se trouve que les permutations sont les mêmes pour les deux algorithmes.

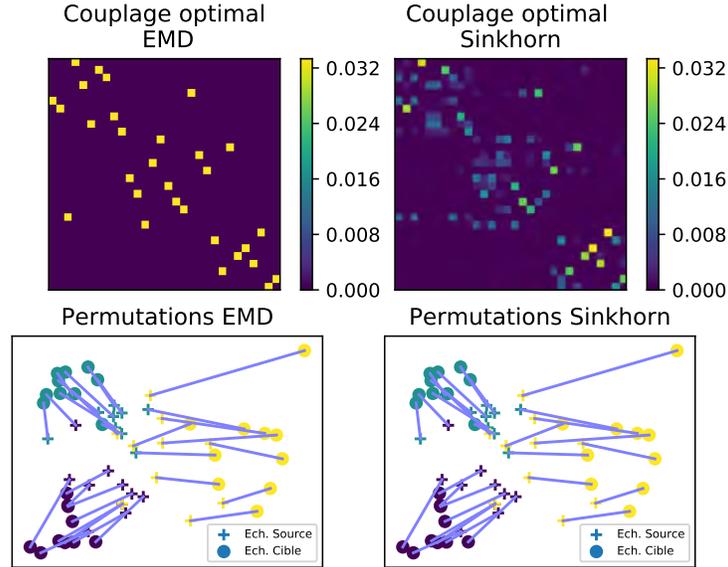


FIGURE 6.4 : Matrices de couplage (ligne supérieure) et permutations associées (inférieure) pour les deux algorithmes de résolution EMD (gauche) et Sinkhorn (droite).

Une fois la matrice de permutation calculée, il ne reste plus qu'à réaliser le transport des échantillons \mathbf{X}_s du domaine Source dans le domaine Cible :

$$\hat{\mathbf{X}}_s = \text{diag}((\gamma_0 \mathbf{1}_{n_c})^{-1}) \gamma_0 \mathbf{X}_c.$$

Les échantillons transportés $\hat{\mathbf{X}}_s$ sont ainsi reliés aux échantillons du domaine cible \mathbf{X}_c . Cette équation permet d'exprimer les échantillons transportés comme des barycentres des échantillons issus de \mathbf{X}_c en fonction des poids fournis dans la matrice de couplage optimale γ_0 . Ainsi, pour la formulation [EMD](#), si tous les échantillons à transporter sont inclus dans l'estimation du transport, il s'agit d'une simple opération de permutation entre les échantillons des deux domaines. La figure [6.5](#) illustre l'effet de ces deux solutions. Dans cette figure, la colonne de gauche illustre l'EMD pour lequel le couplage est parfait : chaque échantillon se superpose parfaitement à son correspondant. Alors que dans la colonne de droite représentant le Sinkhorn, on remarque nettement l'effet de pondération de chaque échantillon cible impliqué dans le transport.

La projection barycentrique permet également de transporter des échantillons non utilisés lors de l'estimation de la matrice de couplage. Notons $X_s = \{X_1 \cup X_2 | X_1 \cap X_2 = \emptyset\}$ l'ensemble des échantillons du domaine Source. Pour des raisons pratiques, $n_s \gg n_c$ par exemple, le TO est estimé uniquement à partir de X_1 , X_2 contient l'ensemble des échantillons en surplus. L'estimation du TO s'avère strictement identique

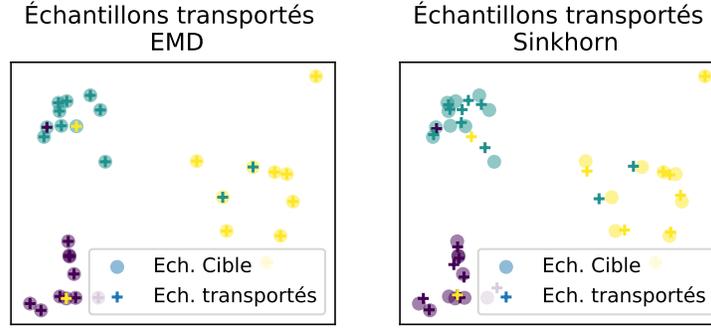


FIGURE 6.5 : Échantillons transportés pour les deux algorithmes de résolution EMD (gauche) et Sinkhorn (droite).

à celui exposé jusqu'ici. Lors du transport des échantillons, nous disposons de deux cas pour transporter l'échantillon x_i :

$$\hat{\mathbf{X}}_s = \begin{cases} \hat{x}_i = \hat{\mathbf{X}}_1[i] & \text{si } x_i \in X_1 \\ \hat{x}_i = \hat{\mathbf{X}}_1[j] + x_i - x_j & \text{si } x_i \in X_2 \end{cases}$$

Où l'échantillon x_j est le plus proche de x_i , au sens de la distance utilisée lors de l'estimation du TO.

6.2.2 L'utilisation des connaissances sur les deux domaines

Nous avons la possibilité d'exploiter des connaissances a priori sur les étiquettes pour les deux domaines. Nous disposons alors du choix entre l'exploitation des étiquettes du domaine Source uniquement, ou bien l'utilisation de celles du domaine Cible également. Concrètement, l'utilisation des étiquettes influence directement la matrice de coût. Au lieu de calculer la distance entre tous les échantillons des deux domaines, le coût n'est calculé qu'entre échantillons d'une même classe. Dans le cas d'échantillons de classes différentes la valeur du coût est définie très grande ce qui réduit les chances pour cette permutation d'être sélectionnée. En d'autres termes on cherche une matrice de permutation par classe. On peut alors envisager de décomposer le problème du TO à N classes, en N sous-problèmes.

L'exploitation des étiquettes du domaine Source uniquement n'a aucun impact sur la matrice de coût. Elles peuvent cependant être utilisées avec une formule de régularisation que nous n'aborderons pas ici [Rakotomamonjy et al., 2015].

La figure 6.6 illustre l'impact de l'utilisation des étiquettes sur la matrice de coût. Nous avons illustré trois cas d'utilisation des étiquettes :

- Non supervisé : seules les étiquettes du domaine Source sont connues
- Semi-supervisé : les étiquettes du domaine Source sont utilisées, et un échantillon sur deux du domaine Cible possède une étiquette
- Supervisé : tous les échantillons des deux domaines ont une étiquette associée

Dans cette figure, la couleur rouge indique la valeur de coût maximale (au moins 100 fois la valeur moyenne), préalablement masquée pour permettre l'affichage de la dynamique des valeurs de coût. Les valeurs rouges impliquent l'interdiction d'un transport entre les échantillons concernés.

Impact des étiquettes sur la matrice de coût

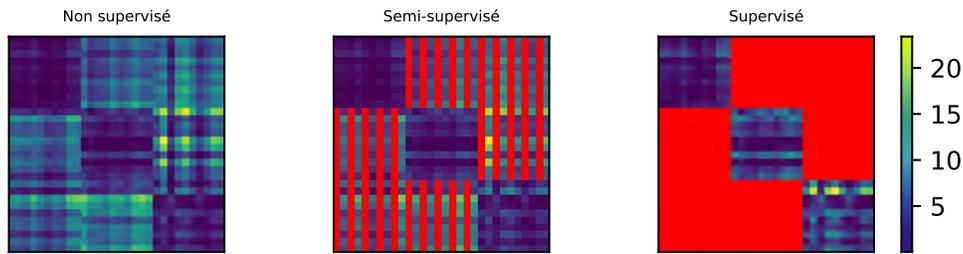


FIGURE 6.6 : Matrice de coût et utilisation des étiquettes. Le cas non supervisé : seulement les étiquettes du domaine Source ; Semi-supervisé : un échantillon sur deux du domaine Cible possède une étiquette ; Supervisé : tous les échantillons ont une étiquette.

Les étiquettes du domaine Source peuvent également être utilisées lors du transport des échantillons. Dans ce cas, nous calculons seulement le coût entre échantillons d'une même classe avant de réaliser la projection barycentrique. Cette utilisation produit alors les mêmes résultats que l'estimation pour un transport de chaque classe indépendamment.

6.3 Utilisation du Transport Optimal pour la production d'OCS

Le TO est une méthode d'adaptation de domaine qui se base sur l'adaptation de données. Elle représente donc une étape préliminaire à l'apprentissage du classifieur. En effet, une fois l'intégralité des échantillons du domaine Source transportée, nous pouvons revenir sur un processus de classification supervisée standard. Nous utilisons les échantillons transportés et leurs étiquettes pour l'apprentissage du classifieur, puis les échantillons cibles sont classés.

Dans cette thèse nous avons opté pour un processus en 5 étapes :

1. La sélection des échantillons du Domaine Source et Cible
2. L'estimation du TO
3. Le transport de la totalité des échantillons d'apprentissage du Domaine Source
4. L'apprentissage du classifieur
5. La classification du Domaine Cible

La donnée de référence d'apprentissage du domaine Source fournit les échantillons utilisés pour estimer le transport. En effet, les données de référence permettent de connaître les échantillons et sont représentatives de la zone d'étude. Cependant, la quantité d'échantillons disponibles constitue un ensemble très volumineux et déséquilibrés. Nous pouvons alors remplir les conditions d'utilisation du TO en sous-échantillonnant les données de référence. Par contre, l'intégralité des échantillons d'apprentissage du domaine Source sont ensuite transportés, à partir des équations présentées dans la section 6.2.1.

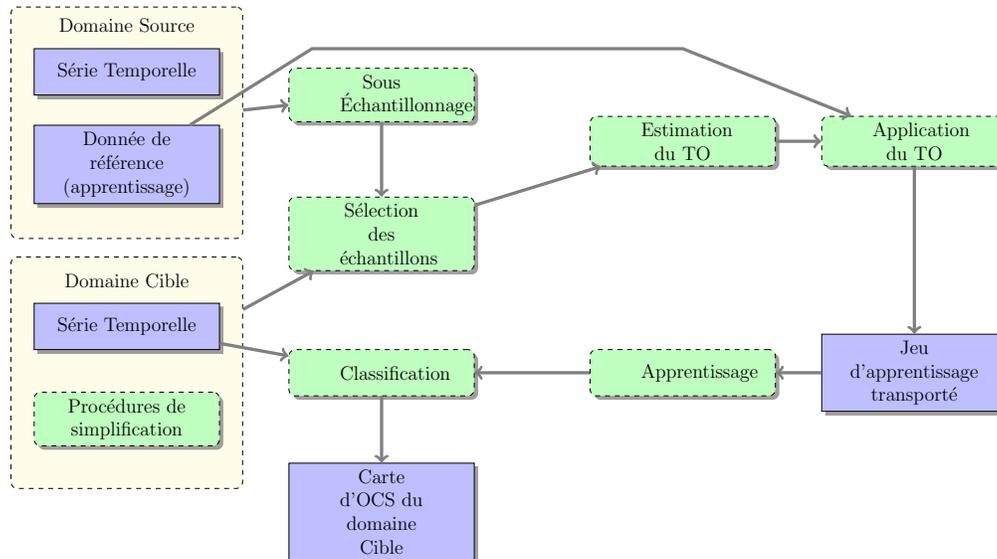


FIGURE 6.7 : Chaîne de traitement utilisant le Transport Optimal pour la production de cartes d'OCS.

Les phases d'apprentissage et de classification ne diffèrent en rien de celles utilisées dans le cas supervisé ou naïf, les paramètres du [RF](#) sont identiques.

Le TO peut être résolu avec l'EMD et le Sinkhorn indifféremment.

La problématique de la sélection des échantillons s'est révélée complexe. Nous présenterons les différentes solutions dans la section suivante.

6.4 Application du Transport Optimal à un vaste jeu de données

À partir des éléments théoriques, nous avons identifiés trois grands facteurs comme des éléments susceptibles d'influer les performances du TO. Ces trois facteurs sont respectivement : le nombre total de classes à traiter en même temps, le nombre d'échantillons à utiliser pour l'estimation du TO, et la méthode de sélection de ces échantillons. L'objectif premier consiste à mesurer leur impact en exploitant un jeu de données contrôlé, en faisant varier ces trois paramètres, préalablement à l'application du TO dans des conditions réelles d'exploitation. Les deux sections suivantes présentent respectivement les différentes études de sensibilité et les performances obtenues sur la production de cartes d'OCS.

6.4.1 Études de sensibilité

Dans cette section, nous présentons les différentes études de sensibilité et leurs résultats. Contrôler le jeu de données s'avère très aisé, il suffit d'exploiter les données de référence des deux domaines. Ceux-ci ayant un taux d'erreur d'environ 2 %, la fiabilité de l'étude est donc très grande. Nous avons fait le choix de restreindre l'étude de sensibilité à des cas représentatifs : c'est-à-dire susceptibles d'être reproduits en situation de production. Ainsi, nous fixons la limite maximale du nombre d'échantillons par le jeu d'apprentissage du domaine Source.

Protocole expérimental

Nous avons mis en place un protocole rigoureux de sélection d'échantillons pour réaliser cette analyse de sensibilité, et ainsi assurer la cohérence des résultats des différentes études.

Une autre approche d'évaluation du TO est généralement exploitée : nous réalisons un premier transport des données du domaine Source vers le domaine Cible. Puis on réalise le transport des données Sources transportées vers le domaine Source. L'évaluation porte alors sur le nombre d'échantillons correspondant exactement entre le domaine Source original et le domaine Source estimé. Cet aller-retour entre les deux domaines ne permet pas d'évaluer la pertinence de la classification des données transformées. Nous sommes donc contraint de réaliser la procédure de classification en intégralité. En effet, une fois la carte produite nous pouvons évaluer à l'aide des métriques présentée dans la section 2.3.1, l'apport du TO, car nous ne cherchons pas à évaluer le TO en lui-même mais son utilité pour la production de cartes d'OCS.

Nous avons indiqué précédemment que l'intégralité des échantillons d'apprentissage sont transportés permettant de comparer les RF entraînés. Pour conserver l'indépendance lors de la validation, les échantillons du domaine Cible utilisés lors de l'estimation du TO n'appartiennent pas au jeu de validation. Pour les cas d'étude de cette section, nous respectons cette contrainte en utilisant uniquement les jeux dédiés à l'apprentissage préalablement séparés des données de validation. Nous réalisons la séparation 10 fois afin de permettre le calcul d'intervalles de confiance à 95%, comme indiqué dans la section 2.3.1.

Les données de référence ne sont pas équilibrées 2.1.1 : nous constatons une variation du nombre d'échantillons par classe. Il devient alors nécessaire d'équilibrer les classes afin de pouvoir utiliser une distribution uniforme des masses de probabilités. Nous réalisons la sélection équilibrée de manière aléatoire parmi les échantillons disponibles jusqu'à obtention de la quantité requise.

Nous devons gérer l'absence des classes dans les deux domaines. En effet, si nous reprenons la visualisation apportée par la formulation de Kantorovitch, chaque classe possède sa propre "direction" lors du transport. S'il manque une classe dans un des domaines, alors l'estimation d'un transport pertinent pour cette classe devient impossible. Pour palier ce problème, nous décidons de ne pas considérer cette (ou ces) classe(s) lors de l'estimation du transport et d'utiliser les échantillons du domaine Source pour l'apprentissage du RF. Il s'agit du même traitement que pour les classes annuelles dans l'approche SAP. Cette solution ne concerne que le domaine Source. Pour le domaine Cible, nous ne possédons aucun moyen fiable permettant de déterminer les classes présentes.

Approches Supervisées

Cette approche vise à évaluer les performances obtenues lorsque nous utilisons l'EMD avec des données réelles, issues de notre jeu de données Formosat-2. Les échantillons utilisés proviennent des données de référence des deux domaines, le taux d'erreur d'échantillonnage se révèle donc très bas. La classe la moins bien représentée dans les deux domaines impose la quantité d'échantillons requise pour équilibrer les domaines. Afin de conserver un nombre minimum d'échantillons, nous considérons une classe comme non représentée si elle possède un nombre d'échantillons inférieur à 100. En moyenne, avec le jeu de données Formosat-2, le nombre d'échantillons par classe monte à 200 et le nombre de classes à 15 (au lieu de 17). Ainsi, il s'agit d'un total moyen de 3000 échantillons par domaine, sélectionnés aléatoirement dans leurs données de référence respectives. L'EMD est estimé une seule fois avec l'intégralité des classes éligibles, avec une distribution uniforme et une étape de normalisation des échantillons. Ce cas permet donc d'utiliser l'EMD et le Sinkhorn

avec des données parfaites, le résultat attendu constituerait donc la meilleure performance possible pour ces algorithmes. Nous réalisons le test pour chaque paire d'années, et les 10 tirages.

Nous utilisons cette configuration deux fois, une première fois via l'usage standard avec seulement les échantillons puis une seconde fois en utilisant les étiquettes des échantillons lors de l'estimation du TO.

Résultats des approches supervisées

Les résultats globaux (PG) de ces cas sont synthétisés sous forme de matrices. Pour chaque matrice, l'année indiquée sur l'axe des abscisses représente le domaine Source et sur l'axe des ordonnées le domaine Cible. Chaque figure se compose de trois matrices représentant des informations différentes. Celle de gauche représente la PG obtenue, l'échelle de couleur varie du rouge (PG = 0) au bleu (PG = 1), une teinte jaune indique une performance moyenne à 0.5. Celle du milieu représente la différence de précision entre le cas naïf et la méthode analysée, une teinte rouge est synonyme de dégradation à l'inverse du bleu. L'intensité de la teinte situe l'importance de l'écart, pour une valeur nulle représentée en blanc. Enfin la matrice de droite représente les intervalles de confiance à 95 % obtenus pour les différents tirages. L'échelle de couleur n'utilise pas de teintes continues afin de maximiser le contraste. Là encore une teinte bleue témoigne d'une bonne performance avec des intervalles de confiance petits et une teinte rouge pour une variation de l'ordre de 20 %.

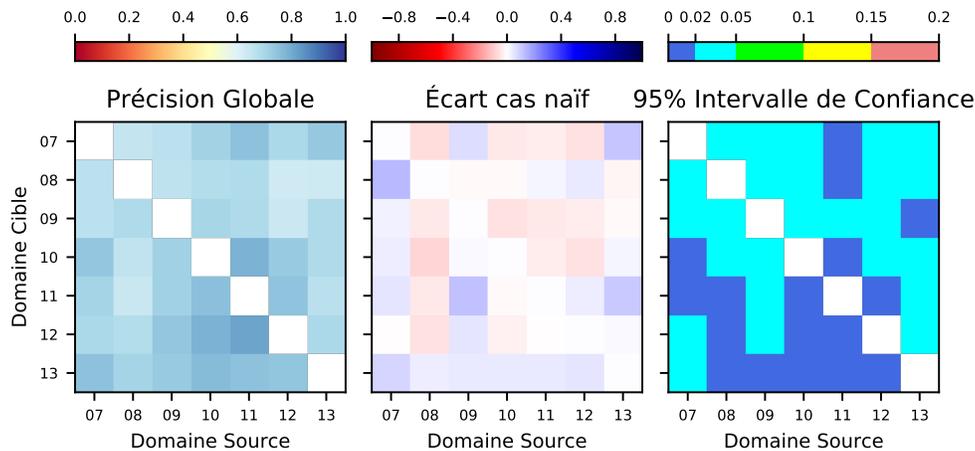


FIGURE 6.8 : Matrices de performances de l'EMD supervisé, l'axe des abscisses représente le domaine Source et l'axe des ordonnées le domaine Cible. La matrice de gauche représente la PG celle au centre représente la différence avec le cas naïf et celle de droite les intervalles de confiance à 95% (Les valeurs indiquées représentent la borne maximale de l'intervale de confiance, il faut donc comprendre $PG \pm I_{c_{max}}$).

La figure 6.8 représente les performances obtenues avec l'EMD. Ces résultats traduisent de bonnes performances en moyenne avec une PG supérieure à 60 % pour toutes les paires d'années. Les intervalles de confiance demeurent petits révélant une robustesse de l'approche face au bruit d'échantillonnage. L'écart avec le cas naïf bien que non significatif met en lumière un premier apport de l'EMD. En effet, les cas d'amélioration les plus importants se situent sur les cas ayant une PG faible lors de la classification naïve, présentée dans la table 3.1, comme par exemple pour les cas (2007, 2008) ou (2009, 2011). Globalement, une amélioration correspond à des valeurs de PG de la classification naïve inférieure à 70 %. Inversement, pour

celles supérieures à ce seuil nous observons une dégradation lors de l'utilisation du TO. L'utilisation de 2008 comme domaine Source semble contredire ces constatations, puisque nous observons systématiquement une dégradation des performances. L'information apportée par le FScore permet de mieux comprendre l'impact du TO.

On utilise également ce principe de matrices pour représenter le FScore par classe. Dans les figures de FScore on compte 17 matrices, une pour chaque classe. Les valeurs affichées désignent celles de l'écart par rapport au cas naïf. Elles suivent donc le principe de la barre de couleur divergente, commençant au rouge foncé pour mettre en évidence une chute du FScore, une teinte claire signifie un écart moins important et enfin le bleu une amélioration du FScore.

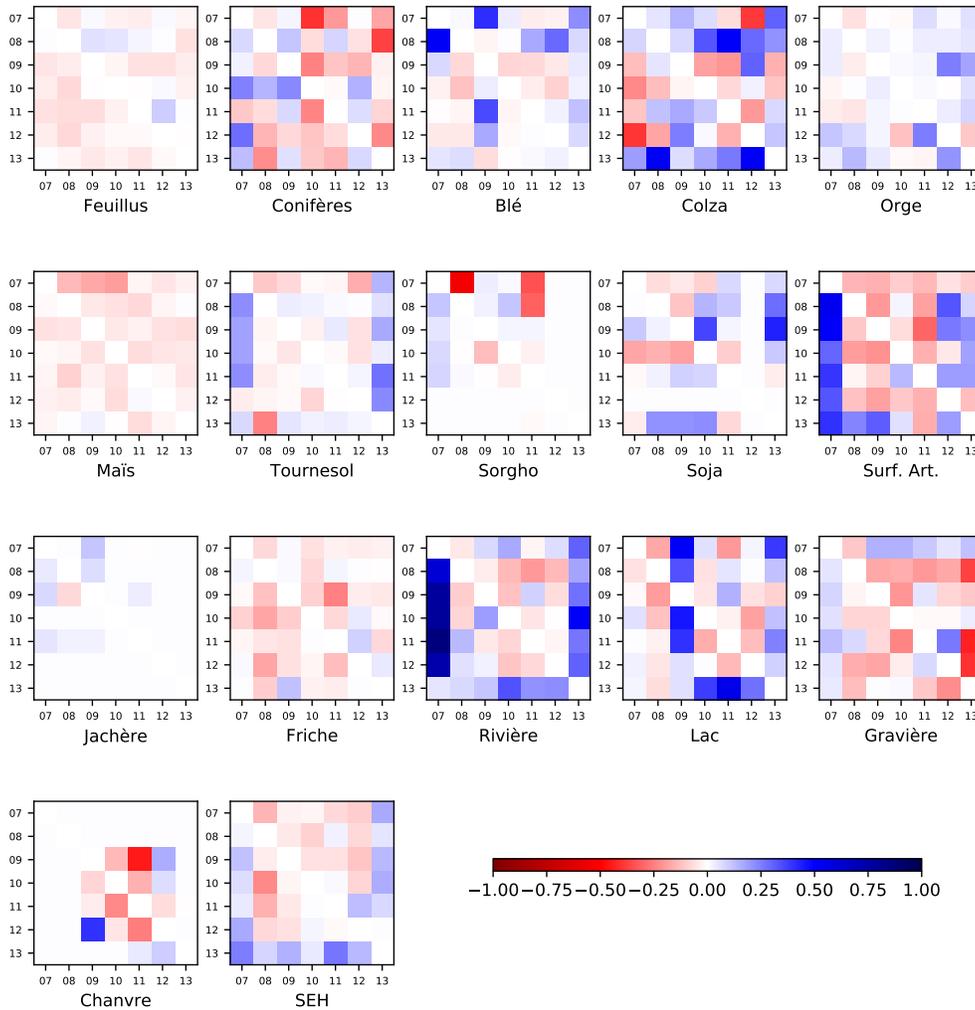


FIGURE 6.9 : FScore par classe obtenu par l'EMD. Chaque matrice représente une classe indiquée en abscisse, les axes sont les années des domaines Source (abscisses) et Cible. L'échelle de couleur varie du rouge (dégradation) au bleu (amélioration). Une valeur blanche signifie que l'écart est nul.

La figure 6.9 présente le FScore obtenu par l'EMD. Nous pouvons constater qu'il existe un fort contraste

inter mais également intra-classes. Alors que les classes de Feuillus et Maïs souffrent d’une légère dégradation (le FScore naïf avoisine les 90%), toutes les autres classes révèlent un comportement variable. Par exemple, pour les Rivières, on observe une très forte amélioration lorsque 2007 et 2013 sont les domaines Source et dans les autres cas, c’est plutôt une dégradation qui est observée. Le Colza représente également un cas intéressant puisque les améliorations et dégradations ne semblent dépendre d’aucune combinaison de domaine Source et Cible. Avant d’aller plus loin dans les conclusions, regardons les performances des autres approches.

La figure 6.10 présente le Sinkhorn. Le constat apparait sans équivoque : les performances s’avèrent mauvaises, avec une dégradation vis-à-vis du cas naïf de plus de 20%. L’observation des performances par classes dans la figure en annexe D.1, nuance cette première impression. Certes les performances sont sérieusement dégradées mais dans quelques rares cas nous observons une amélioration vis-à-vis du cas naïf. Il semble donc que le Sinkhorn soit en mesure d’offrir un transport pertinent mais que le problème en l’état semble trop complexe. Le grand nombre de classes (17) et la ressemblance entre certaines d’entre elles (Blé et Orge ou Lacs et Rivières) constitue un facteur probable permettant d’expliquer ces résultats.

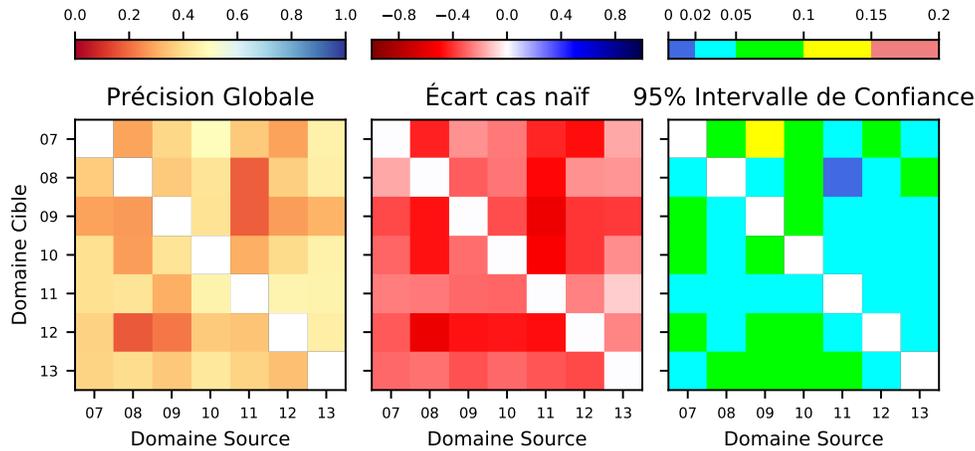


FIGURE 6.10 : PG du Sinkhorn avec les échantillons issus des données de référence.

L’utilisation des étiquettes de classes dans les deux domaines conforte cette dernière hypothèse. Que ce soit pour l’EMD ou le Sinkhorn, l’amélioration obtenue reste significative. Les figures 6.11 et 6.12 présentent respectivement ces performances. L’EMD s’approche des performances du supervisé avec des PG proches des 90 %. Dans le cas du Sinkhorn, l’amélioration se révèle plus modérée, cependant ces résultats montrent qu’il est en mesure de fournir un transport cohérent, lorsque nous réduisons les permutations possibles entre les classes similaires.

L’analyse du FScore pour ces deux cas, disponible en annexe D.2 et D.3 respectivement, met en évidence un effet retors de la PG. Alors que l’EMD offre une amélioration pour toutes les classes, le Sinkhorn améliore seulement le FScore des classes Blé et Colza, qui sont fortement représentées dans la zone d’étude, ce qui améliore la PG.

Ces quatre cas, ne présentent qu’un faible intérêt du point de vue de la production de cartes d’OCS, car les gains demeurent minimes en dehors de l’utilisation des étiquettes des deux domaines avec l’EMD. Cependant ils offrent une bonne appréciation des performances que nous pouvons attendre du TO. La formulation régularisée ne semble pas adaptée pour traiter un problème à 108 dimensions et 17 classes. Pour être compétitif,

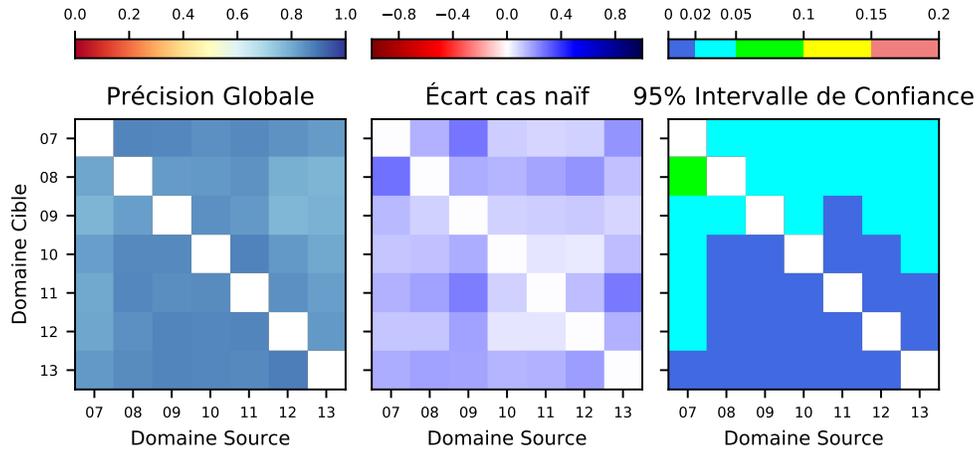


FIGURE 6.11 : PG de l'EMD avec les échantillons issus des données de référence et l'utilisation des étiquettes.

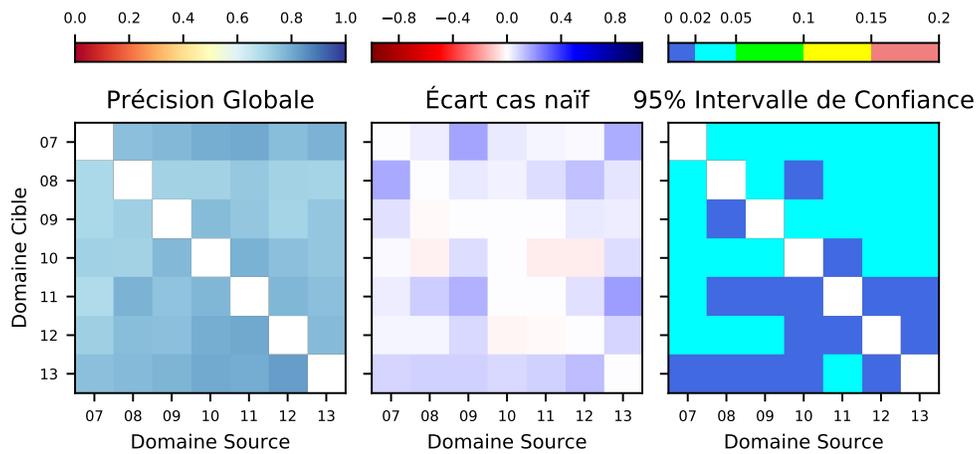


FIGURE 6.12 : PG du Sinkhorn avec les échantillons issus des données de référence et l'utilisation des étiquettes.

l'EMD a besoin d'informations a priori sur les échantillons qu'il exploite. Armés de ces constatation nous pouvons maintenant étudier des cas d'exploitations concrets.

6.4.2 Production opérationnelle

La sélection purement aléatoire

Dans la plupart des travaux d'adaptation de domaine existants par ailleurs, la méthode de sélection des échantillons n'est pas évoquée. Au vu de l'étendue et des spécificités des zones d'études, une sélection aléatoire peut souvent suffire pour obtenir les échantillons requis pour l'estimation du TO. Dans notre cas, nous sommes soumis à deux contraintes fortes :

1. La zone d'étude vaste : l'OCS se révèle donc variée en termes de diversité de classes et de proportions.
2. L'équilibre des classes entre les domaines Source et Cible doit être au maximum préservé pour satisfaire les pré-requis du TO.

Ces conditions impliquent une forte difficulté à maintenir l'équilibre entre les classes et assurer la sélection de toutes les classes. Nous devons exclure de fait une sélection aléatoire des échantillons dans la zone d'étude. Une expérimentation rapide a permis de confirmer la non fiabilité de la sélection purement aléatoire. En effet, après transport la PG obtenue pour de telles cartes ne dépasse pas les 5 %, avec une majorité d'expérimentation proche de 0. Le meilleur cas a été observé pour une sélection de nombreux échantillons de Feuillus dans les deux domaines, ce qui met en évidence la nécessité d'assurer la bonne représentation des classes pour l'estimation du TO.

Une autre approche consiste à découper l'emprise totale en plusieurs petites vignettes. Cependant, cette approche complexifie la conservation de l'équilibre des classes ainsi que la présence de toutes les classes lors du TO.

Sélection par position

Dans une image, les pixels mixtes constituent un facteur pouvant impacter la qualité du TO. Ces pixels représentent la transition entre deux objets d'OCS, en raison de la résolution du capteur, comme l'illustre la figure 6.13. Dans l'optique de sélectionner des échantillons les plus pertinents possibles, il faut alors minimiser la présence de tels échantillons.

À cette fin nous disposons des données de référence des années passées. Même si l'information d'étiquette est obsolète l'information de géométrie reste valable plus longtemps. En effet, les parcelles désignées par les polygones varient très peu au fil du temps. Les facteurs susceptibles de provoquer un changement de géométrie sont essentiellement l'urbanisation, la déforestation ou les changements de cultures (plantation de cultures différentes en divisant la parcelle originale). Si les deux premiers sont des processus lents, pouvant être réduits en minimisant l'écart entre les années des domaines Source et Cible. Le changement des parcelles agricoles s'avère plus imprévisible. Les raisons tiennent aux rotations de cultures, ou la division de parcelles dédiées à des cultures d'été succédant à une monoculture d'hiver, ou encore de mise en jachère. En suivant le principe de la rotation des cultures et la conservation des proportions entre les classes d'une année sur l'autre, nous optons pour une première technique de sélection.

Nous utilisons la donnée de référence pour définir les positions des échantillons du domaine Source. Dans cette approche nous prenons ces mêmes positions pour sélectionner les échantillons du domaine Cible, en supposant qu'il s'agit de la même classe que dans le domaine Source. Nous expérimentons cette approche de sélection avec l'algorithme [EMD](#) et Sinkhorn.



FIGURE 6.13 : Exemple de frontière entre objets associés à différentes classes.

Résultats

Nous exploitons à nouveau le formalisme de présentation des résultats, avec le même type de visualisation pour la PG et le FScore.

Les figures 6.14 et 6.15 représentent la PG et le FScore respectivement pour l'EMD. Comme attendu, les performances se révèlent inférieures à celles obtenues dans le cas supervisé sans étiquettes. Par rapport au cas naïf, nous obtenons, excepté dans 7 cas, une dégradation des performances. Cette différence s'avère peu marquée, généralement inférieure à 10%. Ces résultats mettent à mal l'hypothèse de stabilité des proportions des classes dans les parcelles relevées. Cette idée est confortée par une remarque sur les données de référence issues des campagnes terrains du CESBIO. Entre deux années successives, les polygones d'intérêts ne sont pas forcément tous identiques. D'une part, le nettoyage des relevés occulte certains, mais également car les campagnes terrain sont organisées pour être représentatives de l'OCS de l'année. Or, la rotation des cultures reste à l'initiative de chaque exploitant qui décide quelle culture succède à telle autre. Ces deux facteurs peuvent alors impacter les proportions des classes lors de l'estimation du transport et ainsi dégrader les performances de la classification.

La figure de FScore est particulièrement intéressante, car elle permet de comprendre l'origine des dégradations. Pour des classes très stables géographiquement comme les Rivières et Lacs, cette approche se révèle pertinente car elle offre de meilleures (ou très proches) performances par rapport au cas naïf. En effet, nous espérons que lors de l'estimation du TO, le coût entre des classes très différentes comme eau et végétation sera plus important qu'entre classes appartenant à une même catégorie.

À l'inverse, la classe de Maïs est systématiquement dégradée, alors qu'elle est cultivée sur des parcelles restant en monoculture de nombreuses années ; mais elle appartient au groupe des cultures agricoles. Nous connaissons déjà le handicap posé par cette notion de classe lors de la classification (comme pour le Blé et l'Orge), cependant elle prend encore plus de poids lorsque le TO est impliqué dans l'équation. Les performances du Sinkhorn n'apportent aucune information significative supplémentaire, elles sont présentées en annexes dans les figures D.4 pour la PG et D.5 pour le FScore. On peut tout de même noter que les performances obtenues sont très proches de celles obtenues par le Sinkhorn supervisé, ce qui ajoute du poids en faveur de l'hypothèse de l'inédaquation d'une résolution par régularisation au problème à 17 classes.

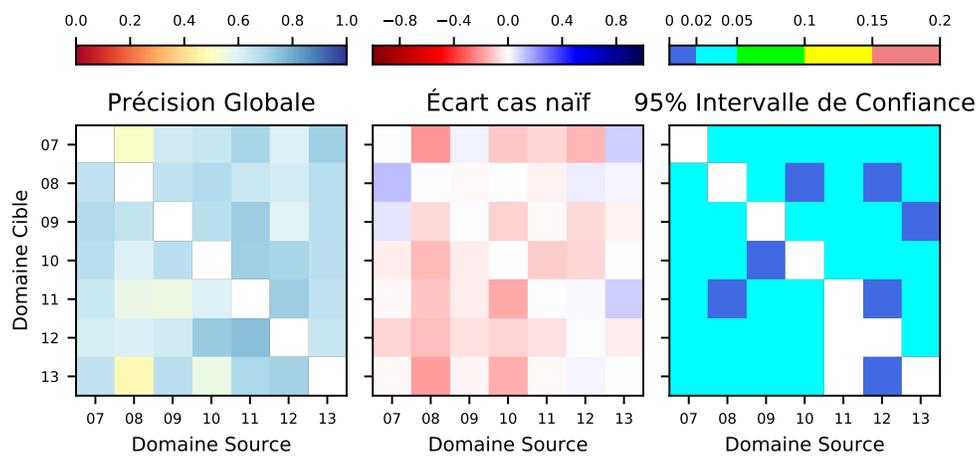


FIGURE 6.14 : Matrices de performances de l'EMD avec la sélection par position. L'axe des abscisses représente le domaine Source et l'axe des ordonnées le domaine Cible. La matrice de gauche représente la PG de l'EMD SP. La matrice de droite représente la différence avec le cas naïf.

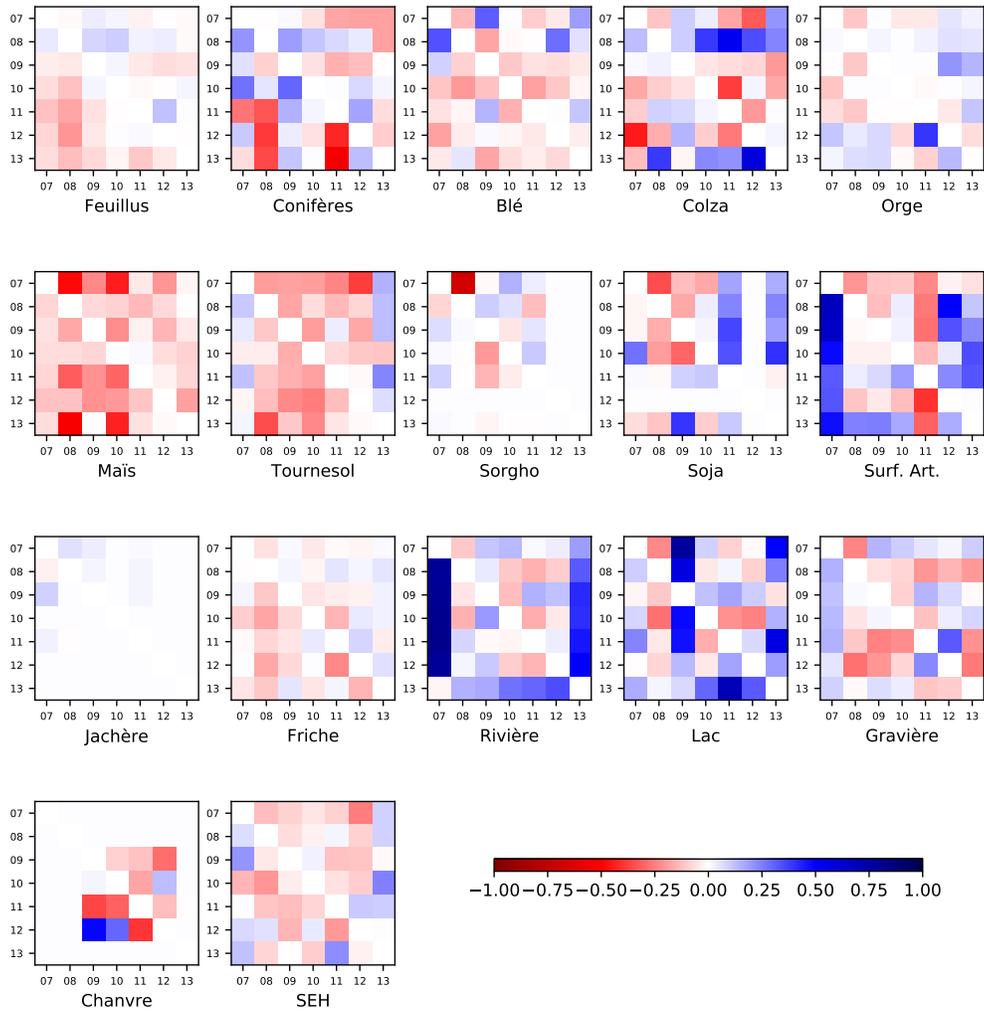


FIGURE 6.15 : FScore obtenus pour le EMD avec la sélection par position. Chaque matrice représente une classe, indiquée en abscises. L'axe des abscisses indique le domaine Source et celui des ordonnées le domaine Cible.

Sélection par étiquette

Par le biais de l'analyse de sensibilité, nous avons montré l'intérêt de l'utilisation des étiquettes dans le domaine Cible sur les performances de la classification après le TO. Dès lors, on peut transformer le problème du TO : Comment utiliser les différentes techniques de classification pour réaliser un premier étiquetage ? Cette problématique instaure les prémices d'un protocole itératif dans lequel la qualité de l'étiquetage impacte le résultat de la classification après TO. Cette dernière, théoriquement meilleure que la précédente peut alors être utilisée pour estimer un nouveau TO ; et ainsi de suite jusqu'à atteindre un palier de performances pour lequel chaque itération ne surpassera pas la précédente.

Nous avons expérimenté l'équivalence entre l'exploitation des étiquettes du domaine Cible, en termes de performance de classification et l'estimation individuelle d'un TO par classe.

Cette approche permet plusieurs améliorations possibles. D'une part, on peut alors appliquer l'approche [SAP](#). Ainsi seules les classes annuelles seront considérées, les positions des échantillons des classes pérennes provenant directement du domaine Source et extraites dans le domaine Cible. D'autre part en estimant un TO pour chaque classe annuelle, on peut augmenter le nombre d'échantillons utilisés et ainsi la diversité de cette classe sera mieux représentée et donc conservée lors du TO.

Un premier étiquetage représentera le point de départ de ce protocole. Nous ne sommes confrontés à aucune contrainte, en dehors de celles définies par le sujet de thèse : le domaine Cible ne dispose d'aucune donnée de référence. Devant le large choix des possibilités, nous retenons deux approches : une classification naïve et une étape de clustering afin de maximiser la variance intra-classes lors du Sinkhorn. En d'autres termes, nous cherchons à maximiser les différences entre les échantillons d'une même classe, afin qu'ils soient bien représentés lors du TO et par conséquent lors de l'apprentissage du classifieur. Ces sous-classes sont déterminées automatiquement à partir d'un algorithme non-supervisé.

Pour l'EMD, le TO est divisée en trois étapes :

1. Classification naïve du domaine Cible avec le modèle du domaine Source
2. Pour chaque classe considérée :
 - (a) Sélection aléatoire dans la carte de l'étape 1, avec limitation à 1000 échantillons pour cette classe.
 - (b) Sélection aléatoire dans le domaine Source, avec équilibrage sur le nombre d'échantillons du domaine Cible.
 - (c) Estimation du TO
3. Reconstitution du jeu d'apprentissage avec classes annuelles après TO, les classes pérennes provenant du domaine Cible (à partir des position du domaine Source, cf [SAP](#) dans la section 3.3) et les classes non traitées qui sont répliqués depuis le domaine Source comme dans la classification naïve.

La première étape consiste à réaliser une classification naïve. Nous savons qu'elle présente des performances tout à fait correctes sur la majorité des classes. Dans certains cas, la classification naïve ne permet pas de prédire une classe, que ce soit parce qu'elle n'est pas correctement classée (Orge) ou parce qu'elle n'existe pas dans le domaine Source (Jachères, Chanvre, etc). Dans le cas où nous ne comptabilisons pas suffisamment d'échantillons pour une classe, nous ne la considérons pas dans la procédure de transport. Les échantillons du domaine Source sont alors utilisés comme dans le cas de la classification naïve.

L'analyse de sensibilité a montré que le Sinkhorn est moins performant que l'EMD y compris avec les étiquettes des deux domaines. Cependant un des avantages du Sinkhorn réside dans sa capacité à conserver proches³ les échantillons d'une même classe, lors du transport.

Nous avons proposé alors de ne pas simplement utiliser les étiquettes mais de maximiser la représentation des sous-classes lors du TO Sinkhorn. Les sous-classes dans ce cas d'utilisation représentent des échantillons ayant des profils temporels différents (amplitude verticale) mais appartenant à la même classe. Pour cela nous décidons d'exploiter un k-moyennes [Hartigan and Wong, 1979] pour extraire les sous-classes à l'échelle de la carte. Un groupement en k-moyennes (*k-means* en anglais) propose une approche de groupement à partir d'échantillons et d'un entier k . Le principe consiste alors à grouper ces échantillons en k groupes de façon à minimiser une fonction de coût : le plus souvent, une simple distance euclidienne entre l'échantillon et le centre du groupe. Généralement, un k-moyennes, comme bien d'autre approches non supervisées s'avère inadapté à un problème multi-spectral ainsi qu'avec de nombreuses classes. Pour améliorer son efficacité, nous utilisons le k-moyennes sur les échantillons d'une seule classe, la figure 6.16 illustre le résultat pour la classe Blé. Nous optons arbitrairement pour 5 groupes, permettant une bonne représentation des sous-classes tout en conservant des temps de calcul raisonnables. On réalise le filtrage des classes avant le k-moyennes sur la classification supervisée pour le domaine Source et la classification naïve pour le domaine Cible. Les étiquettes du clustering sont utilisées uniquement pour assurer la diversité lors de l'échantillonnage. Elle ne sont pas utilisées lors de l'estimation du TO. Pour permettre une telle utilisation, il faudrait réaliser une mise en correspondance entre les clusters des deux domaines.

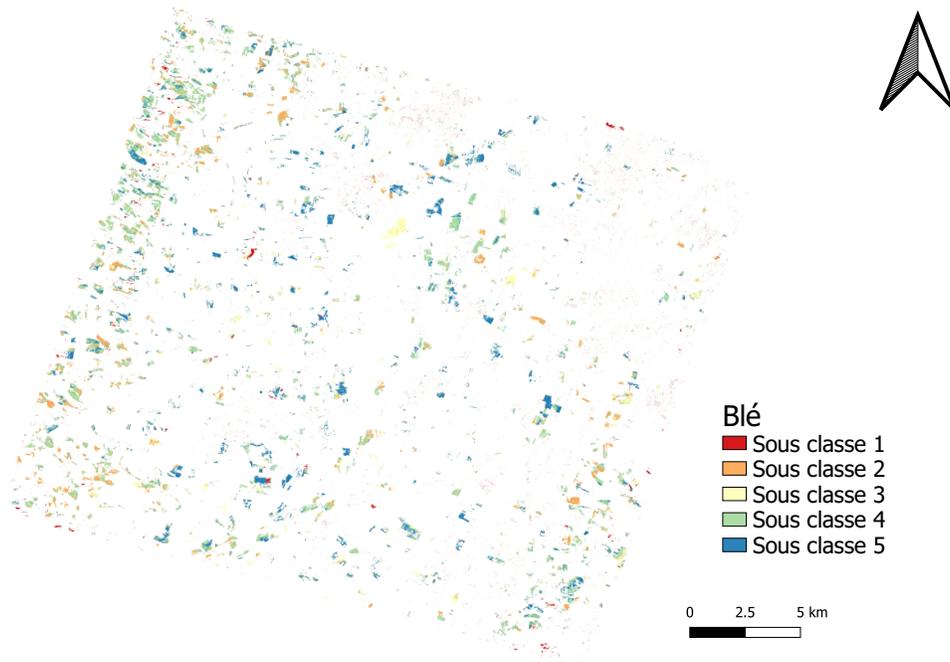


FIGURE 6.16 : Résultat du K-moyennes sur la série temporelle de la classe Blé. Le filtrage est réalisé sur la classification naïve.

³au sens de la distance entre les échantillons dans l'espace de dimension d

Le processus d'estimation du TO avec Sinkhorn est alors divisé en quatre étapes :

1. Classification naïve du domaine Cible avec le modèle du domaine Source
2. Pour chaque classe annuelle :
 - (a) Calcul du 5-moyennes sur la carte filtrée par l'étiquette de la classe.
 - (b) Sélection aléatoire dans la carte de l'étape 1, avec limitation à 1000 échantillons.
 - (c) Sélection aléatoire dans le domaine Source, avec équilibrage sur le nombre d'échantillons du domaine Cible.
 - (d) Estimation du TO
3. Reconstitution du jeu d'apprentissage.

Consécutivement à ces trois étapes, le **RF** est entraîné et la carte produite validée selon les méthodes classiques.

Résultats

Nous attendons beaucoup du découpage du problème du TO par classe. En effet, sous cet angle notre cas d'étude se positionne au plus près de l'idée centrale du transport : un objet possède une direction propre. Elle permet également de traiter plus d'échantillons lors de l'estimation du TO. La figure 6.17 présente les performances globales obtenues par l'EMD par classe, combiné avec la procédure **SAP**. Globalement, nous constatons une amélioration franche par rapport au cas naïf. Le FScore par classe 6.18 confirme cette amélioration. Dans cette figure, nous pouvons voir que toutes les classes sans exception profitent du TO.

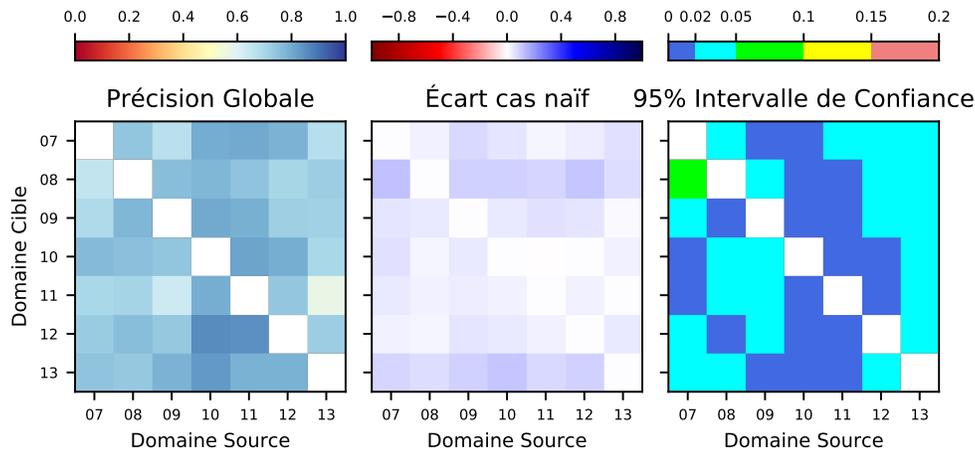


FIGURE 6.17 : Matrices de performances globales de l'EMD SAP. l'axe des abscisses représente le domaine Source et l'axe des ordonnées le domaine Cible. L'histogramme de gauche représente la PG du SAP EMD. L'histogramme du milieu représente la différence avec le cas naïf. Celui de droite représente la borne maximale des intervalles de confiance à 95% (on obtient l'intervalle grâce à $PG \pm I_{c_{max}}$).

Les intervalles de confiance présentés dans la figure 6.17 montrent que les résultats sont statistiquement significatifs. Ils permettent d'affirmer que le processus de sélection aléatoire exploitant la classification naïve et le TO réunis produisent moins de 20 % d'erreurs d'étiquetage [Pelletier et al., 2017]. En effet, si ce taux était dépassé, les performances s'en trouveraient fortement dégradées (à l'image du Sinkhorn supervisé). Pour rappel le taux d'erreur d'une classification naïve, pour notre jeu de données, varie entre 19 % et 49 %. On note une atténuation de ces erreurs d'une part par la sélection aléatoire d'un faible nombre d'échantillons (par rapport au nombre total d'échantillons disponibles dans l'image) et d'autre part par l'estimation du transport en capacité de corriger une partie de l'erreur d'étiquetage du domaine Cible. La combinaison des RF de la procédure SAP et du TO permet d'obtenir un résultat robuste face aux combinaisons possibles de domaines.

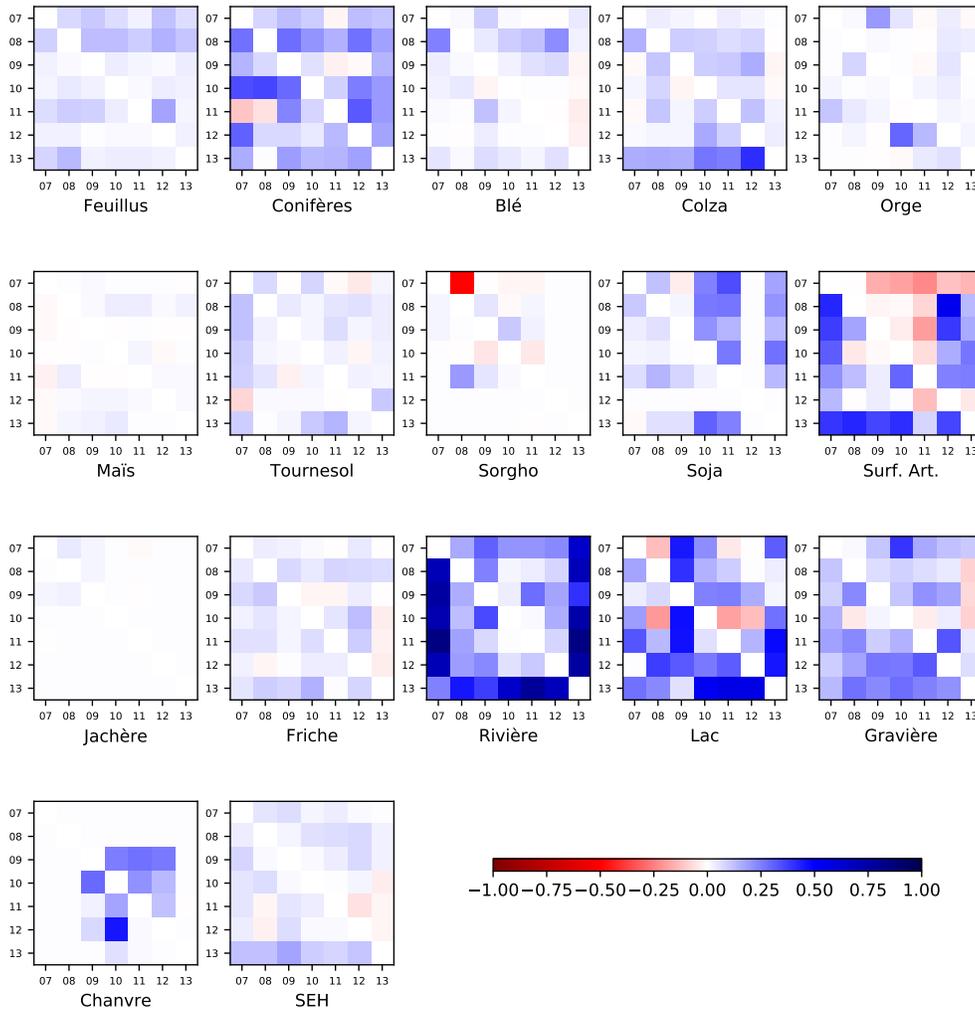


FIGURE 6.18 : FScore obtenu par le TO EMD combiné à la procédure SAP. Chaque matrice correspond à une classe, indiquée en abscisses. L'axe des abscisses représente le domaine Source et l'axe des ordonnées le domaine Cible.

Les performances obtenues par l'EMD par classe (figures D.6 et D.7 en annexes) ne sont pas très différentes de celles obtenues avec la procédure SAP. Ce manque de contraste flagrant s'explique par le fait que la sélection des classes pérennes ne représente pas une difficulté, puisqu'elles bénéficient d'un très bon classement lors de la classification naïve. Les caractéristiques des classes permettent également d'expliquer ce phénomène. En effet, si pour les classes pérennes, les distorsions ne se révèlent pas significatives pour le RF, alors il n'existe pas de raison pour qu'un transport pertinent dégrade cette particularité.

Nous pouvons constater que les cas de dégradations demeurent les mêmes dans les deux approches. Cependant les cas de gains sur les classes pérennes sont plus importants lorsqu'elles ne sont pas transportées. Dans la figure 6.18, nous pouvons voir, dans une minorité de cas, une dégradation du FScore, notamment pour les Surfaces Artificielles. La représentation en matrice permet de visualiser également la composante chronologique. La majorité des dégradations pour cette classes se situent dans la partie supérieure de la matrice. Cela implique que l'historique est utilisé pour produire une carte d'une année antérieure à elle-même.

La matrice de FScore de la classe de Surfaces Artificielles est presque symétrique vis-à-vis de la diagonale. Cet effet se traduit par la tendance d'évolution la plus probable de cette classe qui remplace les autres classes d'OCS.

Ainsi, lorsque nous récupérons les échantillons dans le domaine Cible à partir de la donnée de référence du domaine Source, nous ne disposons d'aucune garantie de correspondance. Alors que dans le sens chronologique c'est le plus souvent le cas.

Tous ces résultats confirment que la sélection des échantillons en amont du TO impacte significativement le résultat obtenu par l'EMD.

Les performances obtenues par le Sinkhorn après clustering sont présentées dans la figure 6.19 pour la PG et 6.20 pour le FScore. Cette décomposition du problème permet au Sinkhorn d'atteindre des performances proches de celles de la classification naïve. Elles restent cependant inférieures à celles-ci. Nous pouvons noter une similitude dans les améliorations avec le cas du TO Sinkhorn supervisé avec étiquette. Bien qu'elles restent inférieures, on voit apparaître les mêmes schémas dans les matrices. Cela signifie que l'apport de la diversité des sous classes lors du TO peut être significatif. Malheureusement nous retrouvons la même problématique que précédemment : la classification naïve n'est pas assez précise pour assurer les besoins de précision imposées par le TO.

6.5 La place du transport optimal dans la production opérationnelle

Au travers des expérimentations présentées, nous avons pu mettre en avant plusieurs obstacles pour une utilisation efficace du TO pour la production de cartes d'OCS.

Le premier concerne les échantillons utilisés pour l'estimation du TO. En effet, les approches proposées ont mis en évidence la relation forte entre l'estimation de la matrice de couplage et la classification résultante.

De plus, nous avons constaté que l'introduction d'une connaissance à priori sur l'étiquette des échantillons du domaine Cible se révèle un atout considérable pour le calcul des permutations. Cependant, ce second obstacle nous ramène au problème initial, à savoir déterminer avec le plus précisément ces étiquettes pour le domaine Cible.

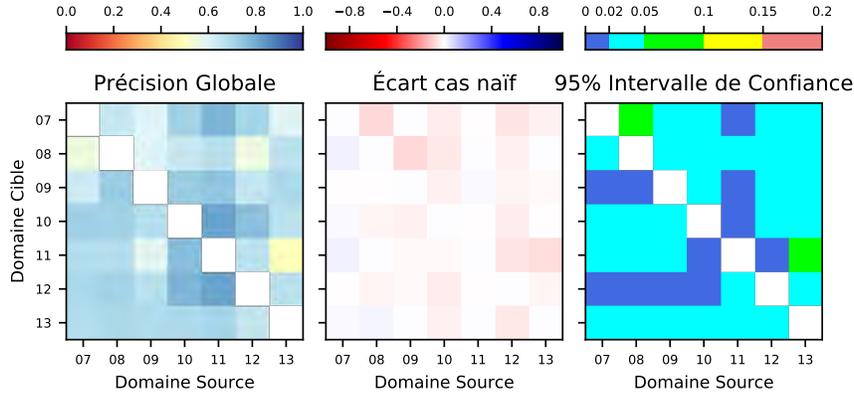


FIGURE 6.19 : Matrices de performances du Sinkhorn avec clustering. L’axe des abscisses représente le domaine Source et l’axe des ordonnées le domaine Cible. La matrice de gauche représente la PG, celle au centre représente la différence avec le cas naïf et celle de droite représente les intervalles de confiance à 95 %.

Une solution envisagée serait de recourir à un étiquetage manuel à la manière de l’apprentissage actif, mais incompatible avec les objectifs de production opérationnelle. Une autre idée serait de réaliser un transfert d’étiquette du domaine Source vers le domaine Cible. Nous n’avons pas eu le temps d’aborder des méthodes poussées de transfert d’étiquettes au cours de la thèse. Cependant, nous avons suivi la piste de la détection de changements lors de l’exploitation du MAD. Face aux problèmes de performances rencontrés et exposés au cours du chapitre précédent, cette piste reste en l’état, faute de méthode efficace.

Nous avons évoqué l’idée d’une procédure récursive de classifications successives. Dans cette optique la première étape consiste à réaliser une classification naïve, puis estimer un transport par classe. À partir de cette classification, nous pouvons alors sélectionner de nouveaux échantillons et estimer de nouveau un transport, et ainsi de suite jusqu’à convergence des performances. Les tests liés à cette approche ont rapidement mis en évidence des problèmes complexes. En effet, les performances d’une itération à l’autre tendent à se dégrader. Il est fortement probable que cela provient de la stratégie d’échantillonnage mise en place. Celle-ci est très, voire trop, simple : à chaque itération nous sélectionnons de nouveau l’ensemble des échantillons dans le domaine Cible. Une stratégie pertinente aurait été d’instaurer une mesure de pertinence des échantillons afin de pouvoir les conserver d’une itération sur l’autre. Un moyen efficace de détecter ce genre d’échantillons réside dans l’utilisation des méthodes de détection d’*outlier* [Pelletier et al., 2017] en analysant les décisions des RF. Une autre idée consiste à exploiter la mesure de confiance fournie en sortie par le classifieur pour guider la sélection d’échantillons pertinents. Pour des questions de temps, il n’a pas été possible d’aller plus en avant sur cette voie.

Lors de l’introduction générale au TO, nous avons évoqué les nombreuses variations possibles dans les modes de résolutions, d’estimation des matrices de coût et de couplage. L’utilisation du TO dans notre cas applicatif s’est révélée particulièrement chronophage. Nous avons, en effet, dû réaliser des tests sous différents paramètres avant d’obtenir nos premiers résultats pertinents. Puis s’est ajoutée la question de la sélection et de l’identification des échantillons.

L’estimation des masses de probabilités reste le troisième obstacle que nous avons rencontré. En effet, la distribution uniforme ne semble pas forcément la plus adaptée à un problème déséquilibré comme l’OCS.

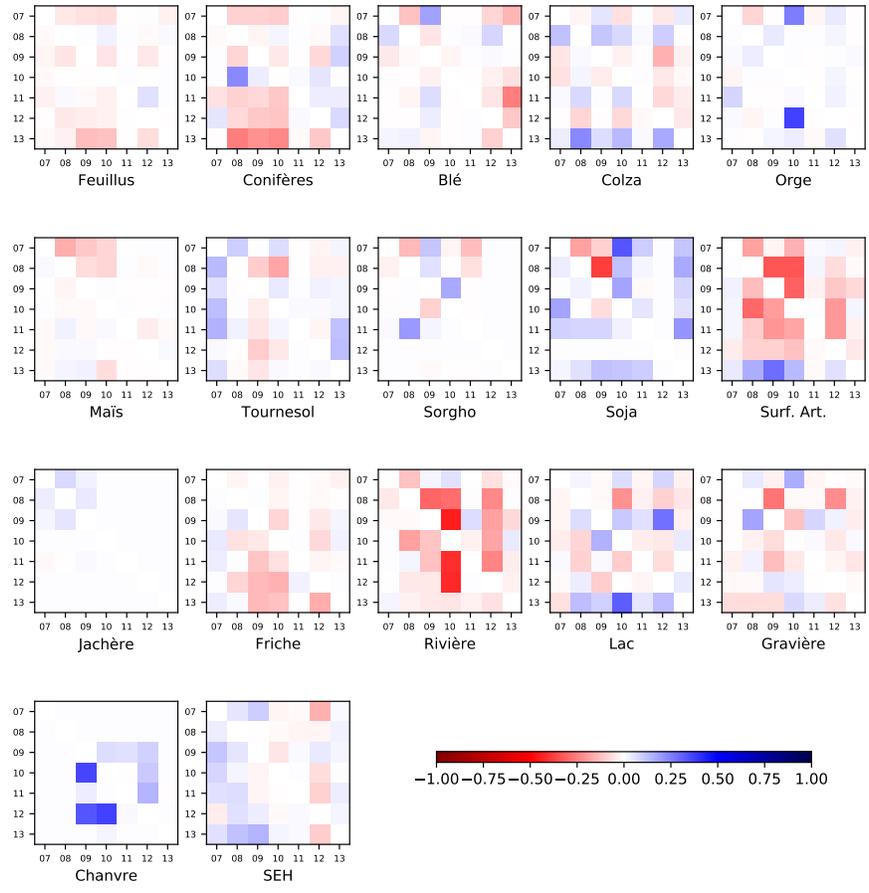


FIGURE 6.20 : FScore obtenus par le TO Sinkhorn avec clustering.

Cependant la meilleure manière d’estimer convenablement les proportions des classes dans le domaine Cible réside dans d’exploitation d’une classification aussi précise que possible ce qui nous ramène à notre point de départ.

Une perspective intéressante concerne le choix de la mesure de coût qui peut aussi impacter les performances du TO. Nous savons que la distance euclidienne perd en précision lorsque le nombre de dimensions est très grand. Dans le cas de séries temporelles, une solution serait de recourir à une mesure de distance plus pertinente comme la [DTW](#), mais cela implique un coût de traitement important qui rapidement révélera les contraintes liées à la production opérationnelle.

Tous ces obstacles représentent autant de perspectives à explorer dans la continuité de ces travaux. Néanmoins, le passage aux données Sentinel-2 pose de sérieuses questions de viabilité du TO, en raison du volume énorme de données à traiter. Ces conclusions s’appliquent également aux méthodes présentées lors de la revue de la littérature, ne laissant qu’une maigre place aux méthodes actuelles d’adaptation de domaine pour la production de cartes d’OCS. En effet l’application du TO représente un effort considérable pour un gain minime, les meilleures performances requérant un premier effort de classification. Si dans le cadre de Formosat-2 la classification ne prend que quelques minutes, il s’agit de plusieurs heures voire jours pour traiter Sentinel-2.

Récemment, [\[Redko et al., 2018\]](#) ont proposé de réaliser le TO en exploitant plusieurs domaines Source. La principale difficulté rencontrée étant l’estimation des proportions des classes dans chaque domaine.

Dans la suite du manuscrit nous présenterons des approches de fusion de données et de classifications permettant ainsi de combiner les informations issues de l’historique.

Quatrième partie

Propositions de méthodes exploitant de multiples domaines Source

Modélisation de l'historique

Dans la première partie des travaux de thèse nous avons exploité l'adaptation des données afin de permettre la classification d'un domaine Cible à partir d'un unique domaine Source. Cependant, nous avons mis en évidence la difficulté d'exploiter ces méthodes, pour n'obtenir qu'un gain minime par rapport à la classification naïve. Dans ce chapitre, nous proposons de combiner les données de l'historique pour produire la carte du domaine Cible. Dans un premier temps, nous exploitons les cartes supervisées, témoins de l'OCS des années précédentes en modélisant la succession des classes. Dans un second temps, nous exploitons les séries temporelles afin d'entraîner des classifieurs aptes à appréhender les distorsions entre les années.

7.1 Modéliser des connaissances

L'historique, tel que nous l'avons défini, contient tous les éléments nécessaires à la production de cartes d'OCS : images de télédétection, données de référence, classifieurs, cartes d'OCS, mesures statistiques de la précision des cartes. Nous pouvons associer à cet historique des connaissances annexes : données météo (pluviométries, températures), pratiques agricoles (ensilage, irrigation) par exemple. On appelle ce genre d'informations des "connaissances expertes". Elles sont généralement fournies par des spécialistes du domaine concerné, et servent en grande partie à l'exploitation des résultats de classification. Cependant, elles demeurent souvent liées à une région spécifique, exigeant un effort conséquent de collecte et d'harmonisation proportionnel à l'étendue de la zone concernée.

Il existe des méthodes permettant de modéliser les connaissances expertes afin de les exploiter lors d'un apprentissage. Ainsi, le rôle du classifieur consiste à comprendre et apprendre ces règles, afin de les généraliser lors de la classification. Pour cela, transformer les différentes connaissances en probabilités permet bien souvent une modélisation simple et exploitable par une règle d'inférence bayésienne. Néanmoins, ces approches ne permettent pas d'établir des liens sémantiques entre les différentes sources d'informations et se révèlent donc trop simplistes. De fait, on retrouve des approches complexes permettant d'établir des connexions entre les informations.

Ontologies

Une première approche exploite les ontologies [Uschold and Gruninger, 1996]. Une phrase simple composée d'un sujet S , d'un verbe V et d'un complément d'objet O constitue une règle. L'ensemble de ces règles constitue une ontologie. Il s'agit de vérités permettant d'établir des liens entre S et V . Par exemple "Blé est une culture d'hiver" ou "une culture d'hiver peut être du Blé". Ainsi, l'ontologie constitue la définition du domaine d'étude, où l'utilisateur crée des liens entre tous les éléments traités. On peut donc confronter, à l'aide d'un raisonneur, l'ontologie à des éléments n'appartenant pas au domaine décrit, afin de vérifier leur cohérence.

Cependant, en fonction de l'ampleur du domaine et la diversité des éléments le composant, l'ontologie atteint rapidement ses limites. En effet, chaque élément impliquant au minimum une règle envers les autres éléments, l'ajout d'un nouvel élément crée autant de nouvelles règles que d'éléments déjà existants. Il devient alors très difficile de conserver la cohérence entre toutes les règles. Le temps de résolution du problème augmente proportionnellement au nombre d'éléments de l'ontologie.

La notion d'incertitude constitue l'un des problèmes majeurs des ontologies liés à l'OCS. En effet, il n'existe aucune vérité absolue sur l'évolution d'une classe d'OCS. De fait, les ontologies constituent un outil puissant, très utilisé dans les problèmes de langages naturels [Reiter and Dale, 2000] ou pour des domaines où l'incertitude n'existe pas (ou peu) [Navigli and Velardi, 2004, Charlet et al., 2004, Le Ber et al., 2006b].

7.1.1 La logique de Markov

Une seconde approche de modélisation d'historique repose sur les chaînes et les champs de Markov.

Chaînes de Markov

Une chaîne de Markov représente un processus aléatoire satisfaisant la propriété de Markov, que l'on peut définir comme : l'état courant contient toute l'information requise à la prédiction de l'état futur et cette prédiction ne dépend pas des états antérieurs. Un état présente la condition actuelle du système étudié, par exemple pour la météo : "soleil", "pluie" constituent deux états d'intérêts.

Inversement, un exemple caractéristique de problème non markovien consiste à réaliser des tirages de boules numérotées dans une urne sans remise, la probabilité d'obtenir une boule donnée dépend alors de tous les tirages précédents.

La première étape d'un processus markovien consiste à déterminer l'équilibre du système : on attribue une probabilité de transition tr à chaque état e_i , vers lui-même et vers n autres états. On regroupe ces probabilités sous la forme d'une matrice de transition :

$$P = \begin{bmatrix} tr_{1,1} & \dots & tr_{1,n} \\ \dots & \dots & \dots \\ tr_{n,1} & \dots & tr_{n,n} \end{bmatrix}$$

Reprenons l'exemple de la prédiction de la météo. On peut exprimer une matrice de transition, avec des valeurs arbitraires, comme :

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

Cette matrice traduit les liens entre les états, ainsi avec une probabilité de 90 % un jour de “soleil” succédera à un autre, contre seulement 10 % de chances de basculer sur un jour de “pluie”. Mais, une fois dans cet état, le système possède la même probabilité de changer ou non d’état. Ainsi, ce problème suit la propriété de Markov, puisqu’il ne prend en compte que l’état actuel.

L’initialisation de la chaîne de Markov consiste à traduire le comportement actuel du système par des probabilités, et ce pour chaque état considéré. Par exemple l’état “aujourd’hui il fait soleil” se traduit par $e_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}$, mais nous pouvons également moduler l’information en attribuant à chaque état une valeur non nulle ainsi : $e_0 = \begin{bmatrix} 0.85 & 0.15 \end{bmatrix}$ qui décrit un jour ensoleillé mais avec la présence de nuages. Cette variation va pondérer la matrice de transition lors du calcul de la prédiction.

À partir de l’initialisation et de la matrice de transition P , la prédiction pour le jour suivant s’obtient par :

$$e_1 = e_0 P,$$

mais on peut également prédire pour le surlendemain :

$$e_2 = e_1 P = e_0 P^2,$$

qui est généralisable à la prédiction du $n^{\text{ième}}$ jour :

$$e_n = e_{n-1} P \text{ ou } e_n = e_0 P^n.$$

Un des avantages considérable des chaînes de Markov réside dans la possibilité de prédire à partir de l’état initial l’état à N étapes dans le futur. Néanmoins, cette modélisation devient vite limitée si le phénomène à modéliser s’observe sur plusieurs états, par exemple les rotations multi-annuelles des classes de cultures. On leur préfère les champs de Markov aléatoires, capables d’établir des liens entre divers états.

Champs de Markov aléatoires

Un ensemble de variables aléatoires respectant les propriétés markoviennes et décrites par un graphe non orienté constitue un champ de Markov aléatoire (*Markov Random Field* en anglais). Dans ce graphe, chaque noeud correspond à un état décrit par une variable. Chaque branche du graphe représente la probabilité de transition d’un état vers un autre. On constate alors que dans sa représentation, un champ aléatoire de Markov ressemble à un réseau bayésien. En effet, un réseau bayésien est également un modèle graphique probabiliste, permettant de représenter des variables aléatoires sous la forme d’un graphe orienté acyclique. Mais la différence majeure réside dans le fait que, contrairement aux réseaux bayésiens, il s’agit d’un graphe non orienté et potentiellement cyclique : une succession de plusieurs transitions à partir d’un état donné ramène le système dans ce même état. De fait, ces graphes représentent un atout non négligeable pour la modélisation des connaissances [Richardson and Domingos, 2006].

En effet, une chaîne de Markov permet de relier deux états, on parle alors d’ordre 1. Ainsi, on considère qu’elle ne possède pas de mémoire. Les graphes permettent d’exploiter des chaînes de Markov d’ordre 2, permettant de considérer l’état présent conjointement à l’état précédent, et ainsi de suite jusqu’à une chaîne de Markov d’ordre N prenant en compte alors tous les états passés. Ces différents degrés permettent alors la modélisation de connaissances complexes, comme la rotation des cultures. Ces approches dominent en fouille de données (*data mining*), mais également dans les applications de transfert d’apprentissage ou de détection de changements.

Modélisation des interactions de l'historique

De nombreux travaux existent exploitant les chaînes et les champs de Markov pour modéliser différentes interactions. Souvent, il s'agit de déterminer les probabilités de transitions entre les différentes classes concernées. Notamment, les travaux présentés par [Castellazzi et al., 2008] mettent en évidence la bonne performance des chaînes de Markov (ordre 1) pour la modélisation des rotations de cultures à partir de séries temporelles. Des connaissances expertes fournissent les probabilités de transitions entre les diverses classes étudiées. Une connaissance experte consiste à exploiter des vérités actuelles impactant ainsi les règles modélisées. Par exemple, la culture du Maïs dans le sud ouest est très répandue en raison des conditions de rentabilité favorables pour l'exploitant. De fait, on peut définir le Maïs comme une culture géographiquement stable, ce qui se traduit par une probabilité de transition faible.

Les travaux de [Dogliotti et al., 2003] proposent un outil ROTAT, permettant à partir de la liste des classes de cultures et de filtres définis par l'utilisateur de calculer toutes les rotations possibles. Ces deux exemples de travaux, requièrent un apport de l'utilisateur, qui joue le rôle d'expert. Cependant, cette notion représente un frein à une utilisation à large échelle, tout simplement car chaque expert maîtrise uniquement une zone restreinte. Néanmoins, une pratique courante consiste à réunir un comité d'experts lorsque la taille de la zone d'étude le requiert [Vancutsem et al., 2012].

Des travaux proposent de déterminer, directement à partir des données, les probabilités de transitions. Par exemple le modèle CarrotAge [Le Ber et al., 2006a] permet de détecter les rotations de cultures à partir d'une base de données (Teruti) ou CropRota [Schönhart et al., 2011] intégrant des critères agronomiques lors de l'estimation des transitions.

Les travaux de thèse de Julien Osman [Osman, 2015] considèrent ces différentes approches pour la production en temps réels d'une carte d'OCS. La finalité de ces travaux porte sur la possibilité d'une mise à jour de la carte au fil de l'arrivée des images, en exploitant les transitions et les champs aléatoires de Markov contraints par les règles d'un réseau bayésien (pas de cycle et une orientation des relations) : c'est la logique de Markov. Les travaux concernent essentiellement les classes de cultures d'été. Les transitions entre les classes sont extraites des différents millésimes du RPG. Dès lors, nous pouvons tirer plusieurs enseignements de ces travaux :

1. Les ontologies manquent d'amplitude (mesure d'incertitude) pour répondre à un problème comme l'OCS.
2. Les modèles de Markov permettent une amélioration des performances de la carte quand le nombre d'images est suffisant.
3. La possibilité de considérer chaque état indépendamment au moment de la prédiction, au lieu d'une succession de différents états, améliore les performances obtenues.

La logique de Markov constitue une méthode très coûteuse permettant d'estimer des règles à partir de probabilités a priori. Un des grands avantages reste la possibilité d'ajouter de nouvelles règles expertes, qui ne sont pas des probabilités estimées à partir des données, à la demande. Nous ne projetons pas l'exploitation de cet avantage, aussi dans ce cas, la logique de Markov peut être remplacée par un simple comptage d'occurrences de transitions [Osman et al., 2012].

Des approches exploitant des classifieurs afin de déterminer les probabilités de transitions existent [Bruzzone et al., 2004, Bruzzone and Serpico, 1997]. Cependant, nous avons choisi de ne pas les aborder dans

cette thèse car elles reposent sur des approches d'adaptation de domaine, dont nous avons eu un aperçu des limitations. Dans ces approches, une fois les liens entre les différentes classes estimées, la prédiction repose également sur une règle de fusion bayésienne.

À partir de ces enseignements, nous proposons une méthode d'estimation de transitions indépendante des données de référence tout en suivant un grand nombre de règles détaillées ci-dessus.

7.2 Modélisation des transitions de classes

Dans notre cas d'application, l'historique pourrait fournir deux informations très utiles. La première concerne les données météorologiques. Ces données pourraient représenter un facteur de similarité entre deux années afin de guider le choix d'un classifieur naïf, par exemple. Devant l'absence de lien évident entre performances de la classification naïve et un lien de similarité entre les séries temporelles (les performances de la table 3.1 ne sont pas symétriques), nous avons mis de côté l'exploitation des données météo, pour nous intéresser à l'évolution de l'OCS au fil des ans.

Il faut cependant noter que les données de référence que nous exploitons sont éparées et géographiquement instables d'une année sur l'autre. Ainsi nous n'avons pas la stabilité disponible dans le RPG par exemple, qui permet d'estimer les transitions entre classes depuis les données de référence. Nous proposons alors de modéliser les transitions entre classes directement depuis les cartes supervisées disponibles dans l'historique.

7.2.1 Principe

Les cartes d'OCS supervisées permettent d'établir l'évolution des classes au fil des ans, avec un taux d'erreur de classification faible. Cette évolution traduit les changements de classes apparaissant en fonction des cycles naturels ou artificiels (activité humaine). Les cartes fournissent cette information pour chaque pixel de la zone d'étude, permettant une modélisation au cas par cas des transitions entre les classes.

Nous exploitons la [Matrice de Transition \(MT\)](#) qui recense toute transition de classes détectée dans les cartes supervisées. Dans cette matrice chaque colonne représente une liste exhaustive des classes de la nomenclature. Chaque ligne représente la succession des classes observées dans les cartes supervisées. La taille (le nombre d'années) de l'historique définit le nombre de classes successives vues.

Construction de la matrice de transition

La première étape de l'estimation des transitions consiste à trier les n cartes supervisées présentes dans l'historique. La séquence de valeurs ainsi obtenue représente les changements d'OCS d'un pixel au fil des ans.

Ainsi, les $n - 1$ plus anciennes valeurs définissent la clef de transition k_i , il s'agit d'une simple succession des étiquettes. La valeur la plus récente représente la classe c_j succédant à cette clef de transition. k_i et c_j définissent une ligne et une colonne de la MT. La table 7.1 offre un exemple simplifié de MT. En parcourant l'intégralité de l'emprise de la carte, nous obtenons une matrice d'occurrences de chaque transition observée.

Pour passer des occurrences à des probabilités, deux possibilités sont envisageables : normaliser chaque cellule de la matrice par le nombre total d'éléments de la matrice, soit normaliser par la somme des lignes. Si nous divisons par le nombre total des éléments, alors nous obtenons des probabilités de transition multipliées par la probabilité a priori de chaque transition. En divisant par la somme des lignes, nous obtenons des probabilités conditionnées par la séquence observée dans l'historique. Au vu de l'utilisation de la MT que

nous réalisons, ces deux calcul de probabilités sont équivalents, car nous sélectionnerons le maximum sur une ligne, ce qui correspondra la classe la plus probable pour une transition observée.

La MT associe à chaque succession de classes une probabilité de transition vers chaque successeur potentiel. La nature des classes impacte les valeurs de probabilités. Les classes géographiquement stationnaires (forêts, eau, etc) possèdent des probabilités fortes vers elles-mêmes et très faibles vers les autres classes. À l'inverse, pour les classes soumises à des changements géographiques annuels, comme la rotation des cultures, les probabilités sont distribuées entre les différents acteurs de la rotation. Le principe accompagnant la MT consiste à pouvoir exploiter la détection des cycles de transitions lors de la prédiction. Cette approche permet d'envisager une production du millésime courant dès la production du millésime précédent ce qui représente un avantage conséquent pour la production opérationnelle.

Exemple d'utilisation de la matrice de transition

TABLE 7.1 : Exemple d'une matrice de transition simplifiée utilisant 4 années d'historique.

Transition \ Classe	Classe		
	1	2	3
1 → 1 → 1	0.20	0.05	0
1 → 2 → 1	0.15	0.10	0
2 → 1 → 2	0.05	0.20	0
3 → 3 → 3	0.01	0	0.24

Pour illustrer le fonctionnement de la MT, considérons l'exemple suivant. Soient 4 années successives constituant l'historique de 2007 à 2010, avec 3 classes numérotées de 1 à 3. L'objectif étant de classer 2011. La clef de transition possède 3 (4-1) classes, on parle ici de la taille de la clef de transition.

Dans cet exemple, les classes 1 et 2 sont des cultures annuelles. Deux comportements caractérisent la classe 1, soit elle est stationnaire (comme le Maïs par exemple) ou alors elle suit une rotation de culture binaire avec la classe 2 (l'une succède à l'autre d'une année sur l'autre). Nous simulons volontairement cette double évolution possible pour la classe 1 afin de montrer les différentes clefs de transition que cela implique dans la matrice. Nous émettons alors l'hypothèse suivante : dans le cas d'une rotation de culture, l'alternance des classes impliquées se poursuit au fil des années futures.

La classe n°3 est une classe pérenne simulant, par exemple, la classe de Surfaces Artificielles. La table 7.1 présente la MT estimée dans cet exemple, les valeurs sont fixées arbitrairement. Cette matrice possède donc 4 clefs de transition, représentant les différentes transitions possibles, définies par cet exemple. Nous observons ici deux échelles de valeurs, des probabilités hautes (> 0.15) représentant le comportement attendu et des probabilités basses (< 0.1) représentant des anomalies, dues par exemple à des erreurs de classification. En conditions réelles, des grands écarts de probabilités peuvent être observés entre les transitions impliquant des classes majoritairement représentées (Blé, Maïs, etc) et celles peu représentées (Soja, Orge, etc).

Pour l'estimation de la MT nous disposons donc de 4 cartes supervisées dans l'historique. Néanmoins, lors du processus de décision, nous utilisons les trois plus récentes pour générer la clef de transition. En effet, nous supposons la stationnarité des transitions : une transition ne dépend pas des années utilisées pour l'estimer. Elles témoignent en effet d'événements produits et qui peuvent réapparaître au gré des années.

Par contre, il est primordial d'assurer la continuité chronologique entre les millésimes utilisés pour la production et le millésime souhaité. Par exemple, la succession d'années 2008 → 2009 → 2010 constitue un usage correct de la MT pour prédire la carte de 2011, mais incorrect pour prédire celle de 2012. La clef de transition permet d'identifier la ligne correspondante dans la MT, l'étiquette de la classe prédite correspond à la plus haute valeur de probabilité détectée. De la même manière que pour l'estimation de la MT, la prédiction s'effectue pour chaque pixel de la carte.

7.2.2 Expérimentations

Protocole

L'objectif de cette méthode est de produire une carte d'OCS, sans utiliser une seule image de l'année correspondante, mais uniquement à partir des cartes supervisées des années passées. La MT est totalement indépendante des données ayant servi à son estimation. En effet, elle ne possède aucune information liée aux cartes, car elle ne contient que des successions de classes et des probabilités. Néanmoins, une MT est rattachée à une zone d'étude car elle transcrit les transitions propres à celle-ci.

Notre protocole d'estimation à partir des cartes supervisées nous permet d'obtenir des clefs de transitions dont la taille varie entre 1 (2 années d'historique) et 5 (6 années d'historique). Nous pouvons donc produire une carte à partir d'une année d'historique, sous réserve d'avoir accès à une MT. Dans cette expérimentation nous distinguons volontairement l'étape d'estimation de la MT et son utilisation. En effet, nous avons vu qu'il existe plusieurs approches permettant d'estimer une telle matrice, et nous ne fermons pas la porte à de telles perspectives.

Performances

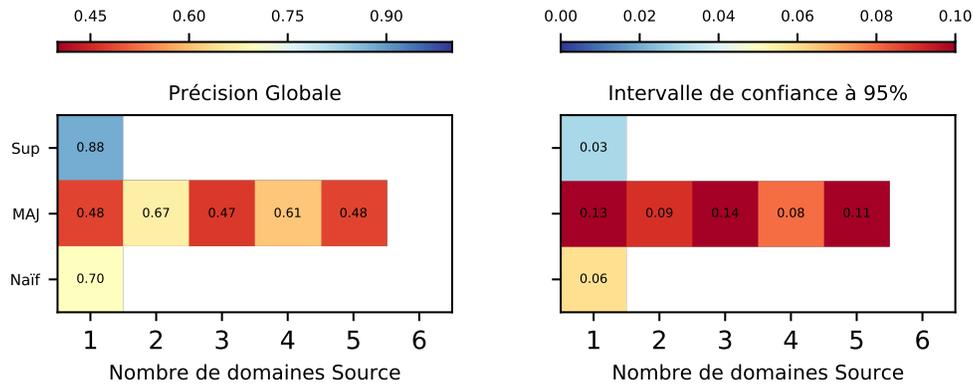


FIGURE 7.1 : PG et intervalles de confiance à 95 % obtenus par la prédiction à partir des probabilités de transition (MAJ).

La figure 7.1 présente la PG (gauche) et les intervalles de confiance à 95 % (droite) obtenus pour les différents cas d'utilisation de la matrice de transition. Nous présentons ici des résultats sous la forme de matrice. L'axe des abscisses indique le nombre d'années requises dans l'historique pour exploiter les méthodes indiquées en ordonnées. L'échelle de couleur varie du rouge (PG < 45 %) à bleu (PG > 90 %). La couleur

blanche implique que la méthode n'est pas applicable pour cette combinaison de paramètres. L'échelle de couleur est inversée pour les intervalles de confiance, car plus l'intervalle est petit, plus la stabilité de la méthode est grande. Ainsi, la figure 7.1 indique des performances irrégulières, qui oscillent alternativement entre $\pm 50\%$ et 65% . La faible performance du cas n'exploitant qu'une année demeure prévisible.

En effet, dans ce cas, nous n'avons toujours le choix que d'une seule classe, car la clef de transition est unique. Cependant, cette configuration n'autorise aucune flexibilité vis-à-vis des variations susceptibles d'arriver dans l'alternance des classes. Les autres performances, certes plus complexes à appréhender, constituent cependant une surprise, en particulier ce qui concerne le comportement alternatif des performances. Nous pouvons tout de même constater que pour 2 années de l'historique, les performances se situent dans l'intervalle de confiance de la classification naïve.

Pour comprendre ces variations dans les performances, nous analysons les performances par classes indiquées dans la figure 7.2. Cette figure représente une matrice dans laquelle l'axe des abscisses recense les 17 classes d'OCS étudiées, et l'axe des ordonnées le FScore pour le cas naïf (en bas) et supervisé (en haut), indiquant les bornes hautes et basses, et les différentes tailles d'historique disponibles (1 à 5) ajouté en suffixe du nom de la méthode (MAJ).

Cette figure éclaire sous un angle nouveau le comportement de cette méthode de mise à jour d'une carte existante. Effectivement, nous constatons que les classes géographiquement stables sont très bien classées avec des FScore toujours supérieurs à 75% . Pour ces six classes (Feuillus, Conifères, Maïs, Rivières, Lacs et Gravières) nous reproduisons les effets de la procédure SAP qui constituent le gain obtenu systématiquement dans les cas du **Transport Optimal (TO)**. La différence réside dans l'absence de classification de l'année courante puisque nous n'utilisons pas les images de celle-ci. Pour ces six classes, l'application des transitions peut s'apparenter à un simple report des classes de l'année précédente. Le fait que ces 6 classes soient toujours correctement prédites explique la limite minimale de PG aux alentours de 50% . En effet, il s'agit de classes dont les proportions sont importantes dans la zone d'étude comme nous l'avons vu dans la table 2.1. La classe de Surfaces Artificielles, comme précédemment, se démarque du comportement des autres classes pérennes, pour les raisons déjà évoquées (fusion de classes très différentes lors du choix de la nomenclature).

Nous expliquons ce phénomène de performances alternatives de la méthode de mise à jour grâce au FScore de cinq classes. Celles-ci sont toutes impliquées dans la rotation des cultures : Blé, Colza, Tournesol, Soja et Chanvre. En effet, pour les deux cas d'augmentation de PG, on constate le même effet sur le FScore de ces classes. Le Blé semble profiter le plus de cette approche, probablement en raison de sa forte présence dans la zone d'étude et sa très bonne prédiction lors des classifications supervisées. En revanche, pour les autres classes, le gain observé demeure moindre. Les performances obtenues mettent en évidence la complexité de la modélisation des rotations de culture. Les performances indiquent que dans notre application, la modélisation la plus adaptée requiert une clef de transition de taille 2. En effet, dans notre zone d'étude, les transitions les plus exploitées cultivent deux classes alternativement. De fait, nous avons besoin de trois ans pour détecter une telle transition. Cependant, de précédents travaux démontrent que certaines transitions s'étalent jusqu'à 4 années [Osman, 2015], ce qui explique que les performances entre les tailles de clef de transition 2 et 4 soient différentes.

Néanmoins, les erreurs accumulées de classifications constituent une explication valable à la dégradation des performances entre ces deux cas, que l'on peut cumuler avec la dynamique des transitions observées. En effet, bien que faiblement présentes dans les classifications supervisées ($< 10\%$), lorsque nous créons la série temporelle de cartes d'OCS, le taux d'erreur devient proportionnel au nombre de cartes exploitées. Ce

problème trouve son origine dans le caractère aléatoire des erreurs générées. En effet, on constate que d'un millésime à un autre, les erreurs ne sont pas obligatoirement localisées aux mêmes coordonnées. Les erreurs de classification entraînent une multiplication des transitions possibles dans la MT, impactant la pertinence de celles-ci, de même que la capacité de généralisation de la MT et également le temps de calcul requis.

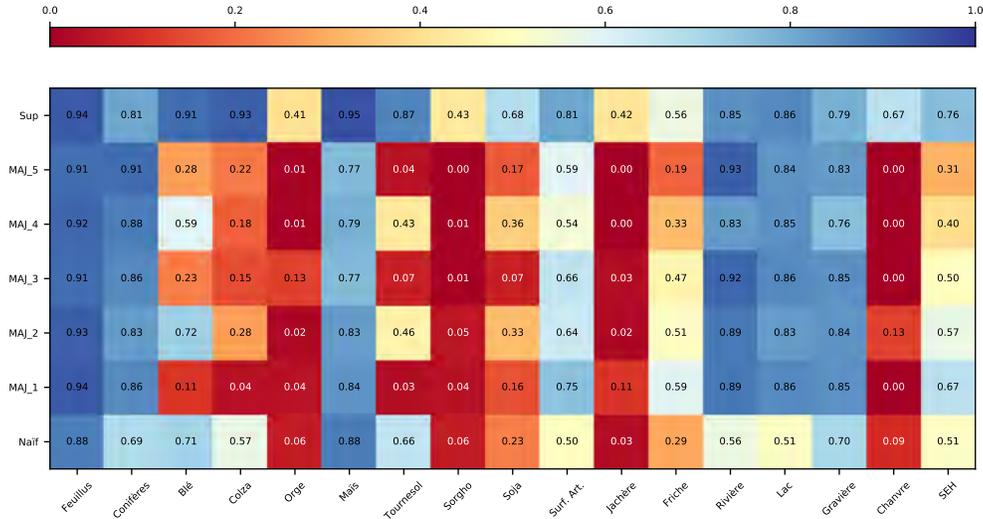


FIGURE 7.2 : FScore par classe obtenus lors de la production de la carte en exploitant les matrices de transition.

7.2.3 Limitations et perspectives d'amélioration

La préservation de la continuité temporelle représente un des points les plus importants dans l'utilisation de la MT. Cependant, aucune contrainte ne relie les millésimes utilisés pour l'estimation de la MT et le millésime que l'on cherche à produire. Ce concept est très important, car il permet d'exploiter une MT provenant d'une source annexe. En effet, sous réserve que la nomenclature ainsi que la taille de la clef de transition soient cohérentes, aucun élément n'interdit d'exploiter une matrice déterminée par exemple à partir des données de référence ou de CLC, plutôt que des classifications. Elle permet également l'utilisation d'une matrice estimée à partir de cartes exploitant une résolution ou un capteur différent, car l'information de transition ne dépend pas d'un pixel mais d'une succession de classes.

Les résultats obtenus mettent en évidence, qu'une clef de transition de taille 2 correspond le mieux à notre zone d'études, car nous étudions majoritairement des rotations tri-annuelles. Les performances notamment sur les classes de cultures montrent que nous arrivons à prendre en compte ces rotations à partir des cartes produites par classification.

Cependant les performances obtenues ne représentent pas une solution viable pour la production d'une carte d'OCS sans utiliser les images et les données de référence de la période correspondante. À partir de nos résultats, nous pouvons identifier plusieurs sources d'erreur expliquant les performances.

L'absence totale de guide lors de la prédiction se ressent dans la faible capacité à prendre une décision valide. Dans les travaux présentés en introduction de ce chapitre, la matrice de transition s'associe à un

classifieur (réseau bayésien, champs de Markov, réseau de neurones), qui offrent une fonction de décision affinée.

L'accumulation des erreurs de classification représente un problème majeur. En effet, dans cette approche chaque succession de classe implique une transition unique. La moindre erreur de classification cause l'apparition d'une nouvelle transition. L'introduction d'un paramètre de flexibilité peut permettre d'améliorer les performances, que ce soit lors de la création de la matrice ou de son utilisation. Dans cette optique, la présence des bords de parcelles mais surtout les pixels mal classés, qui participent à l'estimation des transitions constituent un facteur d'erreur important, car leur classification reste toujours problématique même dans le cas supervisé. Nous pourrions contourner ce problème à partir de trois approches différentes. La première consisterait à exploiter des données de référence, donc des polygones représentatifs du centre des parcelles. Cette solution pourrait permettre d'éliminer une grande partie des erreurs induites par les bords de celles-ci. Dans ce cas précis, nous reviendrions aux méthodes déjà proposées et implémentées, mais également plus complexes, présentées dans la première partie de ce chapitre (CropRota, CarrotAge). Une deuxième solution consisterait à effacer les pixels de bords en recourant à un algorithme de segmentation ou un opérateur morphologique permettant d'opérer un lissage sur la classification pixel, et ne calculer les transitions non plus à l'échelle d'un pixel mais d'un objet. Cela permettrait également d'éliminer le cas de pixels isolés résultant d'un artefact (nuage mal détecté, etc). Enfin, une troisième solution, compatible avec les deux précédentes, consisterait à exploiter la carte de confiance afin de filtrer les pixels non fiables.

Un autre problème potentiel provient du décalage entre les millésimes utilisés pour estimer la MT et ceux utilisés lors de la prédiction. En effet, certaines classes n'étant pas récurrentes annuellement, cette situation peut provoquer une absence de la transition dans la matrice (par exemple le Chanvre ou les Jachères). De la même manière, une transition impliquant ces classes, ne sera jamais vue lors de la prédiction si celles-ci n'existent pas dans les cartes supervisées. Nous avons contourné ce problème, en supposant que les transitions sont stationnaires dans le temps, cependant cette hypothèse ne peut être vérifiée.

Pour remédier à ce problème et autoriser un peu de souplesse dans la prise de décision, nous pourrions envisager un algorithme de recherche moins strict en autorisant une fausse mise en correspondance. Ce type d'approche permettrait de sélectionner une transition selon un pourcentage de correspondance entre la clef recherchée et celles existantes. Par exemple, on cherche une transition de la forme $1 \rightarrow 2 \rightarrow 3 \rightarrow 1$ qui n'existe pas dans la MT. Cependant, seules les transitions $3 \rightarrow 1 \rightarrow 2 \rightarrow 1$ et $1 \rightarrow 2 \rightarrow 3 \rightarrow 2$ existent dans la MT. Une forte contrainte sur l'exactitude de la correspondance ne permet pas de sélectionner une transition. Relâcher la contrainte, autorisant une classe différente dans la clef de transition, permet de sélectionner la transition $1 \rightarrow 2 \rightarrow 3 \rightarrow 2$. Dans le cas où plusieurs transitions correspondent, un calcul pondéré par le degré de correspondance entre les transitions pourrait être réalisé.

Dans le même esprit, au lieu de générer une MT dans laquelle toutes les clefs de transitions possèdent une taille identique, nous pourrions envisager de recenser toutes les transitions dont la taille varie entre 1 et le nombre maximal fixé par l'historique ou l'utilisateur. Dans ce cas d'utilisation, un pixel est associé à au moins une transition, et la décision est prise par une fonction d'inférence bayésienne ou un vote pondéré.

Ces approches que nous n'avons pas pu aborder au cours de cette thèse permettraient de limiter l'impact des erreurs sur la prédiction, tout en offrant une prédiction plus fine pour la production de la carte. Nous utiliserons la MT dans la suite des travaux comme source de secours en cas d'égalité dans un vote.

7.3 Modélisation des signatures temporelles

La modélisation des transitions s'avère insuffisante pour prédire efficacement une carte d'OCS. Cependant, elle met en évidence l'intérêt d'exploiter plusieurs domaines Source. Or, l'historique contient également toutes les données originales : images de réflectance, données de référence, etc. Ainsi, nous pouvons exploiter ces informations pour mieux appréhender les distorsions entre les domaines. Dans de précédents travaux, [Flamary et al., 2015] ont proposé d'exploiter un classifieur (SVM) unique entraîné avec des échantillons provenant de différentes périodes. Ils ont montré que cette approche était aussi performante que des approches d'ensembles de classifieurs et d'apprentissage semi-supervisé. Les auteurs exploitent 8 années de profils temporels de NDVI, calculé à partir d'images MODIS. Cependant, ils cherchent à différencier deux classes, les Feuillus et les Pins, sur une zone forestière, alors que nous souhaitons exploiter 17 classes. Il devient alors intéressant d'évaluer cette approche sur notre jeu de données. Nous exploitons ce principe pour directement combiner les données de l'historique sans distinction entre les domaines Source.

7.3.1 Le classifieur unique

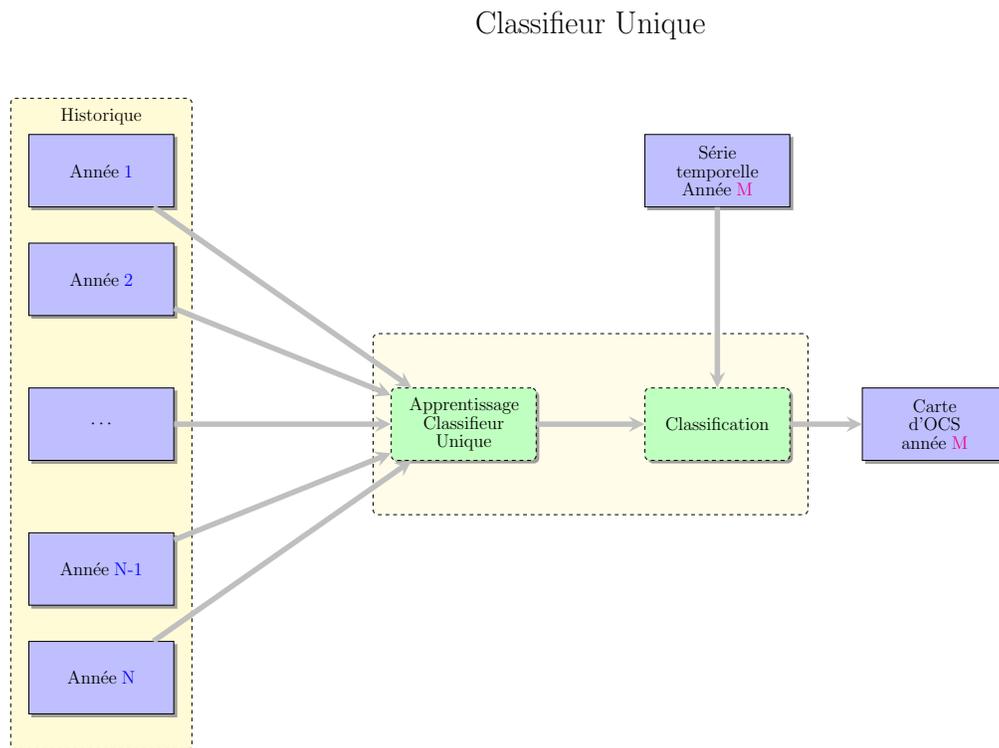


FIGURE 7.3 : Conception de la chaîne de traitement pour le classifieur unique.

Nous avons montré que chaque classifieur naïf possède ses propres caractéristiques, permettant ainsi une variation dans la qualité de prédiction de chaque classe. Ces différences proviennent des échantillons du domaine Source utilisé lors de l'apprentissage du classifieur. Le principe du classifieur unique consiste à laisser celui-ci déterminer les caractéristiques les plus pertinentes dans chaque domaine afin d'entraîner un

classifieur plus performant que chaque classifieur naïf. Ainsi, nous entraînons un classifieur associé à plusieurs domaines Source. Nous baptisons cette approche classifieur unique, puisqu'un seul classifieur est entraîné. Nous pouvons envisager le classifieur unique comme une amélioration de la classification naïve, il constitue alors un classifieur "multi-naïf". Les pré-requis sur les données s'avèrent identiques à ceux de la classification naïve, qu'il s'agisse des séries temporelles ou des données de référence. Au moins deux jeux d'apprentissage entraînent le classifieur provenant de deux périodes différentes de l'historique. Si on ne considère qu'une seule année, alors il s'agit de la classification naïve.

Chaque période de l'historique utilisée pour l'apprentissage représentera une évolution (un comportement) des classes d'OCS. En effet, le climat annuel (température, précipitations) impacte le comportement des classes de végétation. Ainsi, la phase d'apprentissage du classifieur unique permet de laisser le RF choisir parmi les différentes périodes les primitives discriminantes, favorisant une séparation optimale des classes. Le classifieur ainsi entraîné possédera une forte capacité de généralisation, capable de s'adapter aux variations de comportement dues au climat.

7.3.2 Résultats

Dans cette section, nous allons présenter les performances obtenues par le classifieur unique lors de l'utilisation de toutes les combinaisons de domaines Source possible. À partir des données déjà préparées dans le cadre des autres études présentées dans ce manuscrit, l'apprentissage du classifieur unique devient transparent. Il suffit de fournir un jeu d'échantillons d'apprentissage pour chaque domaine Source en entrée de la chaîne de production naïve, présenté dans la figure 3.1. L'harmonisation des données grâce au ré-échantillonnage temporel 2.1 lève d'éventuels conflits entre les primitives de chaque domaine. La phase de classification ne change en rien par rapport à la classification naïve, qui ne concerne qu'un domaine cible.

Pour cette expérimentation, nous utilisons une approche de un contre tous, où une année parmi les sept disponibles constitue le domaine Cible et les autres le domaine Source. Nous pouvons également faire varier le nombre de domaines Source en effectuant toutes les combinaisons possibles de taille variant entre 2 et 6.

La figure 7.4 présente la PG et les intervalles de confiance obtenus. Ces performances se révèlent très intéressantes, partant du cas naïf à 0.70, le cas 2 offre déjà une amélioration de près de 6 % avec des intervalles de confiances minimales. Nous constatons que l'ajout d'un nouveau domaine Source améliore les performances obtenues par le classifieur unique. Nous pouvons également noter le dépassement du seuil des 80% dès le cas 4 (4 domaines Source).

L'analyse du FScore s'avère plus intéressante. En effet, comme le montre la figure 7.5, toutes les classes bénéficient de l'augmentation du nombre de domaines Source.

Seules 4 classes restent complexes à prédire : l'Orge, le Sorgho, les Jachères et le Chanvre. Ces 4 classes sont déjà difficilement prédites par le cas naïf d'une part, mais également par le cas supervisé avec des FScore inférieurs à 50% d'autre part. Nous avons déjà abordé la question de ces classes lors de la présentation des performances du cas naïf : les confusions entre classes similaires expliquent ces chutes de performances. Nous pouvons néanmoins distinguer le Chanvre des trois autres classes. En effet, celui-ci bénéficie d'une amélioration importante entre le cas naïf et le cas 6 passant de 0.09 à 0.26 de FScore. Cette amélioration met en évidence qu'il souffre essentiellement d'un manque de représentation dans les échantillons d'apprentissage : il est à la fois absent de certains domaines et peu représenté dans les autres.

Sans égaler les performances de la classification supervisée, le classifieur unique fournit de bonnes performances, que ce soit en termes de PG mais également de FScore. On peut expliquer ces performances par

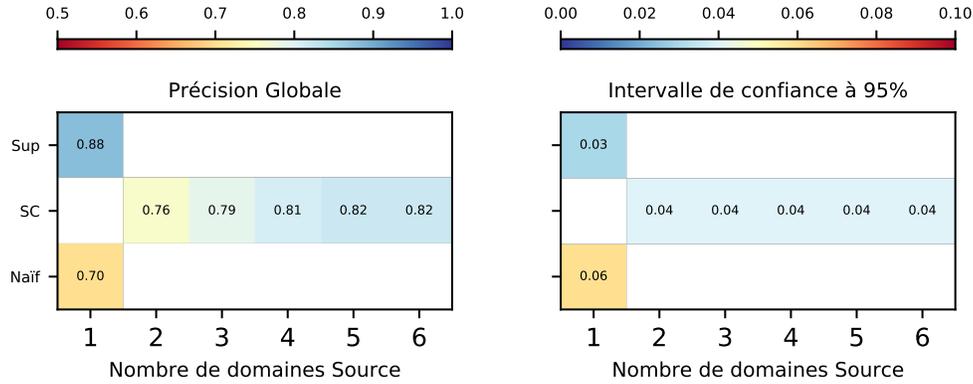


FIGURE 7.4 : PG du classifieur unique.

la capacité du **RF** à sélectionner les primitives les plus utiles à la séparation des échantillons des classes différentes. En offrant une vue sur les distorsions existantes entre les années, nous permettons au **RF** de les apprendre afin de s'adapter lors de la rencontre d'une distorsion d'une ampleur différente. Cela laisse supposer que les distorsions proviennent d'une même origine entre les domaines, et que seules leurs amplitudes varient. Ainsi la forte diversité fournie à partir du cas 4 permet au classifieur unique d'obtenir une bonne capacité de généralisation.

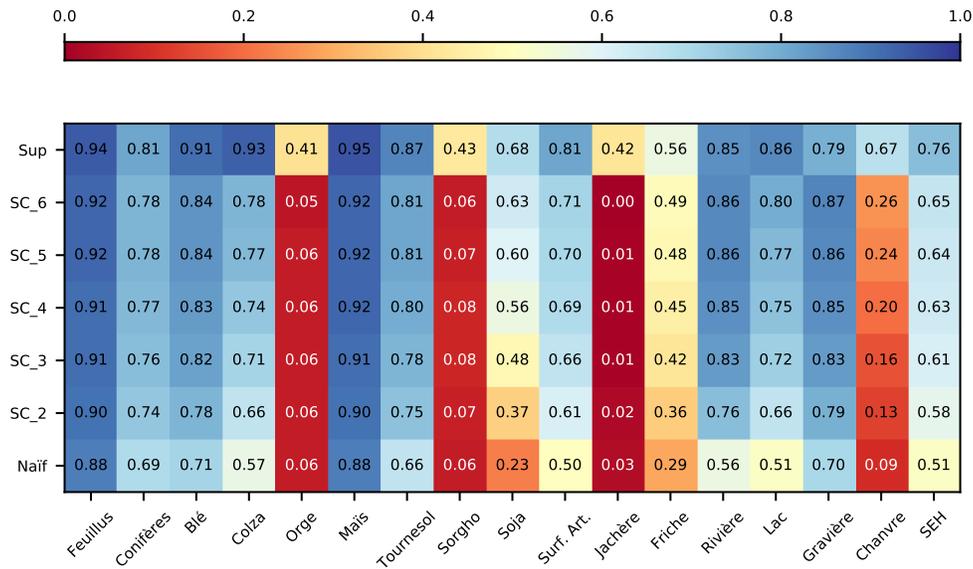


FIGURE 7.5 : FScore par classe obtenus par le classifieur unique pour différentes tailles d'historique. La classification naïve et supervisée sont indiquées pour information.

Cependant, ses bonnes performances représentent un coût de traitement considérable. Il convient d'harmoniser toutes les données afin de supprimer les conflits entre primitives susceptibles de ne pas correspondre. La

plupart des méthodes proposées dans ce manuscrit impliquant cette phase d'harmonisation, celle-ci n'entre donc pas dans un compte négatif. De plus, le coût de la phase d'apprentissage s'avère proportionnel au nombre de domaines Source utilisés. En effet, la diversité introduite présente un avantage majeur pour la généralisation des arbres du [RF](#), mais implique une augmentation de la quantité d'échantillons, ainsi que potentiellement la profondeur des arbres. La profondeur impacte directement le temps d'apprentissage, mais également de classification, puisque le temps de parcours dans un arbre dépend de ce paramètre. Le classifieur unique s'apparente à une fusion "précoce" des données, car elle est réalisée en amont de l'apprentissage. Dans le chapitre suivant, nous proposons d'exploiter une fusion tardive, après classification.

La fusion de classifications

Les performances obtenues par le classifieur unique confirment que l'exploitation de plusieurs domaines Source constitue un avantage pour la production de cartes d'OCS, sans utilisation des données de référence de l'année produite. Néanmoins, nous pouvons constater que, dans le cas des classes mal représentées, le classifieur unique ne permet pas de reproduire les bonnes performances observées via la classification naïve. Cet effet provient du fait que le classifieur unique ne reproduit pas l'ensemble des classifieurs naïfs possibles. Ainsi, nous ne pouvons pas profiter de paires de domaine favorables, visibles dans la figure 3.2. Pour maximiser l'exploitation de tels cas, nous proposons l'utilisation d'approches de fusion de classifications. Au cours de chapitre, après une rapide introduction à la théorie des votes et ses intérêts pour la fusion de classifications, nous mettons en oeuvre plusieurs systèmes de vote de complexités variables. Dans une seconde partie, nous proposons d'exploiter la fusion par Dempster-Shafer, en estimant les performances associées à chaque classification naïve utilisées dans la fusion. Enfin, nous proposons une approche de fusion bayésienne permettant d'exploiter des informations supplémentaires, autres que les classifications naïves.

8.1 Les systèmes de vote

Les premiers témoignages de votes remontent à l'antiquité grecque avec l'apparition de la démocratie. L'étymologie de ce mot, *dêmos* ("le peuple") et *krátos* ("la puissance, le pouvoir") indique que les décisions sont prises par le peuple. Les Grecs fondaient leur système sur la sagesse de la communauté pour prendre les meilleures décisions qui gouverneraient leurs vies. Ce système a considérablement évolué aux cours des siècles, attirant l'intérêt de nombreux mathématiciens notamment sur la meilleure prise en compte des divergences d'opinions et des préférences au sein de la population.

8.1.1 Généralités

Il existe deux grandes théories des systèmes de vote qui ont marqué l'histoire : la méthode de Condorcet et la méthode de Borda. Ces deux mathématiciens du XVIII^e siècle sont considérés comme les piliers fondateurs de la théorie des votes. Ces théories sont fondées sur l'impossibilité de connaître la préférence de la majorité lorsqu'elle choisit un candidat.

Le jury de Condorcet

Les travaux de Condorcet avaient trait à l'application juridique. La question n'était plus de savoir quel candidat avait la préférence des votants, mais plutôt de connaître le taux d'électeurs estimant une situation vraie ou fausse.

Ces travaux ont mené à la constitution des jurys populaires en lieu et place d'un unique magistrat pour établir un verdict. Il a en effet montré que plus le nombre de votants était important, plus la décision prise prenait en compte les faits et non pas uniquement l'intime conviction personnelle. Ces constatations l'ont conduit à prouver que le vote majoritaire simple pouvait ne pas être représentatif de l'opinion des votants. Il propose alors de demander aux électeurs d'établir une liste préférentielle de candidats. Ainsi pour un scrutin à 3 candidats (a, b, c) impliquant 60 votants, nous obtenons les résultats suivants, classés par ordre de préférence :

- 20 votants : a, b, c
- 17 votants : c, b, a
- 10 votants : b, c, a
- 3 votants : b, a, c

Selon la règle du vote majoritaire le vainqueur est le candidat a avec 20 voix, suivi par c avec 17 voix et b avec 13 voix. Cependant, ce résultat ne reflète pas l'opinion des votants. En effet, si l'on compare le nombre fois où un candidat est placé en tête par rapport à un autre on obtient :

- 23 votants préfèrent a à c contre 27 préférant c à a
- 33 votants préfèrent b à c contre 17 préférant c à b
- 20 votants préfèrent a à b contre 13 préférant b à a

Ce qui conduit au classement suivant : b, c, a , ce qui s'avère l'exact opposé du résultat par vote majoritaire. Il s'agit du paradoxe de Condorcet. Ainsi, la seule solution permettant de prendre en compte l'opinion des votants consiste à réaliser une comparaison par paire de tous les candidats, comme nous venons de faire. Néanmoins, il paraît évident que cette solution soit difficile à mettre en place en situation réelle devant le nombre de combinaisons possibles pour des votes avec de nombreux candidats. De fait, en règle générale, la méthode de Condorcet est couplée avec celle de son grand rival de l'époque : Borda.

Le comptage de Borda

La méthode de Borda reste très utilisée encore à l'heure actuelle. Elle s'applique très bien aux élections menant à l'attribution de points (comme l'eurovision) ou de sièges comme les élections européennes. La motivation de Borda est la même que celle de Condorcet, il souhaite prendre au maximum en compte l'opinion des votants lors de la prise de décision. Pour cela il réduit le problème en fixant un nombre n inférieur au nombre de candidats potentiels. Chaque votant classe par ordre de préférence les candidats, comme précédemment. Lors de la comptabilisation des voix, le candidat en première position reçoit n points, le suivant $n - 1$ et ainsi de suite. Les candidats classés à partir de la $n^{\text{ième}}$ position reçoivent un score de 0. Le résultat du vote est alors la somme des poids pour chaque candidat. Dans l'exemple précédent, pour $n = 2$ le résultat du vote est présenté dans la table 8.1, les résultats obtenus sont les mêmes. Cependant, il est fréquent qu'ils soient différents, alimentant le débat sur la méthode la plus juste. Dans le cas où $n = 1$, nous retrouvons le cas du vote par majorité.

TABLE 8.1 : Exemple de pondération des votants par la méthode de Borda.

Poids	Candidats			Total
	2	1	0	
a	20	3	17+10	43
b	13	37	0	63
c	17	10	23	44

De l'opinion à la classification

Les méthodes de Borda et Condorcet reposent sur l'opinion d'un acteur humain qui donne son avis sur une question précise. La transition vers l'AA s'avère simple, en remplaçant les électeurs par des classifieurs, chacun apportant une voix à la classe choisie. En effet, un classifieur produit souvent un résultat qui appartient à un ensemble fini de classes $C = 1, \dots, c_n$. Les règles que nous venons d'introduire s'appliquent alors très simplement au résultat d'une telle classification.

Ainsi, l'objectif devient alors d'exprimer la diversité des opinions possibles grâce à des algorithmes d'AA. Ceux-ci doivent respecter les contraintes d'indépendance exigées par Condorcet et intégrer un système capable de centraliser les différents résultats de classification : on parle de système de fusion de classifieurs.

8.1.2 La fusion de classifieurs par vote

Rencontrer une étape de fusion de classifications en post-traitement s'avère assez fréquent. En effet, elle permet de combiner les décisions issues de différents classifieurs. Il devient alors possible de séparer la tâche de classification en réduisant le volume de données traitées par chaque classifieur, ou de laisser libre cours à différents classifieurs pour choisir les primitives utilisées. La complexité de la règle de fusion permet d'exploiter simplement les étiquettes ou alors d'associer une pondération afin d'influencer la décision du vote.

On retrouve également les votes dans les ensembles de classifieurs, afin de choisir l'étiquette associée à l'échantillon en fonction des résultats des différents classifieurs de l'ensemble. Par exemple, le RF exploite un [Vote Majoritaire \(VM\)](#) pour fusionner la décision de chaque arbre.

Les systèmes de vote présentent de nombreux avantages :

- Une indépendance vis-à-vis du volume de données.
- L'assurance d'une meilleure performance par rapport à chaque classifieur individuellement.
- L'utilisation de classifieurs faibles, souvent excellents en fouille de données.
- La possibilité de réunir des classifieurs spécialisés.
- Une grande liberté dans la fonction de décision.

L'indépendance vis-à-vis du volume de données

La quantité d'échantillons d'apprentissage disponibles constitue un facteur important dans le cadre de la classification, dans la mesure où le choix du classifieur utilisé dépend souvent de celui-ci. La grande liberté des systèmes de vote permet de gérer indifféremment ces problèmes car il exploite les résultats de classification.

Dans le cas d'un jeu d'apprentissage peu fourni, l'utilisation des approches de *bootstrapping* (comme le *bagging*) assure la diversité des différents classifieurs, tout en excluant les configurations médiocres [Hashem, 1997, Marcialis and Roli, 2003].

Au contraire, dans le cas d'un jeu d'apprentissage volumineux, une approche consiste à réaliser un partitionnement de l'espace des primitives [Rastrigin and Erenstein, 1981] afin que chaque classifieur ne traite qu'une seule partition. Une règle de combinaison [Polikar, 2012] permet alors de fusionner les décisions de chaque classifieur.

L'assurance d'une meilleure performance

Nous retrouvons ici une propriété partagée avec les ensembles de classifieurs et notamment le RF : le respect de la règle d'indépendance entre les votants permet d'assurer que le résultat du vote sera toujours meilleur que le résultat de chacun des votants pris individuellement [Clemen, 1989, Xu et al., 1992]. Ce prédicat a été prouvé analytiquement par [Tumer and Ghosh, 1996] dans le cadre de la fusion par un VM de réseau de neurones linéaires.

Les classifieurs faibles, des ressources à exploiter

Les classifieurs dits "faibles" comme les arbres de décisions, constituent des alternatives intéressantes pour la fouille de données. En effet, ils permettent d'identifier les primitives discriminantes, mais ne disposent pas de capacités de généralisation suffisantes les rendant rapidement peu performants. Cependant, dans le cadre d'un système de vote ils représentent une source d'informations très utile, car la composition des primitives exploitées permet de moduler l'influence de ce classifieur lors de la prise de décision [Quinlan, 1993]. Nous avons déjà abordé l'utilité des classifieurs faibles lors de la présentation des RF dans le chapitre 2.2.1.

La spécialisation de classifieur

Généralement, dans le cadre d'une classification à plusieurs classes, nous essayons d'éviter la spécialisation du classifieur. On parle de classifieur spécialisé lorsque celui-ci est très performant pour la détection d'une seule classe. En effet, ce genre de spécialisation se fait au détriment de toutes les autres classes. Cependant, dans le cadre d'un système de vote, spécialiser un classifieur peut représenter un avantage. En effet, nous

savons que certaines configurations offrent à des classifieurs la possibilité de reconnaître très précisément une classe donnée au détriment des autres [Wolpert, 2002]. Une règle de fusion exploitant cette information lors de la prise de décision, permet alors d'extraire le meilleur de chaque classifieur. Cette approche repose donc sur une complémentarité forte entre les différents classifieurs inclus dans le vote.

La règle de décision

La règle de fusion constitue le cœur d'un système de vote, permettant de rassembler les différents résultats des classifieurs afin de fournir une étiquette à un échantillon. On retrouve trois catégories de fonction de décision :

1. La fusion des étiquettes, chaque électeur vote pour une seule classe.
2. L'utilisation d'une fonction de support, chaque électeur associe un poids à son vote.
3. L'utilisation d'un classifieur entraîné sur la sortie des classifieurs électeurs.

Les règles de fusion d'étiquettes représentent les approches les plus simples. Elles dépendent de la notion de majorité, divisée en trois règles distinctes :

- L'unanimité : tous les votants indiquent la même étiquette
- La majorité simple : plus de la moitié des votants ont choisi cette étiquette
- Le **Vote Majoritaire (VM)** : l'étiquette est choisie par le plus de votants

Ces règles rappellent la situation initiale des théories de Condorcet et Borda, elles ne traduisent pas l'ordre de préférence des différents votants. C'est pourquoi, on leur préfère l'utilisation d'une fonction de support.

Une fonction de support permet d'associer un poids à chaque votant. On trouve une grande diversité de poids dans la littérature : la confiance d'un classifieur dans sa décision [Van Erp et al., 2002], ou utiliser une pondération spécifique en fonction du classifieur [Wozniak and Jackowski, 2009], par exemple. L'utilisation d'une fonction de support permet de considérer l'intégralité des étiquettes au moment de la prise de décision. L'idée générale consiste à associer à chaque classe un score, avant la fusion des différents résultats. Ainsi, on se rapproche de la méthode de Borda, utilisée par exemple dans [Ho et al., 1994] afin de fournir un poids à chaque classe en fonction de son classement réalisé par le classifieur. Ainsi, il est possible de réaliser un vote par probabilité, qui donne un ordre de préférence par le biais de poids relatifs entre les classes. Là encore on retrouve une grande variété de sources potentielles de scores, en exploitant par exemple les probabilités a posteriori [Biggio et al., 2007].

On retrouve également des méthodes d'agrégations exploitant diverses règles, comme l'identifiant du classifieur (et donc sa spécificité) seul, ou combiné avec la liste des primitives utilisées, ou de manière plus complexe en mélangeant les informations d'étiquettes de classe, l'identifiant du classifieur et des primitives utilisées [Wozniak and Jackowski, 2009].

Enfin, des approches proposent de formuler le choix d'une fonction de décision optimale comme un problème de classification [Inoue and Narihisa, 2002, Hashem, 1997]. À propos de ce principe, [Shlien, 1990] propose d'exploiter l'approche de Demspter-Shafer, qui exploite des connaissances a posteriori sur la performance du classifieur associé, pour combiner des arbres de décision. Une autre approche consiste à sélectionner le sous-ensemble optimal de classifieurs à l'aide d'algorithme de fouille de données [Opitz and Shavlik, 1996, Ruta and Gabrys, 2005].

8.1.3 Conception du système de vote

Partant du même concept que pour le classifieur unique, nous avons conçu le système de vote de sorte qu'il puisse accepter un nombre variable de domaines Source. La figure 8.1 illustre la conception du système de vote. Toutes les approches de vote présentées dans ce chapitre exploitent cette même structure. La différence entre elles se situant au niveau de la fonction de décision qui fusionne le résultat de chaque classifieur selon des règles différentes.

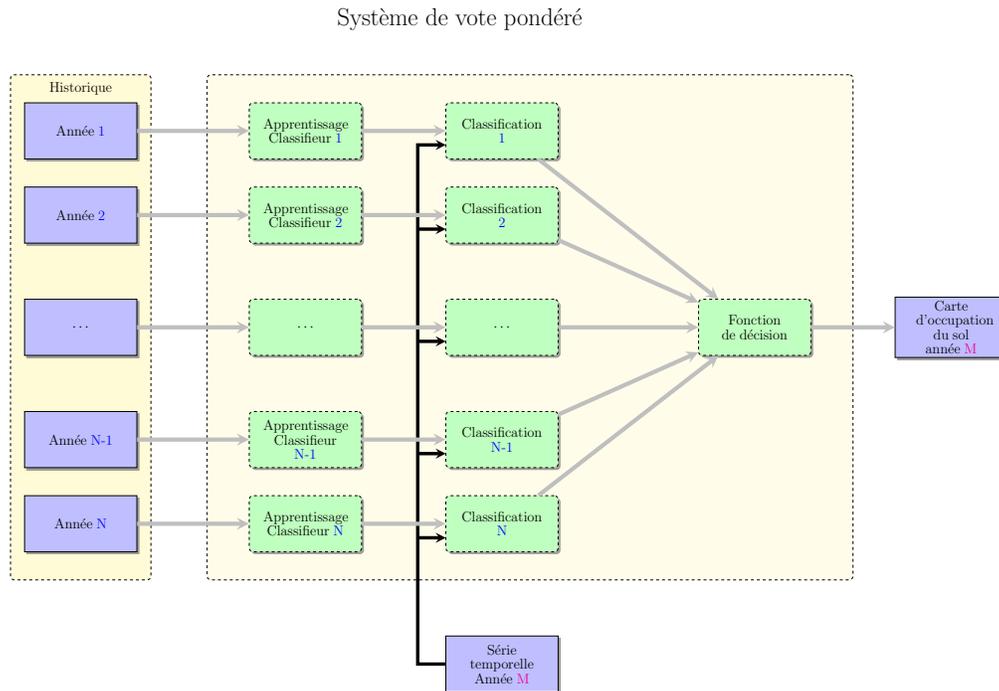


FIGURE 8.1 : Conception du système de vote pondéré.

Dans ce système de vote, les différentes années composant l'historique constituent les données d'entrée. Inversement au classifieur unique, nous entraînons un classifieur pour chaque domaine Source disponible. Les classifieurs ainsi entraînés permettent de produire une version de la carte d'OCS du domaine Cible par la classification de la série temporelle. Notons que le domaine Cible n'appartient pas à l'historique, il n'y a donc pas de classification supervisée dans ce processus (domaines Source et Cible identiques). Enfin, nous fusionnons ces différentes versions de la carte du domaine Cible, afin d'obtenir la carte d'OCS finale.

8.1.4 Choix des entrées

Avant d'aborder les différentes fonction de décisions, notons que nous disposons d'un large choix sur l'origine des votants. En effet, la seule contrainte consiste à respecter la condition d'indépendance entre les votants [Xu et al., 1992]. En résumé, chaque approche de classification, répondant aux critères imposés dans cette thèse, peut être utilisée comme votant.

Dans ces travaux nous étudierons uniquement la classification naïve et la classification après le TO. Pour le TO, nous ne considérons que la résolution EMD exploitant la procédure SAP qui offre les meilleures

performances jusqu'à présent, en tant que données d'entrée des systèmes de vote. Nous ferons référence à cette approche sous le terme générique : TO.

8.1.5 Règles de décision

En exploitant ces différents niveaux de décision, nous pouvons générer un vote pondéré. Avec un poids neutre fixé à 1, le vote basé sur les étiquettes détermine un simple **Vote Majoritaire (VM)**. Deux autres types de pondération peuvent être exploités, la confiance d'une part et les probabilités par classe comme poids dans le vote d'autre part. Ce qui définit respectivement le **Vote par Confiance (VC)** et **Vote par Probabilités (VP)**.

Ces règles de fusion s'avèrent assez simples, mais chacune possède des particularités :

Vote Majoritaire

Le VM est un processus très léger à mettre en place, n'utilisant que les étiquettes. Nous pouvons définir le VM comme un vote pondéré, où le poids associé à chaque classe est calculé par une fonction indicatrice :

$$f_k^c = \begin{cases} 1 & \text{si le classifieur a choisi la classe } c \\ 0 & \text{sinon} \end{cases}$$

Vote par confiance

Le VC pondère les étiquettes par la confiance du **RF** dans sa décision. Par exemple, une classe choisie par 50 arbres pèsera moins que cette même classe sélectionnée par 80 arbres.

Dans ce cas la fonction de pondération est définie par :

$$f_k^c = \begin{cases} \text{conf}_c(x) & \text{si le classifieur choisit la classe } c \\ 0 & \text{sinon} \end{cases}$$

Vote par probabilité

Le VP peut être interprété comme une amélioration du VC. En effet, un votant apporte un poids pour chaque classe et non pas uniquement à la classe majoritaire. Dans ce cas, la fonction de pondération est définie par :

$$f_k^c = \begin{cases} \text{prob}_c(x) & \text{nombre d'arbres ayant choisi la classe } c \end{cases}$$

Notons que ces trois approches sont imbriquées. En effet, le VP représente le cas le plus général, à partir duquel les deux autres peuvent être retrouvés. Ainsi, en ne considérant que le maximum de probabilité pour chaque votant, nous retrouvons le VC. En considérant que le classifieur ne commet pas d'erreur, la confiance vaut alors 1, nous retrouvons le VM.

8.1.6 Le problème des non-décisions

Effets des non-décisions

Avant de réaliser les comparaisons des méthodes de vote proposées, nous allons quantifier l'impact des non-décisions dans le calcul des métriques d'évaluation des cartes produites par vote. La figure 8.2 présente les **PG** obtenues par le VM, le VC et le VP dans les cas de figure où les non-décisions ne sont pas comptabilisées en erreur (en pointillés) ou inversement (courbes pleines).

Le VP et VC présentent l'avantage de pouvoir exprimer à la fois leur accord et leur désaccord sur un choix de label. Cette prise de position peut grandement influencer sur le résultat du vote. Dès lors le nombre de votants impactera fortement les résultats du vote. Les poids utilisés étant des nombres réels entre 0 et 1, les cas de non-décision se trouvent largement éliminés pour ces deux votes, contrairement au VM. Nous avons défini un seuil ϵ à 10^{-3} . Ce seuil représente la différence minimale entre les votes pour deux classes afin de considérer qu'il n'y a pas d'égalité.

Le VC exploite les cartes naïves en entrée et la confiance associée, et comme attendu on ne distingue qu'une légère variation qui témoigne de la présence en faible quantité de pixels non décidés. Nous réalisons le même constat avec le VP. Pour le VM, la présence des non-décisions entraîne une diminution forte des performances obtenues. Nous ne présentons ici que le vote à partir des classifications naïves, l'effet est similaire pour le vote exploitant le TO. L'effet que nous souhaitons mettre en évidence ici étant similaire pour les deux types d'entrée, nous ne présentons que les performances à partir du cas naïf. Nous reviendrons sur la comparaison entre les entrées plus tard dans ce chapitre. L'écart causé par la prise en compte des non-décisions avoisine les 30 % pour le cas 2 (où on exploite deux années en entrée du vote, par extension le cas N indique que nous utilisons N années dans le vote), qui constitue le cas présentant le plus de non-décisions. Cet écart réduit fortement lorsqu'au moins on utilise trois années. En effet les cas de non-décisions sont moins susceptibles de se produire.

Résolution des cas de non-décision

La propriété du VM a ne pas prendre de décision lors d'une égalité dans le vote pose un dilemme. D'une part les décisions prises se révèlent correctes en grande majorité, mais de l'autre la carte ainsi générée demeure inexploitable pour un utilisateur final, en présence d'un trop grand nombre de cas de non-décisions. Il devient alors primordial de résoudre ces cas de non-décisions avant d'écarter les approches de vote de la liste des solutions opérationnelles.

Une première solution proposée consiste à reconstruire une carte à partir des résultats du vote et de la classification naïve. Par exemple, lors d'une égalité dans le vote, on sélectionne l'étiquette de la carte naïve générée avec le classifieur le plus récent. Cette décision s'avère totalement arbitraire car aucun élément n'indique que cette carte soit la meilleure parmi les votants. On peut assimiler cette approche à un transfert d'étiquette depuis la classification naïve dès qu'un cas de non-décision apparaît. Nous avons baptisé cette approche "canevas" pour l'analogie avec le support de broderie, où l'absence d'étiquette s'apparente aux motifs à broder.

Cette décision pourrait alors être motivée par la confiance, dans ce cas il résulte un mélange entre le VM et le VC. L'intérêt de cette approche est alors minime car il suffit de réaliser un VC qui présentera très peu de cas de non-décision. Forcer une dépendance à la confiance ou aux probabilités par classes rend ces approches inutilisables, si le classifieur choisi ne fournit pas ces informations.

Pour palier au problème de non-décision, nous allons utiliser la [MT](#) définie lors du chapitre [7.2.1](#). Les conditions d'utilisation de la MT demeurent les mêmes :

- La continuité temporelle doit être respectée entre l'historique et l'année courante.
- Les cartes utilisées pour générer la clef de transition sont produites grâce à la classification supervisée.

En cas de non-décision, l'historique de ce pixel détermine la clef de transition.

Nous exploitons la MT de deux manières :

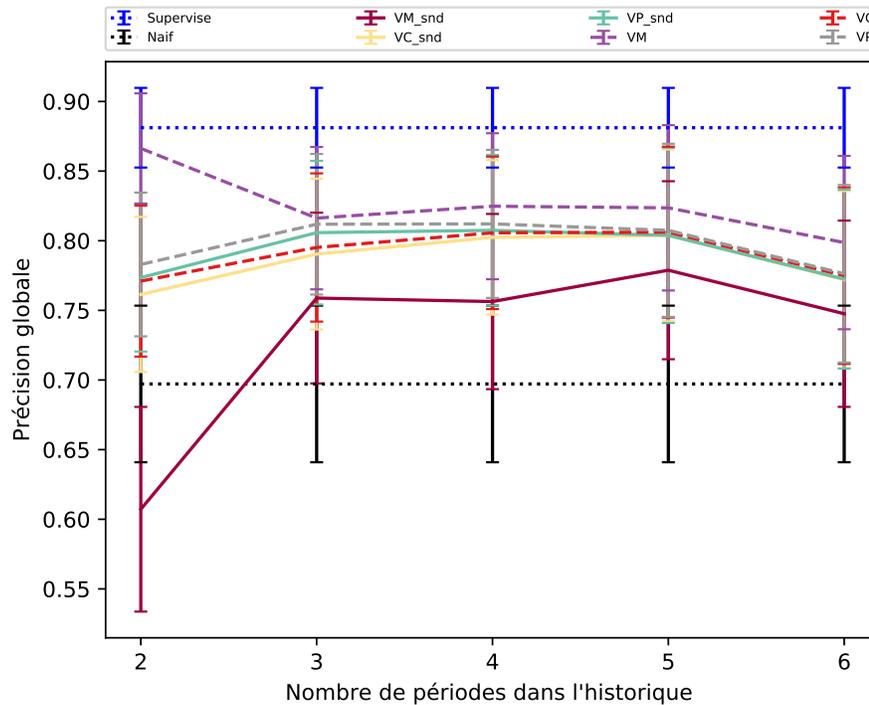


FIGURE 8.2 : PG obtenues lorsque les non-décisions ne sont pas comptabilisées (courbes en pointillées + snd) ou sont considérées comme des erreurs dans la validation (courbes pleines) pour le VM, le VC et le VP. Les courbes horizontales représentent les performances du cas naïf et supervisé pour référence. Les performances présentées exploitent les classifications naïves. Les liaisons entre les différents points permettent d'illustrer les variations.

- la sélection libre par maximum de probabilité
- la sélection guidée par maximum de probabilité

La sélection libre consiste à exploiter de manière standard la MT : la clef de transition indique la ligne concernée, et l'étiquette choisie correspond à la valeur maximale de probabilité.

La sélection guidée consiste à restreindre le choix possible des étiquettes en ne considérant uniquement celles proposées par les votants (et donc responsables de la situation de non-décision).

8.1.7 Performances du vote majoritaire

La figure 8.3 présente sous forme de matrices les performances obtenues par les différentes méthodes de résolution des cas de non-décisions dans le vote majoritaire (en ordonnées) selon le nombre d'années de l'historique utilisé (abscisses). Dans cette figure, la matrice de droite présente la PG tandis que celle de gauche les intervalles de confiance à 95 %. On relève, comme premier constat, que les performances des systèmes de vote s'améliorent avec le nombre d'années utilisées. Pour le VM, nous avons déjà illustré l'impact des non-décisions sur le calcul des performances, ici, elles sont comptabilisées comme des erreurs, d'où le faible

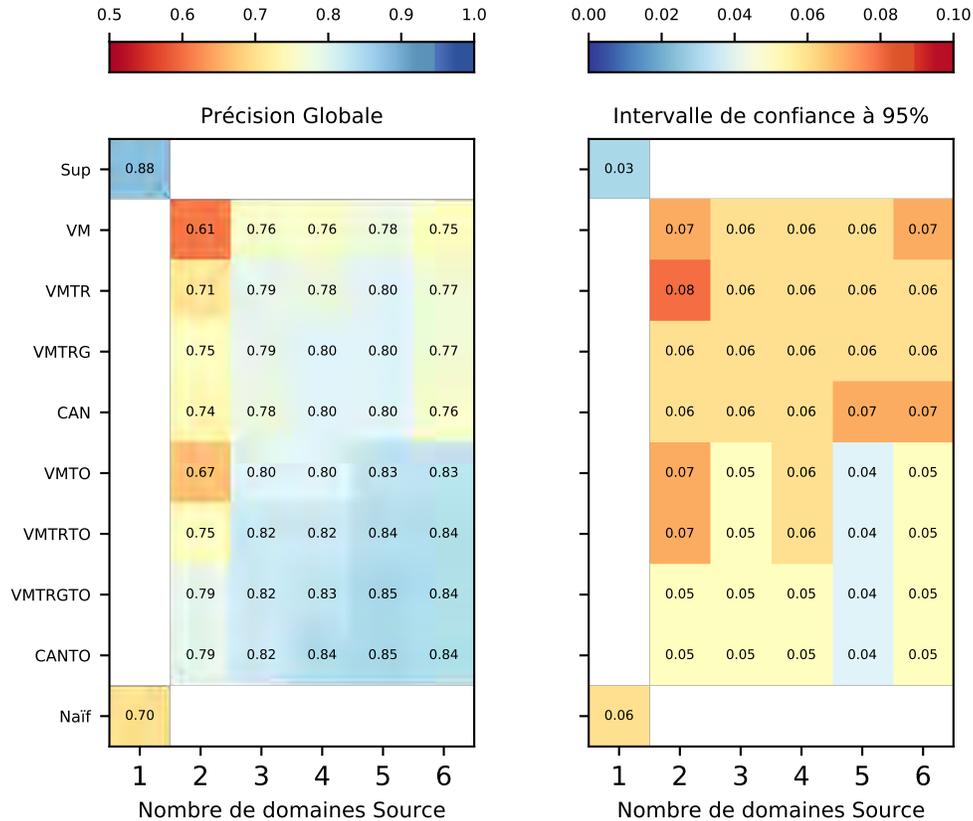


FIGURE 8.3 : PG obtenue lorsque les non-décisions sont corrigées par l’approche des transitions (TR) ou canevas (CAN). On représente les performances sous la forme d’une matrice, où l’axe des abscisses représente le nombre d’années de l’historique utilisées, l’axe des ordonnées les différentes méthodes pour le utilisant les cartes naïves ou après transport (OT).

score obtenu pour deux années d’historique. Pour cette configuration, les performances se révèlent inférieures au cas naïf (indiqué en bas à gauche). On repère le même phénomène pour le VM avec transport (VMOT). Le cas 6 enregistre une faible chute de performances, cela provient de la contrainte chronologique imposée par la MT (une seule combinaison possible contre plus de 4000 pour le cas 2).

Intéressons-nous exclusivement à l’impact des différentes méthodes de prises de décision pour le VM exploitant les classifications naïves. Nous avons proposé trois approches : le canevas (CAN) et l’utilisation de la matrice de transition avec toutes les classes (VMTR) et en guidant la sélection seulement entre les classes concernées par l’égalité (VMTRG). Ces trois approches fournissent des PG similaires, dans les intervalles de confiance, mais améliorent fortement les performances du VM notamment pour le cas 2 avec une amélioration jusqu’à 14 % de la PG pour le VMTRG. Dès le cas 4, ces différences sont atténuées en raison du faible nombre de non-décisions.

Nous pouvons observer des variations entre ces différentes approches en analysant le FScore par classe, présenté dans la figure 8.4. Il s’agit également d’une matrice, dans laquelle l’axe des abscisses indique la classe concernée et l’axe des ordonnées la méthode utilisée. On présente uniquement le cas 2 ici, on dispose

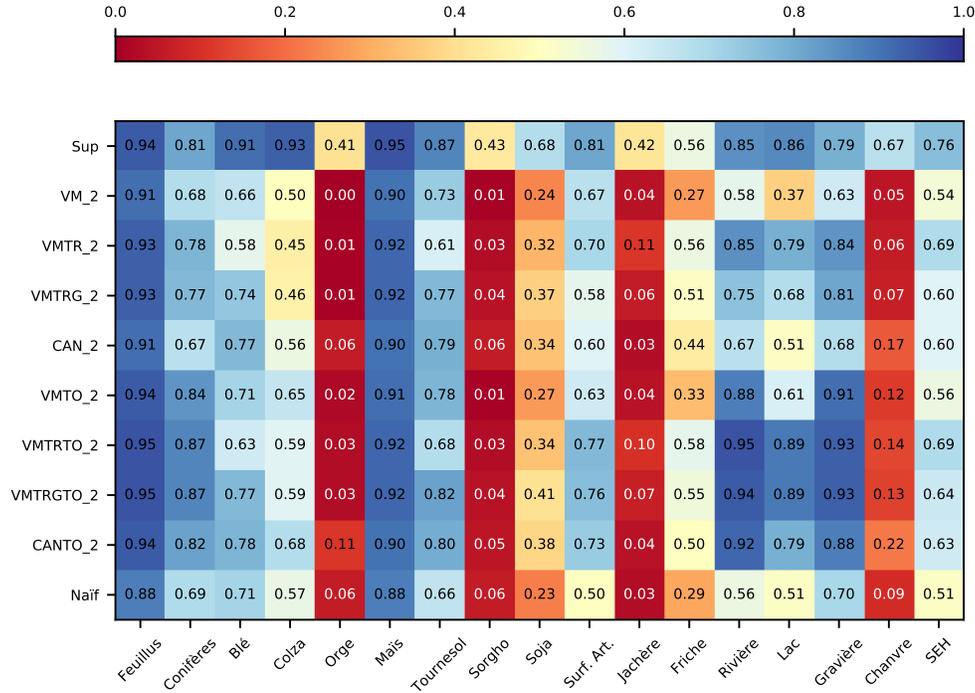


FIGURE 8.4 : Matrice de FScore présentant les méthodes de vote et les différentes approches de prises de décisions pour le cas 2. L'axe des abscisses indique les différentes classes et l'axe des ordonnées les méthodes concernées. L'échelle de couleur varie entre 0 et 1 (FScore maximal).

de l'intégralité des performances en annexe (les figures E.1, E.2, E.3, E.4). On ne relève pas de différences importantes parmi les diverses performances. Nous pouvons cependant relever quelques exemples d'écart significatifs entre l'utilisation de la matrice de transition, comme pour le Blé qui présente un FScore de 0.58 pour le VMTR et de 0.74 avec le VMTRG ou inversement avec la classe Jachères où le VMTR s'avère plus performant. Les approches exploitant les transitions offrent de très bonnes performances pour les classes géographiquement stationnaires, comme les Lacs et Rivières, où l'on relève une amélioration de plus de 20 % de FScore. Pour ces mêmes classes, l'approche CAN reproduit les performances du cas naïf. Nous pouvons relever le FScore de 0.17 pour la classe Chanvre obtenu par l'approche CAN alors que les transitions plafonnent à moins de 1 %. Cette différence peut s'expliquer par la stationnarité des champs de Chanvre et leur faible représentation dans la zone d'étude qui implique une faible probabilité dans la matrice de transition.

Intéressons-nous à présent à la seconde partie des figures 8.3 et 8.4 qui présente les performances lors de l'utilisation des cartes générées par le TO (avec SAP). Nous retrouvons l'amélioration moyenne de 5 % entre le cas naïf et le TO observé précédemment. Nous retrouvons les schémas observés pour le VM, les différences entre les approches demeurant minimales. L'approche CAN bénéficie le plus du TO. En effet, il fournit des cartes avec une meilleure précision, automatiquement répercutée lors de la prise de décision [Lam and Suen, 1997]. Nous pouvons observer ce gain pour les classes de Lacs et Conifères par exemple.

Ces résultats mettent en évidence la problématique des classes mal prédites par la classification naïve. En effet, pour l'Orge et le Sorgho, aucune méthode ne parvient à dépasser celle-ci. De plus, ces classes

sont également mal prédites par la classification supervisée, ce qui impacte obligatoirement la matrice de transition.

8.1.8 Comparaison des systèmes de vote pondéré

Dans cette section nous allons confronter le VM et sa variante utilisant les transitions guidées (VMTRG), avec le VC et le VP. La figure 8.5 utilise la même présentation des performances que précédemment. Les votes exploitant les cartes générées via le TO présentent le suffixe correspondant.

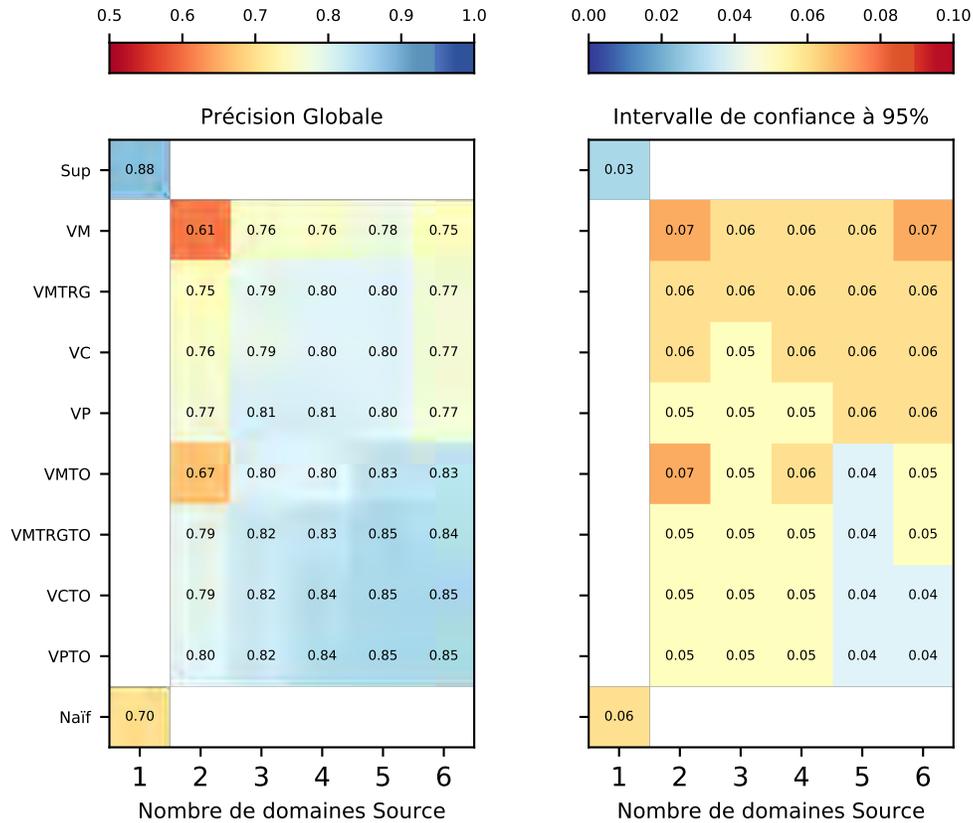


FIGURE 8.5 : PG obtenues par les différentes méthodes de vote pondéré. La matrice de gauche présente les performances, celle de droite les intervalles de confiance à 95 %. L'axe des abscisses présente le nombre d'années de l'historique utilisées, et l'axe des ordonnées les différentes méthodes.

Nous retrouvons une situation de forte similarité entre les performances des méthodes proposées. Aucune ne se démarque nettement des autres, car les intervalles de confiance contiennent la plupart de ces performances. Nous retrouvons également l'apport de quelques points de PG supplémentaires offerts par le TO. Le VM reste moins performant que tous les autres votes présentés, en raison des nombreux cas de non-décisions comme vu précédemment, néanmoins l'apport des transitions guidées réduit ces différences. Les VP et VC bien que très proches ne sont pas équivalents, démontrant la pertinence de l'information apportée par les probabilités par classe.

Nous pouvons constater des différences plus marquées via le FScore présentées dans la figure 8.6. Nous retrouvons le schéma du gain apporté par la procédure SAP pour le FScore du TO sur les classes pérennes notamment. Les votes exploitant le TO parviennent à prédire certaines classes minoritaires comme l'Orge, ce qui représente un point d'intérêt majeur en comparaison aux méthodes qui ne prédisent jamais cette classe (classification naïve, par exemple).

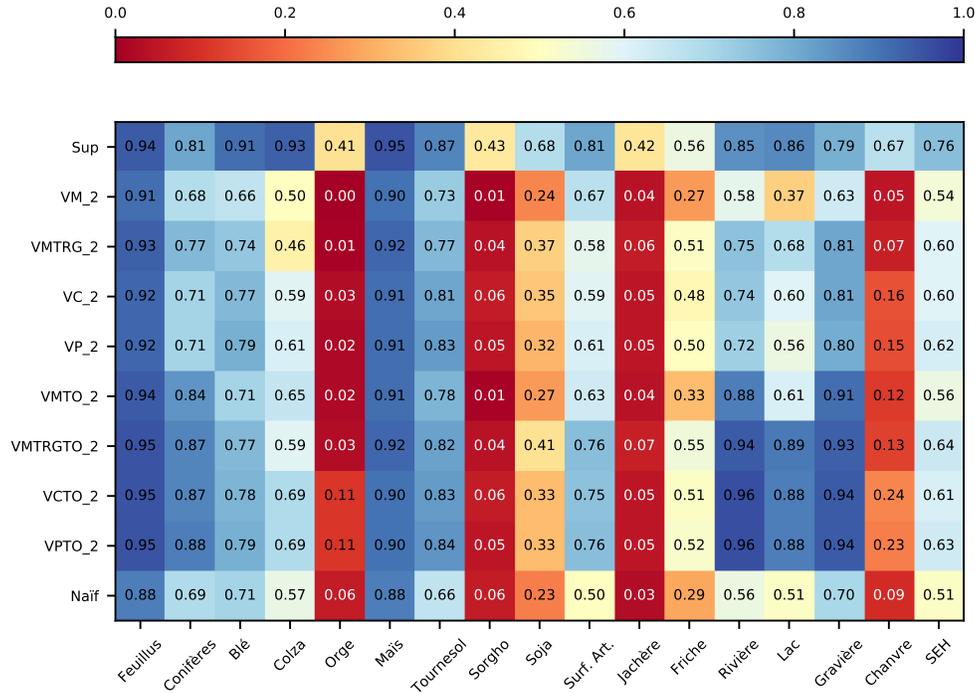


FIGURE 8.6 : FScore obtenu par les différentes méthodes de votes pondérés, pour le cas 2. Les performances des VM, VP, VC ainsi que le VMTRG sont présentées selon l'axe des ordonnées, l'axe des abscisses représente les différentes classes.

Les systèmes de vote offrent un grande liberté de choix dans les composantes utilisées. Il devient alors facile de multiplier les approches de votes. Cependant, les variations de pondérations n'apportent pas de grandes différences sur les performances mesurées. Aussi, nous pouvons nous tourner vers des approches de fusion plus élaborées, comme la fusion par Dempster-Shafer.

8.2 La fusion par Dempster-Shafer

Les approches de vote majoritaire pondéré offrent de bonnes performances. Néanmoins, dans ces approches chaque classifieur possède le même poids. Or, les performances de ceux-ci varient en fonction des combinaisons d'années. Cependant, nous avons mis en évidence que, dans la majorité des cas, les performances obtenues par un classifieur restent stables lorsqu'il est appliqué à différentes séries temporelles. La table 3.1 illustre cet effet, où les performances restent peu dispersées autour de la moyenne. L'objectif de la méthode proposée dans cette section, consiste à exploiter la fusion de classifieurs par Dempster-Shafer, en estimant des performances

à partir des classifications naïves de l'historique.

8.2.1 Théorie de Dempster-Shafer

La théorie Dempster-Shafer ou théorie de l'évidence désigne une fusion de données fondée sur la notion de preuves distinctes, permettant de calculer la probabilité d'un évènement p en fonction de masses de croyance (*mass of belief* en anglais).

Définissons X comme l'ensemble des évènements p possibles (dans notre cas p représente le choix d'une classe d'OCS) :

$$X = \{p_1, \dots, p_n\},$$

alors l'ensemble des parties de X est représenté par :

$$2^X = \{A | A \subseteq X\} = \{\emptyset, \{p_1\}, \dots, \{p_n\}, \{p_1, p_2\}, \{p_1, \dots, p_n\}\}.$$

L'élément A peut être une hypothèse simple comme $\{p_1\}$ qui indique que la classe choisie vaut 1, ou un ensemble d'hypothèse comme l'élément $\{p_1, p_2\}$ qui signifie que la classe est 1 ou 2.

Les masses de croyance possèdent les propriétés suivantes :

$$m(\emptyset) = 0,$$

et

$$\sum_{A \in 2^X} m(A) = 1.$$

La masse $m(A)$ d'un évènement A de l'ensemble des parties représente la proportion de toutes les preuves exclusivement en faveur de A .

À partir des masses de croyance Shafer [Smith and Shafer, 1976] détermine deux mesures : la croyance $bel(A)$ et la plausibilité $pl(A)$ qui vérifient la contrainte :

$$bel(A) \leq pl(A).$$

On définit la croyance $bel(A)$ comme la somme des masses de tous les sous-ensembles d'un ensemble A .

$$bel(A) = \sum_{B | B \subseteq A} m(B).$$

On définit la plausibilité $pl(A)$ comme la somme des masses de tous les ensembles B qui intersectent l'ensemble d'intérêt A .

$$pl(A) = \sum_{B | B \cap A \neq \emptyset} m(B).$$

Ces deux mesures s'opposent :

$$pl(A) = 1 - bel(\bar{A}),$$

de telle sorte que généralement, on peut alors estimer la masse, la croyance ou la plausibilité à partir des deux autres grandeurs. L'intervalle de confiance de l'évènement A est fourni par $bel(A), pl(A)$.

Pour installer le système de décision, ou combinaison de Dempster, on construit un jeu de masses pour les N domaines Source (les votants). Soit S_1, \dots, S_N les N domaines Source, alors chacun dispose d'une fonction de croyance m distincte : m_1, \dots, m_N . On définit alors la valeur de la masse de croyance comme :

$$m(A) = \frac{1}{1 - \kappa} \sum_{B_1 \cap \dots \cap B_N = A} \left(\prod_{j=1}^N m_j(B_j) \right),$$

où κ désigne un terme de normalisation nommé "terme de conflit", qui permet de lisser les fortes divergences entre sources.

$$\kappa = \sum_{B_1 \cap \dots \cap B_N = \emptyset} \left(\prod_{j=1}^N m_j(B_j) \right).$$

L'évènement A disposant de la plus haute masse de croyance associée détermine la classe choisie. Pour la fusion de classifications, les masses associées à chaque votant peuvent être identiques pour toutes les classes et exploiter la PG ou le Kappa. Ces masses peuvent également être spécifiques à chaque classe en exploitant la précision ou le rappel de chaque classe. Ces quatre fonctions de croyance sont utilisées et présentées dans les résultats.

Généralement utilisée avec des classifications supervisées, pour lesquelles on dispose de données de validation, on peut facilement exploiter ces différentes métriques. Lors de la définition des contraintes liées au sujet de thèse, nous avons convenu que nous disposions des données de validation pour réaliser l'évaluation des méthodes. Cependant, nous excluons d'exploiter ces évaluations pour la production. Par conséquent, nous relevons l'impossibilité d'exploiter les métriques réelles associées à chaque votant et donc d'estimer sa capacité de généralisation à des années différentes.

On peut néanmoins contourner cet obstacle en exploitant l'historique. Chaque votant est issu d'une classification de la série temporelle de l'année courante par un classifieur différent. Nous réalisons alors la classification naïve de toutes les séries temporelles de l'historique pour chacun des classifieurs ; nous pouvons alors réaliser la validation de ces cartes obtenir une matrice de confusion pour chaque cas naïf. Cette matrice de confusion estime la capacité de chaque classifieur naïf à généraliser correctement aux autres années. Cette astuce nous offre la possibilité d'estimer une matrice de confusion moyenne permettant de calculer les différentes métriques associées à chaque votant.

Par exemple, posons 2011 l'année que l'on cherche produire. Soient 2007, 2008 et 2009 disponibles dans l'historique. Neuf classifications sont alors produites, trois représentent les votants et six utilisées pour générer une matrice de confusion moyenne par classifieur, comme indiqué dans la table 8.2. Pour chaque classifieur, les deux matrices issues de la validation (provenant du même classifieur et donc de la même ligne du tableau) sont moyennées et on procède à l'association des métriques au votant correspondant (en ligne dans la table 8.2). Ainsi, dans notre exemple, la matrice de confusion moyenne issue de la ligne 2007 sert à calculer les masses associées à la classification de la série temporelle de 2011 par le classifieur issu de 2007. Il en est de même pour les classifications de la série temporelle de 2011 par les classifieurs 2008 et 2009, auquel on associe les matrices de confusion moyenne. Ainsi nous sommes en mesure de fournir les masses requises par la fusion par Dempster-Shafer pour chaque votant.

TABLE 8.2 : Classifications générées pour réaliser la fusion Dempster-Shafer

Classifieur \ ST	2007	2008	2009	2011
2007		Validation	Validation	Votant
2008	Validation		Validation	Votant
2009	Validation	Validation		Votant

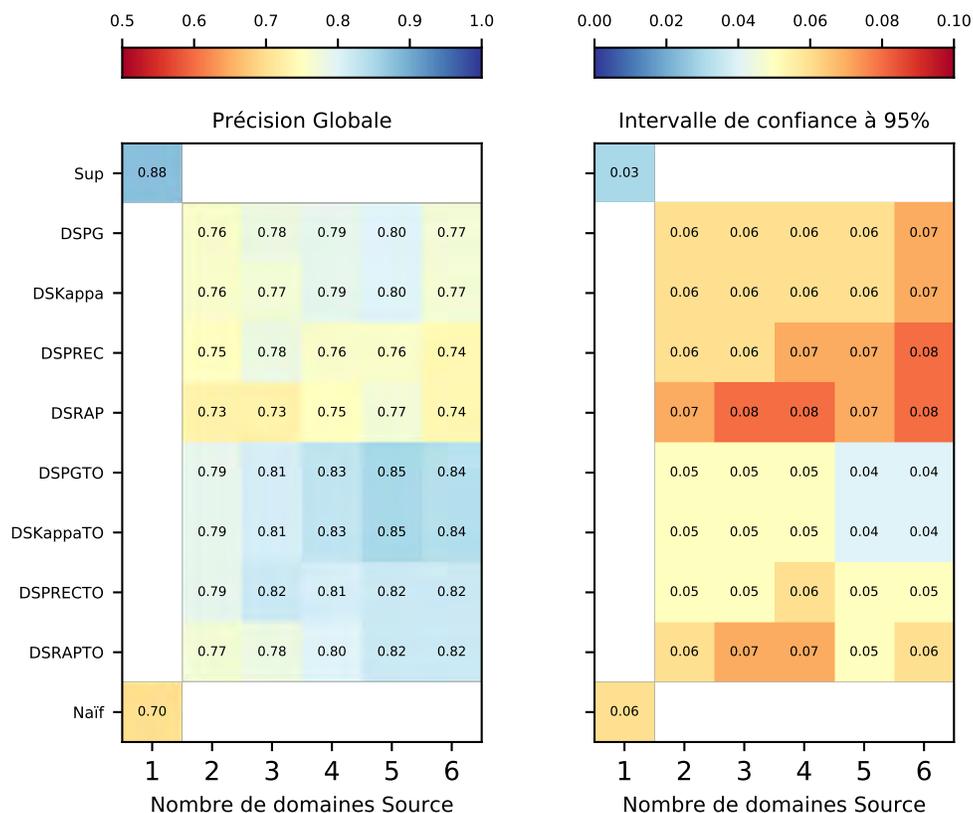


FIGURE 8.7 : PG obtenue pour les différentes approches de fusion par Dempster-Shafer. On représente les performances sous la forme d'une matrice où l'axe des abscisses présente le nombre d'années de l'historique utilisées, et l'axe des ordonnées les différentes méthodes pour les fusions utilisant les cartes après transport (suffixe TO) et les cartes naïves (sans suffixe).

8.2.2 Résultats

Dans cette expérimentation, nous allons comparer les différentes métriques exploitées dans la fusion Dempster-Shafer. Au nombre de quatre, on distingue deux indices globaux (Kappa et PG) et deux indices par classes (précision et rappel). La figure 8.7 utilise le formalisme introduit précédemment pour présenter les PG obtenues, encore une fois pour les deux types d'entrée disponibles (naïf et TO).

Nous pouvons constater que les deux indices globaux produisent des performances identiques. Elles suivent

la même logique d'amélioration lorsque le nombre d'années utilisées augmente. Nous retrouvons l'écart entre le naïf et le TO toujours d'environ de 5 %. L'utilisation de la précision et du rappel fournit des performances symétriques avec une performance maximale du rappel pour le cas 3, et le cas 5 pour la précision. Ces performances restent néanmoins très proches dans les intervalles de confiance, ne mettant en évidence aucune différence significative en faveur de l'un ou l'autre indice. L'analyse du FScore, ne révèle aucune particularité pour le cas 2. Le cas 3 disponibles en annexe dans la figure E.1, met en évidence de légères variations dans le FScore. Ces variations minimales estimées à 0.01 pour certaines classes peuvent s'expliquer par la différence entre la PG et le Kappa. Ces différences s'accroissent lorsque le nombre d'années utilisées augmente. Le calcul de la matrice de généralisation constitue une des sources probables de ces variations. Le nombre d'années disponibles dans l'historique influe sur le nombre de votants, mais également sur le nombre de cas naïfs utilisés pour calculer la matrice. Une variation brutale dans les performances d'un classifieur pour un domaine cible peut alors dégrader la pertinence de la matrice de généralisation, causant ainsi une différence entre les valeurs de Kappa et de PG utilisées lors de la fusion.

Une approche judicieuse consisterait à utiliser une matrice médiane ou réaliser une moyenne dans laquelle les valeurs extrêmes ne sont prises en compte. Nous pourrions également envisager d'introduire les différentes mesures vues précédemment dans la fusion, telles que les transitions, la confiance, les probabilités par classe, etc. Néanmoins, toutes ces mesures peuvent être assimilées à des probabilités, aussi l'approche de fusion bayésienne semble la plus immédiate.

8.3 Approche de fusion bayésienne

Nous avons proposé et évalué, durant nos travaux de thèse, de nombreuses méthodes, avec un taux de réussite variable. La plupart de ces méthodes utilisent des procédés simples, mais sans obtenir de résultats probants dans des cas d'utilisation complexes. À l'inverse, les méthodes plus complexes, soit en termes de processus ou de mise en oeuvre, se sont révélées inadaptées face à un problème de très grande dimension. Dans tous les cas nous avons essayé de palier aux différentes problématiques en exploitant des informations à-priori, comme les rotations entre classes d'OCS, ou sur la nature des classes : annuelles ou pérennes. Cependant ces moyens supplémentaires impactent peu les performances générales, révélant la suprématie de la décision du classifieur sur les autres méthodes de décision. Ainsi, nous exploitons systématiquement une chaîne de traitement, durant laquelle, après la décision du classifieur et l'identification du problème, (non-décision par exemple) vient l'utilisation de la méthode de correction (matrice de transition dans ce cas). Or chaque grandeur utilisée, que ce soit la matrice de transition ou les connaissances a priori sur la proportion des classes dans la zone, s'avèrent exploitables afin d'utiliser une variante d'inférence bayésienne. Ainsi nous pouvons retarder au maximum la phase de décision de chacune des instances pour obtenir une décision plus fine.

8.3.1 Modélisation

Nous définissons alors $P(c|V, T, \pi)$ notre inférence bayésienne (qui caractérise un cas particulier d'un classifieur par maximum de vraisemblance). Nous avons retenu trois facteurs susceptibles d'interagir entre eux pour améliorer les performances obtenues jusqu'à présent :

1. V : représente l'ensemble des votants, issus du cas naïf ou du TO.
2. T : représente une transition entre classes, issue de la matrice de transition.

3. π : représente les proportions a priori des classes dans la zone d'étude.

Nous avons déjà exploité les grandeurs V et T dans les systèmes de vote pour palier au problème des non-décisions du VM. La grandeur π se révèle complexe à déterminer pour le domaine Cible dans la mesure où le seul moyen pertinent de l'estimer consiste à réaliser la classification supervisée. Cependant, nous pouvons partir de l'hypothèse selon laquelle la rotation des cultures conserve les proportions entre classe à l'échelle de la zone d'étude. En considérant plus d'une année pour calculer ces proportions, nous pouvons obtenir une estimation moyenne probablement proche de la réalité. Cette grandeur peut impacter fortement les résultats de l'inférence bayésienne. C'est pourquoi nous évaluerons deux critères de vraisemblance, où la proportion π est optionnelle. Le théorème de Bayes permet ce genre de manipulation.

On définit alors le critère de vraisemblance que l'on cherche à maximiser par :

$$P(c|V, T, \pi) = P(V|c)P(T|c)P(\pi),$$

ou

$$P(c|V, T) = P(V|c)P(T|c),$$

pour la version sans les proportions par classe. On définit alors :

- $P(V|c)$: la probabilité sachant la classe c d'obtenir le votant V .
- $P(T|c)$: la probabilité sachant la classe c d'obtenir la transition T .
- $P(\pi)$: la probabilité a priori (proportion) π

Le classifieur prédit la classe c pour un échantillon, où c maximise la valeur du critère de vraisemblance.

Un des avantages majeurs d'une inférence bayésienne demeure sa capacité à exploiter un nombre variable d'informations. Ainsi, nous projetions d'exploiter des informations annexes comme la similarité entre les années d'un point de vue climatique (précipitation, température). Les matrices de confusion estimées pour la fusion par Dempster-Shafer constituent également une piste intéressante à intégrer dans l'inférence bayésienne. Cependant, nous n'avons pu explorer plus en avant ces approches faute de temps.

8.3.2 Résultats

Contrairement aux autres approches proposées dans cette partie, le classifieur bayésien peut fournir une classification sur la base d'un seul votant. En effet, nous rappelons que la MT provient d'une source lambda, l'harmonisation obligatoire des étiquettes au problème posé caractérisant la seule contrainte. Ainsi une MT ne disposant que d'une seule classe en clef de transition peut alors être utilisée conjointement à l'unique cas naïf issu de l'historique. Dans notre cas, le processus commence par la classification de 2009, exploitant la classification naïve de 2008 avec la MT déterminée à partir des cartes supervisées 2007 et 2008. Ensuite, la continuité chronologique nous impose une nouvelle contrainte. La figure 8.8 présente les PG obtenues avec le classifieur bayésien lorsqu'on utilise les cartes naïves issues du TO. Nous indiquons également à titre de comparaison les performances du classifieur unique, pour lequel nous considérons uniquement les cas respectant la chronologie, expliquant ainsi des performances légèrement différentes comparativement à la section 7.3.2. À l'inverse de toutes les autres méthodes, le classifieur bayésien brille lorsque on utilise peu d'années. Il est également la seule méthode utilisable avec une seule année d'historique.

Ce comportement s'explique par l'accumulation des erreurs dans les différentes composantes, entraînant ainsi une forte probabilité associée à celles-ci. Par contre pour un faible nombre d'années, elle permet d'approcher, dans les intervalles de confiance, les performances du classifieur unique.

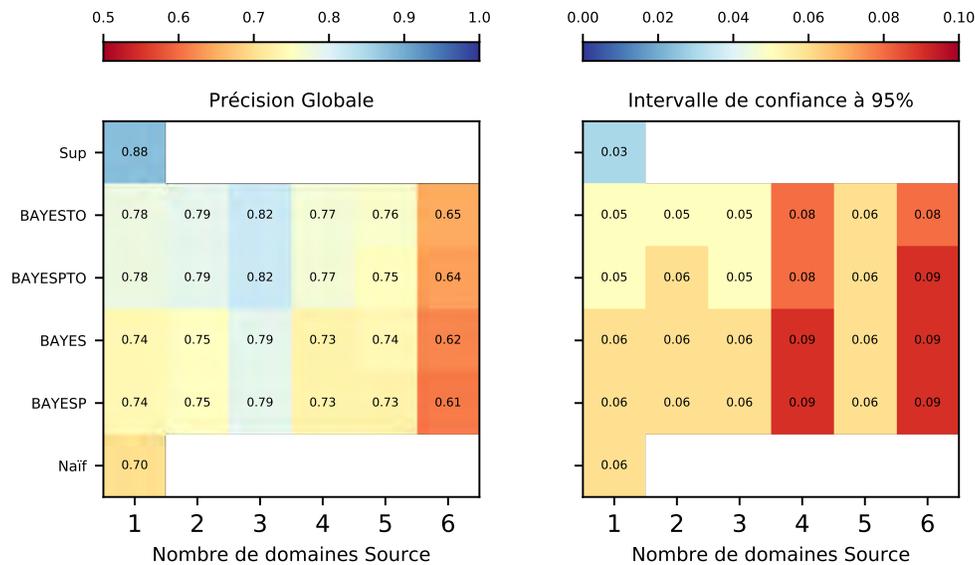


FIGURE 8.8 : PG bayésien avec l'exploitation optionnelle du TO et des proportions π .

Nous retrouvons la sensibilité du classifieur face aux erreurs dans les différentes sources dans la figure 8.9. Tous les FScore chutent lorsque le nombre d'années utilisées augmente, à l'inverse du classifieur unique. Ces dégradations sont nettement plus marquées pour les classes difficiles à classer, soumises à un taux de mauvaises classifications plus important : Orge, Soja, Chanvre . . .

Le classifieur bayésien ne conserve pas les gains apportés par l'utilisation du TO, contrairement à toutes les autres méthodes proposées jusqu'à présent.

Le cas 1 du classifieur bayésien représente une opportunité intéressante, que ce soit à partir du cas naïf ou du TO. En effet, les valeurs de FScore égalent ou surpassent celles du cas naïf, notamment pour les classes difficiles à prédire comme le Chanvre, l'Orge grâce au TO, ou encore le Soja.

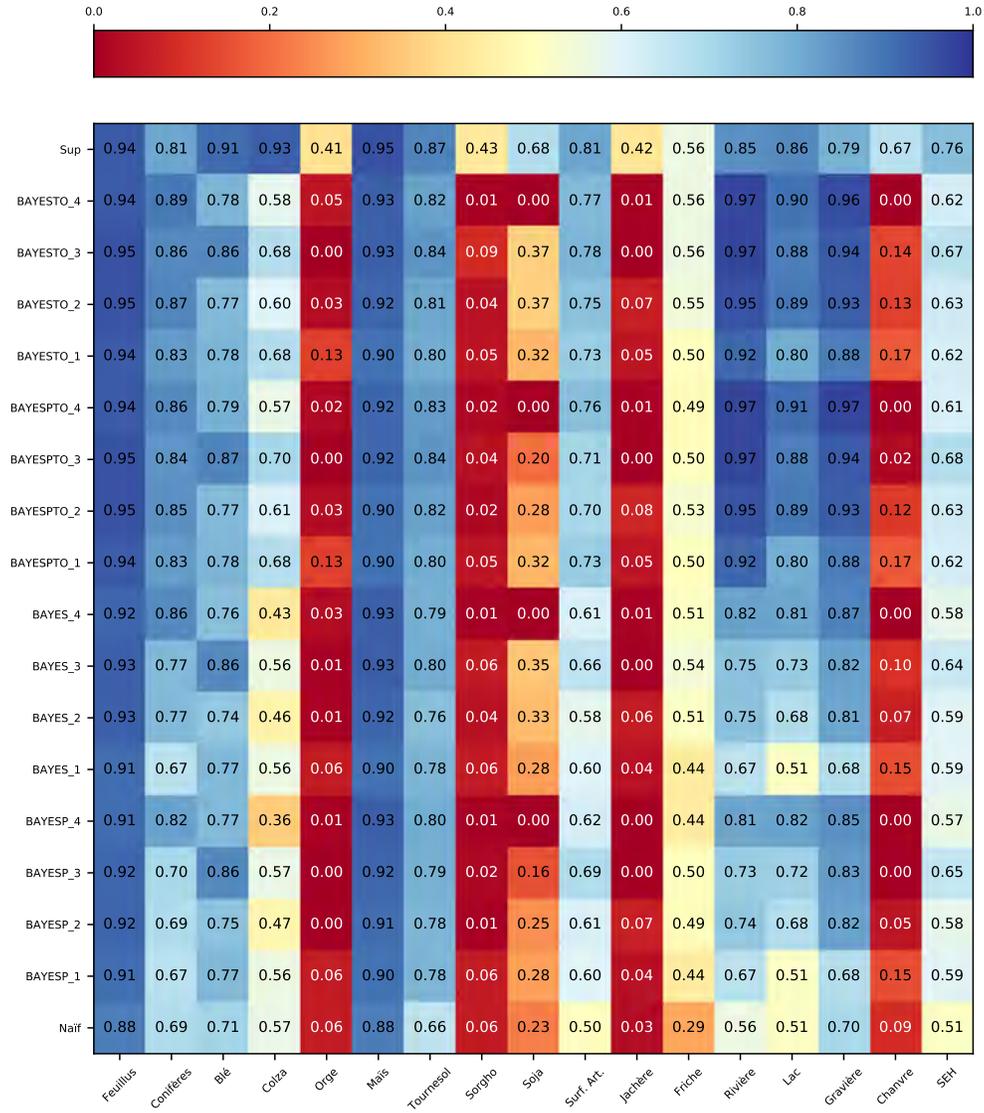


FIGURE 8.9 : FScore obtenu pour les classifieurs bayésien pour les cas 1,2,3 et 4 qui sont les plus significatifs pour ces méthodes. L'utilisation des proportions π est indiquée par la lettre P dans le nom de la méthode.

Comparaison des approches multi-annuelles

La seconde partie du manuscrit est consacrée à différentes approches permettant de combiner les informations issues de l'historique. La première méthode consiste à modéliser les transitions de classes d'OCS à partir des cartes supervisées des années précédentes. Bien que les résultats de cette approche demeurent peu convaincants, ils mettent en évidence l'intérêt d'exploiter plusieurs domaines Source en parallèle. Ainsi, la seconde approche proposée consiste à entraîner un classifieur unique à partir d'échantillons issues de plusieurs années. Cette approche offre de bons résultats, mais ne permet pas de reproduire les cas de gain observés, pour des combinaisons spécifiques de domaine, lors d'une simple classification naïve. Pour exploiter les meilleurs prédictions des classifications naïves, nous proposons l'utilisation de systèmes de vote pondéré. Ces systèmes offrent des performances correctes, mais il reste difficile de déterminer un critère d'arrêt sur les informations utilisées lors de la pondération. Ainsi, nous avons exploité le cadre de la fusion Dempster-Shafer, afin d'évaluer l'apport des métriques dans la fusion de classification. La principale difficulté demeure l'estimation des matrices de confusion utilisées pour déterminer les masses de croyances. De plus, la fusion Dempster-Shafer ne permet pas d'exploiter des informations annexes, qui ont déjà prouvé leurs intérêts comme les probabilités par classes ou la confiance. En dernière approche, nous proposons l'exploitation d'un modèle intégré de fusion, dans le cadre d'une inférence bayésienne. Ce type d'approche permet alors d'ajouter, sous forme de probabilité un grand nombre d'information et de retarder la décision de chaque élément.

Au cours de ce chapitre, nous proposons de comparer ces différentes méthodes. Une première comparaison porte sur les performances globales et par classes obtenues par les différentes approches. Dans un second temps, nous analysons les coûts de production associés à chacune des méthodes avec différents critères.

9.1 Comparaison statistique

La figure 9.1 recense les PG obtenues pour toutes les méthodes présentées dans cette partie. Les deux matrices de cette figure présentent les PG obtenues lors de l'exploitation de la classification naïve à gauche, et à droite celles exploitant les classifications obtenues après TO. Les performances de la classification supervisée et

naïve ne suivent pas cette règle, car elles représentent les limites de performances de nos cas de référence. Une absence de couleur correspond à la non-exploitation de la méthode pour la taille d'historique indiquée en abscisse.

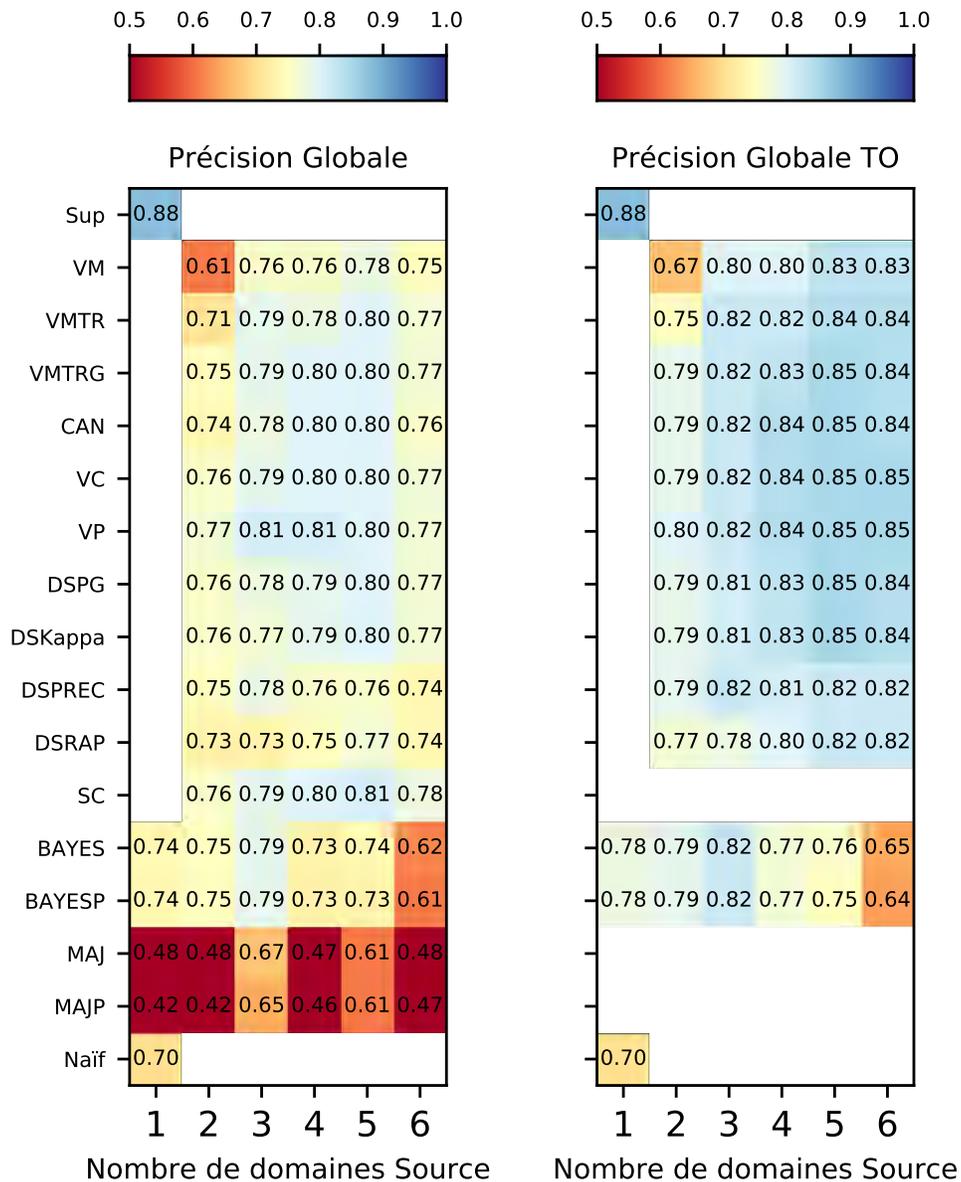


FIGURE 9.1 : PG obtenue pour toutes les méthodes abordées dans ce chapitre.

La figure 9.1 confirme les résultats présentés au cours de cette partie, notamment sur la forte similarité des performances, entre méthodes d'une même catégorie, par exemple les votes pondérés (*Vote par Confiance* (VC), *Vote par Probabilités* (VP)), mais plus globalement pour toutes les méthodes pour une taille d'historique donnée. En effet, dans la matrice de gauche pour la colonne 2, hormis l'approche de mise à jour

(MAJ) et le VM comptabilisant les non-décisions, toutes les PG présentent des valeurs entre 0.71 et 0.76. On retrouve exactement la même stabilité dans la matrice de droite pour des valeurs entre 0.75 et 7.80 pour la même colonne.

Nous constatons que la fusion bayésienne offre des performances très correctes dès qu’une seule année est disponible, sous réserve de disposer d’une MT. L’estimation de la MT représente un point clef d’amélioration de ces performances. En effet, nous avons mis en évidence des défauts dans l’estimation actuelle, par exemple lors de l’utilisation d’un simple comptage à partir des classifications sans pré-traitement. Une étape de régularisation des cartes pourrait grandement améliorer l’estimation des probabilités de transitions. De même l’exploitation des données de référence pour déterminer les transitions, ou corriger les erreurs de classification apporterait une meilleure précision de la MT. Cette amélioration pourrait s’avérer également bénéfique pour l’exploitation de la MT dans les cas de non-décision rencontrés dans les systèmes de vote et notamment le VM.

Globalement la matrice de droite, présentant l’apport du TO, offre une distribution de valeurs très homogène. La principale explication de ces bonnes performances, ainsi que la stabilité des différentes méthodes dans ce cas, se trouve dans la procédure SAP qui permet d’exploiter les échantillons de l’année correspondante pour les classes pérennes. Le rôle du TO demeure important, car il permet de pallier les distorsions causées par la distribution irrégulière des images au cours des années. La mise à disposition des images Sentinel-2 tend à corriger ce problème, grâce à une acquisition plus régulière et dense.

Alors que les différentes méthodes offrent des performances globales similaires, le FScore révèle des écarts parfois importants entre elles. Cependant, ces écarts concernent essentiellement des classes minoritaires, ce qui explique le faible impact observé sur la PG. La figure 9.2 présente la matrice de FScore correspondant à la seconde colonne des matrices (cas 2) de PG présentées dans la figure 9.1, les figures en annexes E.1, E.2, E.3, E.4 présentent les FScore obtenus pour les autres tailles d’historiques (cas 3 à cas 6).

Dans ces figures, nous retrouvons l’amélioration systématique des classes pérennes, par l’utilisation conjointe du TO et SAP, qui permet également une bonne reconnaissance des classes majoritaires de cultures (Blé, Tournesol). Cependant, le gain sur les classes minoritaires reste aléatoire notamment pour le Chanvre et l’Orge.

Les méthodes de fusion exploitant des mesures de probabilités dans leur règles de fusion (Dempster-Shafer, VC après TO) permettent une meilleure reconnaissance de ces classes, comme le montre l’augmentation de 10 % du FScore pour la classe Orge avec ces approches. Cependant l’augmentation de la taille de l’historique fait chuter ces gains. Cet effet met en évidence la difficulté à prendre en compte les distorsions entre les années.

L’amélioration du FScore pour ces classes peu représentées constitue une piste intéressante, car celles-ci demeurent un problème récurrent, y compris pour la classification supervisée. Une piste envisageable pour améliorer cet aspect des méthodes consiste à exploiter les approches d’augmentation de données. L’objectif d’une telle méthode consiste à générer des échantillons supplémentaires à partir de ceux disponibles dans la donnée de référence, permettant ainsi d’améliorer l’équilibre entre les classes.

Enfin, nous pouvons conclure sur la forte corrélation entre la qualité des classifications utilisées en entrée des votes et les performances obtenues. L’amélioration de la classification naïve par le TO se répercute sur les classifications obtenues par les systèmes de vote [Xu et al., 1992]. Dans cette optique l’utilisation du TO présente donc un intérêt fort, cependant les difficultés d’exploitation de celui-ci représentent un frein à sa mise en application à grande échelle opérationnelle.

Les figures 9.3 et 9.4 représentent une zone d'intérêt extraire de l'emprise Formosat-2. Chaque miniature représente une méthode de production différente. Les méthodes exploitant une seule année (supervisé, naïf et TO) exploitent l'année 2009 en unique domaine Source. Les autres méthodes exploitent un historique composé de 2009 et 2010. Le domaine Cible est toujours l'année 2011. Les couleurs utilisées sont indiquées dans la nomenclature indiquée dans la figure 1.4. Une couleur blanche indique les cas de non-décisions lors des votes.

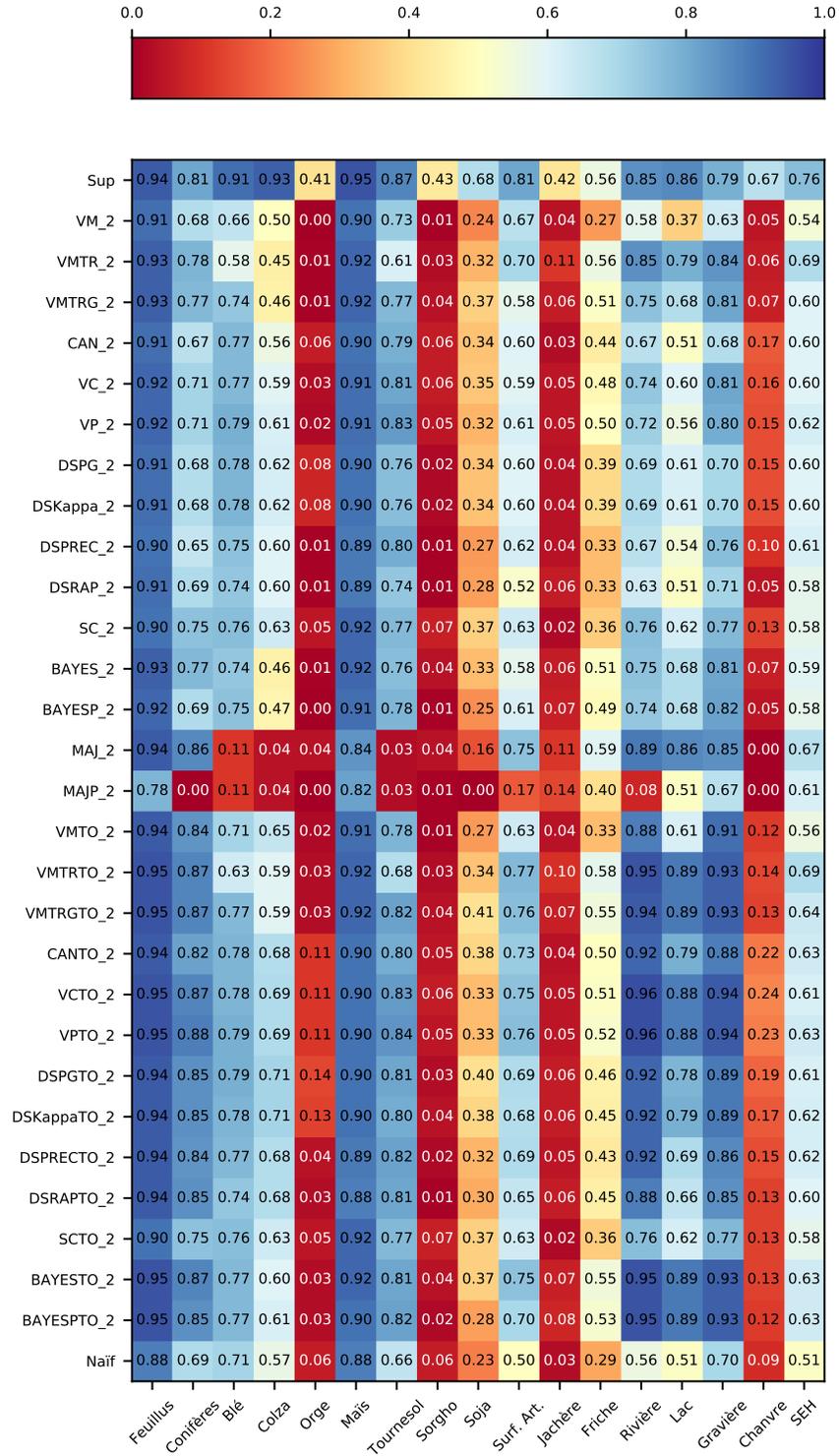


FIGURE 9.2 : FScore obtenus pour toutes les méthodes proposées.

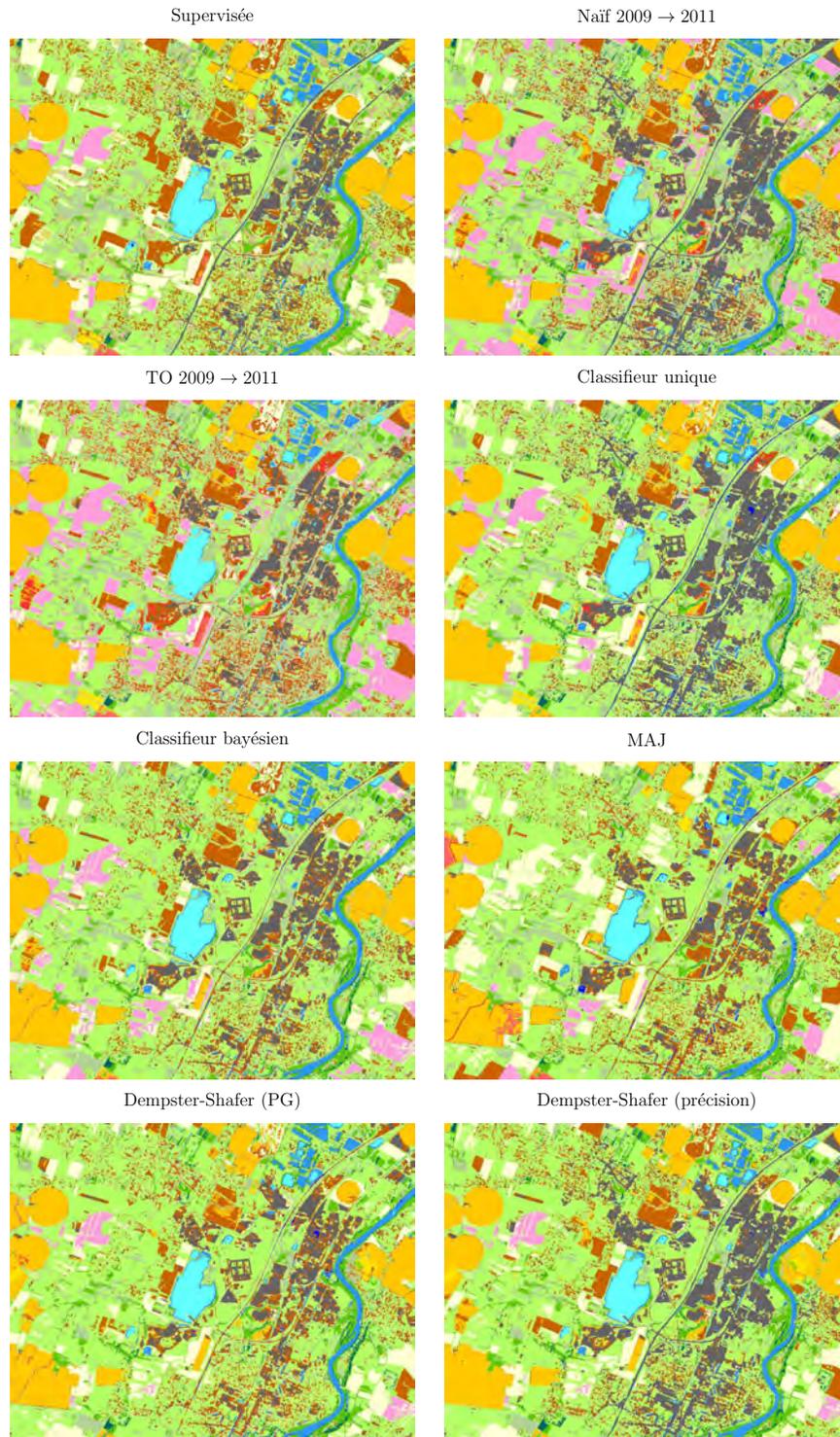


FIGURE 9.3 : Illustration des cartes produites selon différentes méthodes à partir des données Formosat-2.

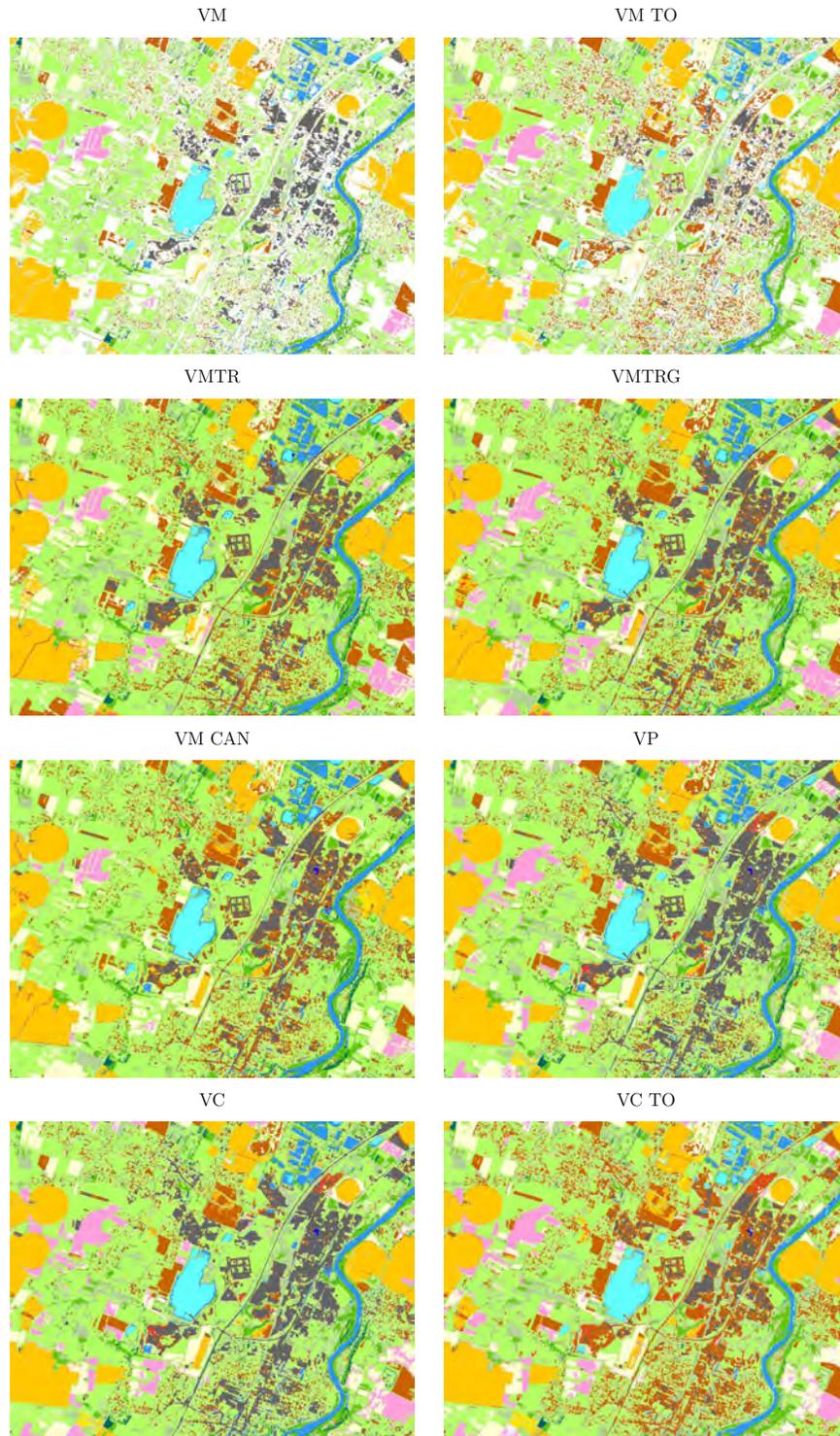


FIGURE 9.4 : Illustration des cartes produites selon différentes méthodes à partir des données Formosat-2.

9.2 Analyse des coûts de production

Les performances présentées aux cours de cette thèse ne permettent d'élire une méthode unique permettant de satisfaire au mieux les demandes des utilisateurs, sans utiliser les données de référence correspondantes à la carte souhaitée. De plus, la vocation des expérimentations proposées cherche à mettre en évidence des comportements et des tendances de performances en fonction de la composition de l'historique.

Dans cette section nous proposons des critères d'évaluation afin de comparer les coûts de production de chaque méthode, mais également des pré-requis à l'utilisation des méthodes proposées. Cette section pourra être exploitée comme un guide d'utilisation des différentes méthodes proposées dans ce manuscrit.

9.2.1 La composition de l'historique

Un premier critère que nous mettons en avant concerne les données disponibles dans l'historique. Au cours des travaux, nous avançons la possibilité d'exploiter des données hétérogènes pour contourner la dépendance aux données de référence. Dans cette catégorie nous considérons huit points d'intérêt :

1. Nombre de domaines Source disponibles (NbDom).
2. Séries temporelles (ST) : exprime la nécessité de disposer de l'information spatiale des images de la série temporelle d'un domaine de l'historique.
3. Données de référence (DRef) : uniquement pour les domaines de l'historique. On accède ici aux étiquettes avec leurs positions géographiques.
4. Échantillons d'apprentissage(EApp) : sous la forme d'une base de données, permettant d'entraîner un nouveau classifieur. On accède ici à des points pour lesquels on dispose des étiquettes et des primitives, donc nous n'avons pas besoin des images.
5. Classifieurs : ceux utilisés pour produire des cartes supervisées (déjà entraînés).
6. Cartes d'OCS : produites par classifications supervisées.
7. Une ou plusieurs matrices de transition : pouvant provenir d'une source quelconque
8. Autres informations susceptibles d'améliorer la méthode concernée (Autres) : intervention de l'utilisateur, connaissances a priori, etc.

Nous utilisons la table 9.1 pour synthétiser les différentes méthodes et leurs liens avec ces critères. Nous utilisons une note fixée entre 0 et 5 pour représenter la dépendance de la méthode (en ligne) vis-à-vis d'un critère (en colonne). Une valeur nulle implique qu'il n'y a aucune dépendance entre ces éléments, à l'inverse la note maximale implique une forte contrainte sur la composition de l'historique. Par conséquent, un score faible est préférable à un score élevé. Pour compléter cette table nous considérons uniquement la nécessité de disposer dans l'historique des données en question, cela ne concerne pas les données de la période courante (domaine Cible). Ainsi, la classification supervisée, qui n'exploite aucune donnée de l'historique, possède une valeur totale de dépendance nulle.

La classification naïve représente la méthode la moins dépendante de l'historique. En effet, le nombre de domaines de l'historique n'a pas d'impact puisque nous n'avons besoin que d'un domaine Source. Par contre,

TABLE 9.1 : Évaluation des méthodes selon le critère de composition de l'historique.

Méthode	Nb dom	ST	D Ref	Ech	Classifieur	Carte OCS	MT	Autres	Score
Supervisé	0	0	0	0	0	0	0	0	0
Naïf	1	0	0	0	5	0	0	0	6
TO	1	2	5	5	0	3	0	3	19
ACC	1	5	0	0	0	0	0	3	9
MAJ	5	0	3	0	0	4	5	3	20
VM, VC, VP	5	0	0	0	5	0	0	0	10
Correction non décision	5	0	3	0	0	4	5	3	20
Dempster-Shafer	3	5	0	0	5	0	0	0	13
Classifieur Unique	3	0	0	5	0	0	0	0	8
Classifieur Bayésien	3	0	2	0	5	4	5	3	22

la disponibilité du classifieur constitue une forte contrainte. En effet, si le classifieur n'est pas disponible, cela implique de réaliser de nouveau toute la phase d'apprentissage pour le domaine Source.

Le TO et la projection ACC possèdent la même contrainte de dépendance vis-à-vis du nombre de domaines Source disponibles. Cependant, l'ACC requiert obligatoirement la disponibilité de l'intégralité de la série temporelle du domaine Source afin d'estimer les matrices de projection. Nous attribuons la note médiane pour la catégorie "Autres" car des techniques non-utilisées ici pourraient améliorer l'estimation de l'ACC, par exemple une mise en correspondance des pixels de mêmes classes, sans tenir compte de leur position dans l'image. À l'inverse, le TO dépend moins de l'accessibilité à la série temporelle, car nous pouvons imaginer une technique de sélection d'échantillons exploitant les séries temporelles, comme une détection de changements par exemple. L'accès aux échantillons d'apprentissage du domaine Source demeure obligatoire, car nous les utilisons pour entraîner le classifieur après TO. Il en est de même pour la disponibilité des données de référence utilisées pour sélectionner les échantillons des classes pérennes dans le domaine Cible. Nous laissons l'opportunité à des facteurs extérieurs qui permettraient d'améliorer les performances, par exemple étiqueter une partie des échantillons du domaine Cible via une technique d'apprentissage actif.

Les classifications naïves et après TO constituent les données d'entrée des fusions présentées dans la deuxième partie des travaux de thèse. La conception des systèmes de fusion utilisée amène à additionner les dépendances de chaque système avec le score des entrées utilisées. Par exemple, utiliser l'ACC impose une contrainte maximale vis-à-vis de la disponibilité de toutes les séries temporelles des domaines exploités.

Pour les systèmes de fusion nous évaluons la dépendance vis-à-vis du nombre de domaines Source disponibles en fonction des performances observées. Ainsi, une forte contrainte associe le VM et le nombre de votants, car pour le cas 2 la quantité de non-décisions représente un problème majeur pour l'exploitation. Globalement, le nombre de votants impacte toutes les performances des votes majoritaires pondérés (VC et VP). Cependant, ces trois votes ne requièrent que l'accès aux classifieurs pour générer les votants (dans le cas de la classification naïve) ce qui le place en première position des systèmes de fusion.

En effet, la fusion par Dempster-Shafer requiert la disponibilité des séries temporelles des domaines Source afin d'estimer les performances moyennes obtenues par le classifieur sur le reste de l'historique, et par conséquent de réaliser la classification de ceux-ci. Cette contrainte augmente donc son score de dépendance.

L'exploitation de la MT contraint l'accès aux cartes d'OCS de l'historique afin de générer la clef de transition. Cependant le score n'est pas maximal, car le nombre de cartes exploitées dépend de la taille de la clef de transition, et non pas de celle de l'historique. Cette dépendance augmente fortement le score de dépendances

des méthodes exploitant la MT : classifieur bayésien, méthode de correction des non-décisions, mise à jour de la carte (MAJ). De plus nous avons ajouté un score “Autres” pour ces méthodes, car l’amélioration de la MT que ce soit pour son estimation ou son exploitation permet d’améliorer les performances. Certaines méthodes, notamment la fusion bayésienne, peuvent être exploitées sans toutes les données. Ici, nous présentons le coût dans le cas où toutes les données d’entrée possibles sont utilisées. Ainsi, nous considérons que la fusion bayésienne exploite les cartes supervisées, la matrice de transition et les probabilités a priori. Selon ces critères, le classifieur bayésien représente la méthode la plus dépendante et la classification naïve la plus libre.

9.2.2 La qualité des données

Un second critère concerne la qualité globale du jeu de données. Nous identifions cinq points d’intérêt permettant de définir la qualité des données de l’historique :

1. La régularité d’acquisition des images (RegIm) : une absence d’images pendant une longue période ne représente pas un obstacle
2. Couverture nuageuse (Nuages) : ce critère désigne si la présence d’artefacts dus à la correction des nuages représente un problème pour une comparaison pixel à pixel (ACC, MAD)
3. Ré-échantillonnage temporel (RéEch) : la nécessité d’aligner les séries temporelles entre elles
4. Stabilité des données de référence (StabRef) : la méthode permet de gérer une absence momentanée de classes ou un léger changement de nomenclature
5. Impact de la zone couverte : l’augmentation de l’étendue demande une adaptation de la méthode, pour la préparation des données ou la production de la carte.

Nous distinguons les points 1 et 2 car même s’il s’agit d’une absence de données qui requiert une interpolation pour être corrigée, celle-ci n’a pas le même impact en fonction des méthodes (notamment la classification naïve et la projection ACC). La table 9.2 présente les scores de dépendances aux différents critères, avec la même logique que précédemment. Un score de 0 signifie que le critère n’impacte pas directement le résultat obtenu. Sous ces critères, la qualité de la classification supervisée ne dépend que de deux facteurs : la régularité d’acquisition des images et la couverture nuageuse. En effet, l’absence d’images implique que l’interpolation est identique pour tous les échantillons, alors que la présence de nuages ne concerne que quelques pixels. Ce score peut varier en fonction du classifieur utilisé, un RF sera moins sensible à ce type d’artefact qu’un SVM, car il est possible que la primitive interpolée ne soit jamais sélectionnée lors de la création du RF. Pour ces raisons nous avons fixé le score des trois techniques de classifications (naïve, TO, et classifieur unique) à une valeur de 3. La classification ne dépend pas de la qualité de l’historique, puisqu’elle ne l’exploite pas, le coût associé est donc nul.

La sensibilité à la régularité d’acquisition des images a été mis en avant par la procédure SAP (cf chapitre 3.3). Cependant elle ne présente pas une contrainte forte dans le cas d’une classification naïve standard, impactant peu ses performances. Néanmoins, la classification naïve dépend énormément du ré-échantillonnage temporel, sans lequel le résultat de la classification devient hasardeux. La stabilité des données de référence impacte également les performances du cas naïf. En effet, une classe présente dans la donnée de validation, mais pas dans le domaine Source, rend impossible sa prédiction par la classification. Inversement, une classe

présente dans le domaine Source peut être prédite, et devient par conséquent une erreur au moment de la validation. Définir laquelle de ces erreurs constitue la plus importante relève de l'utilisation souhaitée de la carte, et ne représente donc pas une contrainte algorithmique importante.

TABLE 9.2 : Évaluation des méthodes selon le critère de qualité des données de l'historique.

Méthode	RegIm	Nuages	RéEch	StabRef	Étendue	Score
Supervisé	0	0	0	0	0	0
Naïf	4	3	5	1	0	13
TO	2	3	5	4	5	19
ACC	0	5	0	0	2	7
MAJ	0	0	0	5	0	5
VM, VC, VP	0	0	0	0	0	0
Correction non décision	0	0	0	0	0	0
Dempster-Shafer	0	0	0	5	0	5
Classifieur Unique	0	3	5	1	0	9
Classifieur Bayésien	0	0	0	2	0	2

Le TO permet d'alléger la contrainte sur la régularité d'acquisition des images, dans la mesure où il permet d'adapter le signal du domaine Source à celui du domaine Cible. Néanmoins une forte absence des données peut entraîner une chute de performances et des comportements inhabituels lors de l'estimation des matrices de distances. Le ré-échantillonnage temporel représente une forte contrainte pour permettre l'utilisation de la procédure [SAP](#). Le volume de données considéré constitue un obstacle à l'exploitation à grande échelle du TO. En effet, l'augmentation de la zone couverte requiert d'identifier plus d'échantillons pertinents pour le transport imposant ainsi une mise en place spécifique afin de gérer cette situation.

Ces critères d'évaluation ne concernent pas équitablement toutes les méthodes. En effet, alors que les approches de fusion dépendront essentiellement de la qualité des données d'entrée, les techniques de production de celles-ci dépendent directement de la qualité des données de l'historique. Ainsi on associe des scores très faibles aux fusions. Cependant, comme précédemment, nous sommes dans une logique d'accumulation des dépendances, ainsi un VM à partir des classifications naïves obtiendra un score (13+0) inférieur à un VM à partir du TO (19+0). La fusion par Dempster-Shafer reste une exception. En effet, il dépend de la stabilité des données de référence, car nous les exploitons, afin d'estimer les performances de chaque classifieur exploité dans le vote. Ces performances, sous forme de matrice de confusion définissent alors les fonctions de masse utilisées par la fusion.

La complexité des méthodes ne dépend pas du volume de données, pour la plupart d'entre elles. Bien sûr celui-ci impacte directement le temps de traitement des différentes étapes, mais ne demande pas d'effort particulier pour exploiter la méthode, exceptions faites du TO pour les raisons déjà énoncées, mais également pour la projection [ACC](#). En effet, nous devons envisager une approche spécifique en vue d'assurer une projection cohérente, dans le cadre d'une exploitation multi-tuiles par exemple. Il devient obligatoire d'exploiter des algorithmes spécifiques aux traitements de grands volumes de données afin d'assurer la cohérence entre les combinaisons linéaires déterminée par l'ACC.

9.2.3 Temps de calcul

Le dernier critère considéré consiste à évaluer les temps de calcul requis pour exploiter les méthodes en fonction de la taille de l'historique. Nous ne pouvons pas, en situation réelle, stocker l'intégralité de l'historique en

permanence. Dans le cadre des seules données Sentinel-2, sur l'ensemble de la France métropolitaine par exemple, nous comptabilisons le volume de données en centaine de teraoctets, uniquement pour les images de réflectances. Ainsi, déployer une des méthodes proposées requiert de produire à nouveau une partie des données de l'historique, notamment les séries temporelles avec leurs indices spectraux et les classifications. Néanmoins, on peut aisément envisager que les MT, matrices de confusion et données de référence soient conservés. Dans cette optique, la table 9.3 évalue le coût d'exploitation de la méthode indiquée en ligne. Nous considérons un historique de taille N , et chaque opération coûte une unité.

Une unité d'apprentissage a correspond au coût de calcul requis pour entraîner un classifieur quelconque. Une unité de classification c correspond au coût de ce classifieur pour traiter une série temporelle annuelle quelconque. Une unité de post-traitement p correspond à un parcours pixel par pixel de l'emprise totale de la série temporelle. Dans cette étape de post-traitement, on retrouve les opérations de validation, par exemple pour déterminer les fonctions de masse de la fusion Dempster-Shafer, l'estimation de la MT mais également un processus de vote comme le VM. Un ensemble de votants de même nature (par exemple les étiquettes) est considéré comme une image unique à plusieurs canaux, chaque canal représentant un votant.

Pour le jeu de données Formosat-2, l'analyse des ressources utilisées, temps de calcul ramenés sur un unique processeur, a mis en avant les relations suivantes :

- $c = 3 \times a$
- $p = a$

TABLE 9.3 : Évaluation des coûts de production des différentes méthodes

Méthode	Apprentissage	Classification	Post-Traitement	Temps (en fonction de a)
Supervisé	a	c	0	$4a$
Naïf	a	c	0	$4a$
TO	$2a$	c	0	$5a$
ACC	a	c	0	$4a$
MAJ	Na	Nc	p	$(4N + 1)a$
VM, VP	Na	Nc	p	$(4N + 1)a$
VC	Na	$2Nc$	$2p$	$(7N + 2)a$
Correction non-décisions	Na	Nc	$2p$	$(4N + 2)a$
Dempster-Shafer	Na	$(N^2 - N)c$	$(N(N - 1) + 1)p$	$(N + 3(N^2 - N) + (N(N - 1)) + 1)a$
Classifieur Unique	Na	c	0	$(4N + 1)a$
Classifieur Bayésien	Na	$2Nc$	$2p$	$(7N + 2)a$

Cette table met en avant que la fusion par Dempster-Shaffer constitue l'approche la plus coûteuse. En effet, elle requiert N apprentissages (un par domaine Source), $N^2 - N$ classifications pour générer les cartes naïves nécessaires, ainsi que celles pour l'estimation des masses. Enfin elle requiert $N(N - 1)$ validations pour obtenir les matrices de confusion moyennes, puis une étape de fusion par vote pondéré qui ne parcourt qu'une pile de votants.

Le classifieur bayésien représente également un algorithme très coûteux. L'origine de ce coût provient de l'exploitation de la MT, qui contraint d'avoir accès aux cartes supervisées pour générer la clef de transition, et aux classifications naïves qui constituent les votants. Viennent ensuite les méthodes de vote qui exploitent les classifications naïves, qui exigent autant de classifications et d'apprentissages que de domaines Source.

Le classifieur unique présente l'avantage de ne comporter qu'une seule étape de classification, par rapport aux systèmes de vote.

La figure 9.5 représente le coût sur l'axe des abscisses et de la PG en ordonnées. Les différentes valeurs de coût sont calculées à partir des équations de score présentées dans la table 9.3.

Les arcs de cercle sont centrés sur le point optimal (0, 1) et intersectent respectivement le point (triangle) de la classification supervisée (bleu), celui de la classification après TO (vert) et naïve (rouge). Ainsi l'espace entre l'arc de cercle et ce point idéal représente une amélioration du coût et de la performance de la méthode vis-à-vis de la référence (arc de cercle). Cette représentation permet de visualiser que malgré les étapes requises pour estimer un transport cohérent celui-ci constitue une amélioration de la classification naïve. Les différentes méthodes présentées dans la table 9.3 sont représentées par les lettres colorées. La couleur de la lettre indique le nombre d'années exploitées dans l'historique. Pour des raisons de clarté, certaines approches ne sont pas représentées dans ce graphe, car le couple coût et performances est très proches d'une autre méthode, par exemple le VP et le VM avec MT.

Cette représentation met en évidence le coût proportionnel de la fusion par Dempster-Shafer, imposé par le grand nombre de classifications et de validations croisées pour estimer les matrices de confusion moyennes.

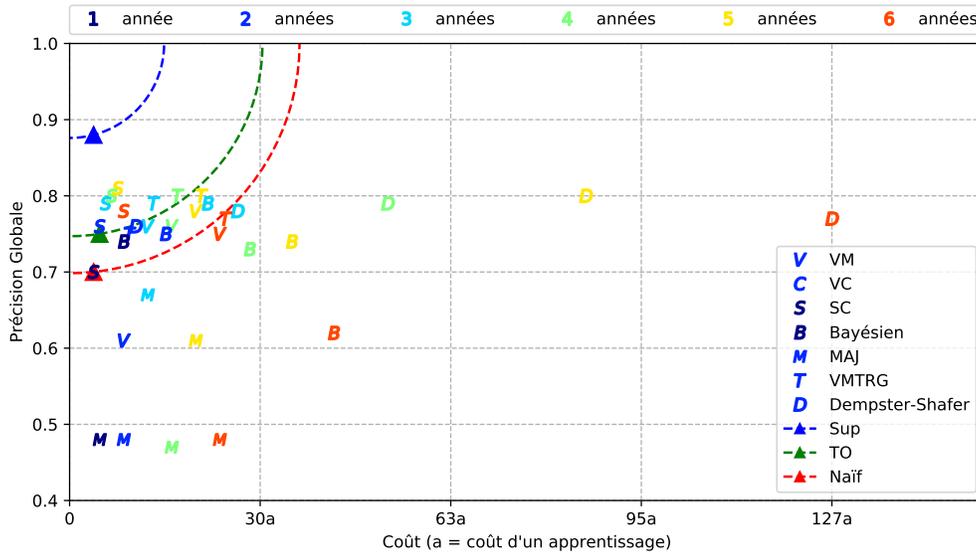


FIGURE 9.5 : Évaluation des performances pour différentes méthodes en fonction du temps d'un apprentissage (en fonction de a) et différents nombre d'années utilisées.

Au cours de ce chapitre nous avons proposé un ensemble de critères : la dépendance à l'historique, la sensibilité à la qualité des données, ainsi que la précision de classification en fonction des coûts de calculs. Ces critères ont pour but de guider l'utilisation des différentes méthodes en fonction de la situation particulière que nous souhaitons traiter.

Cinquième partie

Mise en oeuvre des méthodes pour une production opérationnelle sur de grandes étendues

Production sur des grandes étendues avec Sentinel-2

Les expérimentations menées au cours de cette thèse exploitent des séries temporelles annuelles Formosat-2. Ce jeu de données nous offre des images avec des résolutions spatiales et temporelles adéquates pour la production de cartes d'OCS. Cependant, la zone d'étude reste relativement restreinte avec une emprise de $24km \times 24km$. Au cours de ce chapitre, nous présenterons les résultats des tests de déploiement de la plupart des méthodes proposées, sur un jeu de données Sentinel-2 couvrant le sud de la France métropolitaine continentale. L'objectif de ce chapitre est de démontrer le passage à l'échelle des méthodes, tout en identifiant les contraintes. Le passage à l'échelle a été réalisé en fin de thèse, limitant ainsi le nombre d'expérimentations possibles.

10.1 *iota*²

En télédétection, l'utilisation de données sur de très grandes étendues (par rapport à l'emprise d'une image) demeure un défi. En effet, il faut gérer de nombreux aspects, tels que le recouvrement entre les tuiles, causés par le découpage en segments des acquisitions et la projection en zones UTM, ou encore la différence de dates d'acquisition entre les différentes images couvrant l'intégralité de l'étendue choisie. La complexité de ces problèmes croît avec celle-ci, car le nombre de tuiles augmente proportionnellement à la taille de la zone d'étude, comme le montre la figure 10.1. Cette figure, illustre le recouvrement entre les tuiles et le nombre de tuiles que nous allons exploiter dans ce chapitre (en vert). Le carré bleu au représente l'étendue couverte par les données Formosat-2, exploitées jusqu'à présent.

Dans cette optique, le CESBIO a développé la chaîne de traitement [Infrastructure pour l'Occupation des sols par Traitement Automatique Incorporant les Orfeo Toolbox Applications \(*iota*²\)](#) ¹. La chaîne de traitement, développée en python, se compose de différents modules et classes, qui la rendent adaptable aux besoins spécifiques d'un utilisateur, pour ajouter la gestion d'un nouveau capteur par exemple. Nous utilisons

¹<https://iota2.readthedocs.io/en/latest/index.html>

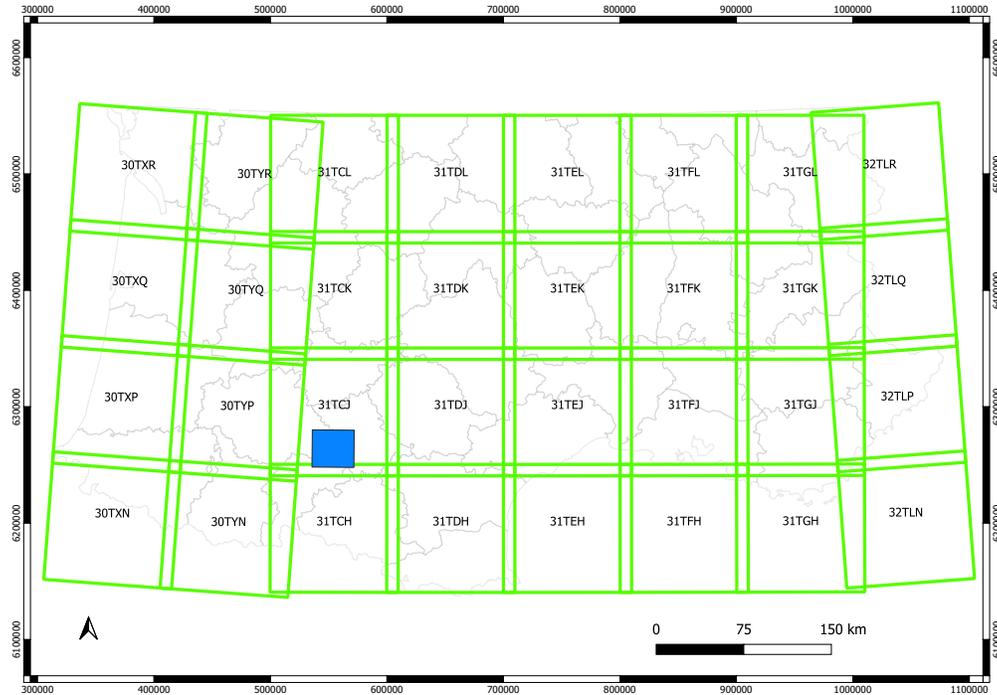


FIGURE 10.1 : Illustration des tuiles Sentinel-2 exploitées dans ce chapitre. Les effets de la projection sur les zones UTM est clairement visible. Le carré bleu représente la zone couverte par Formosat-2.

ici des données Sentinel-2 dont le traitement est pris en charge par *iota*².

Centrée sur un noyau de classification standard, *iota*² facilite la gestion de données multi-tuiles d'images satellites [Inglada et al., 2017].

La première étape consiste à recenser toutes les images disponibles en fonction de la liste des tuiles que nous souhaitons traiter. Ces images constituent les séries temporelles annuelles. Ces images doivent être assorties d'un masque de validité (nuage, saturation, bord). Nous exploitons ceux-ci lors de l'étape de ré-échantillonnage temporel, obligatoire dans le cadre d'une utilisation multi-tuile. En effet, celui-ci assure l'homogénéisation des dates d'acquisition sur l'ensemble de la zone d'étude. Ainsi entre deux tuiles, les dates d'acquisitions virtuelles sont identiques, ce qui assure la cohérence des primitives.

La chaîne *iota*² réalise la plupart des traitements concernant les images tuile par tuile. Néanmoins elle offre la possibilité de gérer des régions éco-climatiques. L'intérêt de ces régions, pour la production de cartes d'OCS, réside dans l'hypothèse qu'au sein d'une même entité, le climat, et surtout son impact sur la végétation, demeurent homogènes. Ainsi, la tâche du classifieur s'en trouve facilitée, en évitant les fortes variations des profils temporels pour ces classes. De même, les modes de fonctionnement et de culture humains demeurent liés au climat et leur position géographique. Ainsi dans une région éco-climatique de montagne nous attendons à voir des toits d'ardoises et non de tuiles comme dans les zones de plaines. Ces différences sont visibles depuis un satellite, mais restent difficiles à gérer pour un classifieur, en raison de différence de signature spectrale de ces matériaux pour une même classe d'OCS. L'exploitation des régions éco-climatiques permet donc de simplifier ces problématiques.

À partir des régions éco-climatiques et des séries temporelles ré-échantillonnées, nous entraînons les différents classifieurs, à raison d'un par région. Une fois les phases d'apprentissage terminées, nous réalisons tuile par tuile la classification de l'intégralité de la zone d'étude. Puis nous fusionnons les différentes cartes pour constituer la carte sur l'emprise définie par la liste des tuiles fournies en entrée de `iota`².

Enfin, nous calculons les métriques (cf section 2.3.1) pour obtenir une évaluation statistique de la carte ainsi produite.

Dans son fonctionnement normal, `iota`² exploite la classification supervisée. Ainsi, en dehors de la gestion des différentes tuiles nous retrouvons la chaîne de traitement présentée dans la figure 2.6.

10.2 Données disponibles

Le jeu d'images Sentinel-2 auquel nous avons accès nous permet de couvrir trois années continues sur l'ensemble du sud de la France métropolitaine. Cependant, ces trois années ne présentent pas la même quantité d'informations. En effet, Sentinel-2 se compose de deux satellites jumeaux, lancés à 2 ans d'intervalles en 2015 puis en 2017. De fait, la série temporelle de 2016 ne dispose que des images issues de Sentinel-2A, acquises à un pas de temps de 10 jours. La série temporelle de 2017 contient des images issues de Sentinel-2B à partir du mois d'avril, commençant ainsi le cycle d'acquisition régulier de 5 jours. Ainsi, la série temporelle de 2018 propose les images des deux satellites sur l'année entière. La table 10.1 indique la répartition moyenne du nombre d'images disponibles pour une tuile par saison. La table en annexe F.1 indique le nombre d'images par tuile et par année, illustrant le déséquilibre dû aux recouvrements entre orbites adjacentes qui doublent les observations.

TABLE 10.1 : Répartition du nombre d'images moyen disponible par tuile et par année

Année	Hiver	Printemps	Été	Automne
2016	7	11	16	15
2017	13	14	18	28
2018	21	20	34	22

Nous disposons également des données de référence pour ces trois années. À l'inverse de celles exploitées avec le jeu Formosat-2, les données de référence utilisées sont extraites de diverses bases de données. Ainsi, pour les classes de cultures, les données proviennent exclusivement du `RPG` de la période correspondante. Les classes pérennes non agricoles proviennent de `CLC 2012`, et des `BD TOPO` et `BD Forêt`.

Pour ce passage à l'échelle, nous disposons d'une nomenclature à 23 classes, représentant des cultures, des espaces naturels et des surfaces artificielles. La table 10.2 présente la nomenclature, ainsi que le nombre d'échantillons disponible par classe sur l'intégralité de la zone d'étude.

Nous exploitons un découpage en 7 régions éco-climatiques, illustrées dans la figure 10.2. Ces régions ne sont pas équilibrées en termes de superficie, mais représentent les différents climats observés dans le sud de la France.

TABLE 10.2 : Nomenclature et nombre d'échantillons disponibles par année.

Classe	2016	2017	2018
Bâti dense	40 759	40 603	40 687
Bâti diffus	396 843	398 593	272 239
Zones industrielles et commerciales	242 357	239 533	227 135
Surfaces routes	26 672	26 662	26 330
Colza	99 217	101 497	108 942
Céréales à pailles	238 329	226 586	232 188
Protéagineux	52 284	39 365	60 587
Soja	246 950	225 806	220 750
Tournesol	706 949	670 224	633 863
Maïs	458 664	440 062	425 297
Riz	37 435	39 229	40 857
Tubercules/Racines	24 195	28 275	23 125
Prairies	401 357	387 472	376 506
Vergers	43 813	44 229	43 720
Vignes	238 902	239 420	237 713
Forêts de feuillus	700 529	707 177	711 576
Forêts de conifères	2 319 135	2 219 223	2 356 542
Pelouses	1 568 094	1 671 980	1 628 090
Landes ligneuses	825 147	852 586	850 047
Surfaces minérales	973 260	681 602	777 577
Plages et dunes	55 195	68 894	86 200
Glaciers et neiges	60 423	70 635	60 931
Eau	1 693 368	2 292 993	2 238 811

10.3 Expérimentations

10.3.1 Préparation des données

Les données de référence ont été préparées au préalable par le [CESBIO](#), nous pouvons donc les utiliser directement.

Pour la préparation des images, nous réalisons le ré-échantillonnage de toutes celles disponibles sur une grille temporelle régulière de 10 jours entre le 1^{er} janvier N et le 31 décembre N . Nous disposons donc de trois années, constituant ainsi un historique composé de deux domaines Source et un domaine Cible.

Nous ne réalisons l'étape de préparation des données qu'une seule fois par domaine. En effet, cette étape n'a aucune incidence sur les différentes méthodes.

Le découpage de la zone étudiée en 7 régions éco-climatiques implique que nous entraînions autant de classifieurs. Avant de réaliser l'apprentissage, nous effectuons la sélection des échantillons par région éco-climatique. Cette étape est également réalisée une seule fois pour chaque année, car elle permet d'accéder aux échantillons requis pour l'application de nos méthodes.

Le volume de données à traiter constitue le principal défi pour ces expérimentations. En effet, pour le nombre d'images présenté dans la table 10.1, nous avons besoin de 5, 6 et 8 teraoctets d'espace de stockage pour l'étape de préparation des données de chacune des années respectivement. Les autres étapes de la chaîne requièrent en moyenne 300 gigaoctets. Bien que ces aspects soient gérés par [iota](#)², nous devons rentabiliser au maximum les différentes exécutions de la chaîne de traitement afin de réduire les temps globaux de calcul. En effet, l'exécution de la chaîne, dans le cadre de la classification supervisée, par exemple sur la structure

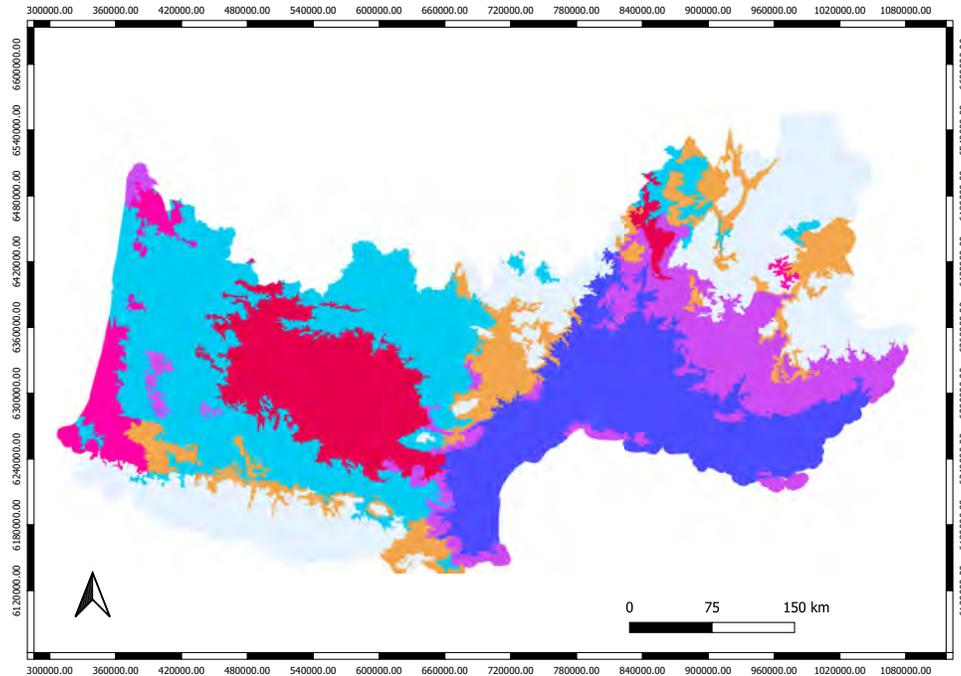


FIGURE 10.2 : Illustration des régions éco-climatiques sur le grand sud

HPC du [Centre National d'Études Spatiales \(CNES\)](#) requiert plusieurs jours.

10.3.2 Utilisation de la chaîne

Comme indiqué précédemment, la chaîne *iota*² permet une utilisation modulaire de ses fonctionnalités. Pour le passage à l'échelle de la plupart des méthodes proposées ici, nous avons mis en place deux protocoles d'exploitation. Le premier déroule toutes les étapes de la chaîne, permettant ainsi la production des cartes supervisées. Le second, consiste à remplacer l'étape concernée par une de nos méthodes, puis de reprendre le processus normal d'exécution de la chaîne.

Parmi les différentes méthodes proposées, toutes n'utilisent pas les mêmes données et ne se situent pas au même endroit dans la chaîne de traitement. Nous avons donc mis en place un plan d'exécution en plusieurs étapes.

Une première étape consiste à réaliser la classification supervisée de chaque domaine. En effet, cette étape nous permet de constituer la plupart des éléments de l'historique mais également de définir les performances obtenues dans ce cas, pour ensuite les comparer avec nos méthodes.

Ainsi nous réalisons trois exécutions de *iota*². Cette opération nous permet donc pour chaque année de réaliser :

1. La préparation des données
2. Le découpage en régions éco-climatiques
3. La séparation des jeux d'apprentissage et de validation

4. L'extraction des échantillons d'apprentissage
5. L'apprentissage des classifieurs
6. La production de la carte supervisée
7. L'estimation statistique des performances

À l'issue de ces sept étapes, nous allons pouvoir dérouler la plupart des méthodes proposées.

Classification naïve

Une fois de plus, la classification naïve constitue la méthode la plus simple à réaliser. En effet, lors de la production des cartes supervisées, nous avons préparé les données de l'année courante, mais également entraîné les classifieurs des années de l'historique. Ainsi, il suffit de classer les données de l'année courante à partir de ces classifieurs.

La projection ACC

Dans ce cas d'utilisation, exploiter la projection par ACC semble très complexe à mettre en oeuvre. En effet, pour assurer la cohérence entre les différentes projections, il faudrait pouvoir traiter d'un seul tenant toutes les tuiles concernées. De fait, la gestion de la mémoire s'avère fort complexe, donc nous ne retenons pas cette approche.

Le Transport Optimal (TO)

Pour exploiter le TO nous disposons plusieurs options de protocoles. En effet, l'extraction des échantillons d'apprentissage nous permet d'accéder facilement aux données du domaine Source. Cependant, la sélection des échantillons dans le domaine Cible s'avère plus délicate. En effet, il faudrait évaluer l'intérêt des régions éco-climatiques pour le TO. Un autre facteur à prendre en considération concerne le nombre d'échantillons utilisés pour estimer un TO pertinent à l'échelle de la région éco-climatique. L'approche du TO par classe serait également envisageable, dans la mesure où nous sommes capables de générer la classification naïve.

Une seconde approche viable serait d'estimer plusieurs transports par région éco-climatique avant de les regrouper pour réaliser l'apprentissage. Une telle approche permettrait d'appréhender les variations au sein d'une même région éco-climatique.

L'application du TO au jeu de données Sentinel-2 n'a malheureusement pas pu être réalisé par manque de temps, car il nécessite, entre autre, une approche de sélection d'échantillons dans le domaine Cible bien plus élaborée que celles proposées précédemment.

La mise à jour de la carte

Bien que les résultats sur le jeu de données Formosat-2 se soient avérés peu concluants, nous avons généré la MT à partir des cartes d'OCS des années de l'historique (2016 et 2017). Puis nous avons réalisé la mise à jour de la carte supervisée la plus récente (2017) sans utiliser les images de l'année courante (2018).

Les systèmes de fusion

Contrairement aux approches d'adaptation de domaine, la légèreté des systèmes de fusion a permis d'exploiter toutes les approches proposées. La seule exception concerne le **VP**, où chaque votant est une image pour laquelle chaque canal correspond à la probabilité déterminée par le RF pour chaque classe. La production d'une telle image entraîne une augmentation du temps de calcul très importante, qui nous a contraint à l'écartier dans un premier temps.

Les **VM** et **VC** exploitent les sorties des classifications naïves déjà réalisées. Ainsi nous devons simplement réaliser la fusion des classifications naïves, puis réaliser la validation à l'aide de **iota**². Ces deux votes ne requièrent aucun traitement particulier, comme la gestion des tuiles ou des régions éco-climatiques. En effet, ces aspects sont pris en considération lors de la classification naïve, et nous pouvons donc travailler directement sur la carte composite finale (qui couvre l'intégralité de l'emprise).

Dans la mesure où nous ne disposons que de deux années d'historique, ces deux approches vont produire un nombre important de non-décisions. Pour ces deux approches de vote (**VM** et **VC**), nous utilisons l'approche de transfert des étiquettes de la carte la plus récente, pour régler le problème des non-décisions. Là encore, la gestion des régions éco-climatiques demeure transparente.

Nous utilisons également la MT pour pallier le problème des non-décisions dans le cadre du **VM**. Pour cette approche, nous devons séparer les traitements par région éco-climatiques afin d'assurer que la diversité et les habitudes liées à ces régions soient les mieux représentées. Nous estimons donc 7 MT et résolvons les problèmes de non-décisions dans l'intégralité de la carte avant de réaliser un filtrage sur les régions éco-climatique pour obtenir la carte finale. Nous comparons ce découpage avec l'estimation d'une seule MT couvrant l'ensemble des régions.

Pour exploiter la fusion par Dempster-Shafer, nous produisons les classifications naïves croisées entre les domaines Source de l'historique (2016 → 2017 et 2017 → 2016). Ces deux exécutions nous permettent ainsi d'associer une matrice de confusion aux classifications naïves du domaine Cible (2018). Devant la forte similarité des résultats nous n'avons expérimenté que les masses de croyance issues de la **PG** et de la précision.

À partir des cartes supervisées, de la MT et des classifications naïves, nous avons exploité le classifieur bayésien. Dans la mesure où celui-ci exploite les MT, nous comparons le découpage en régions éco-climatiques avec une matrice unique pour toutes les régions.

Le classifieur unique

Enfin, nous avons expérimenté l'approche du classifieur unique à partir des échantillons d'apprentissage extraits lors des différentes classifications supervisées. Nous utilisons uniquement les échantillons des domaines Source (et non pas ceux de l'année courante) pour entraîner les sept classifieurs, respectant ainsi le découpage en régions éco-climatiques.

10.4 Comparaison des différentes approches

10.4.1 Performances obtenues

La table 10.3 présente les performances obtenues pour nos expérimentations. Dans cette table, nous trouvons les deux indices globaux : la **PG** et le kappa de Cohen, ainsi que le **FScore** minimal obtenu et la classe associée. La figure 10.3 présente sous la forme d'une matrice, le Fscore pour les différentes méthodes (en

ordonnées) en fonction des classes de la nomenclature (en abscisses). Pour des contraintes de temps, nous n'avons pas pu réaliser différents tirages aléatoires des échantillons d'apprentissage, nous privant ainsi des intervalles de confiance. Nous ne considérerons donc que les tendances observées, sans établir d'ordre strict de performances entre les méthodes.

Nous retrouvons dans la table 10.3, la relative constance des résultats que nous avons observée avec le jeu Formosat. Cependant, nous notons que l'écart entre le cas supervisé et les cas naïfs se réduit considérablement. Cette amélioration peut provenir de deux sources. La première concerne la différence du nombre d'images disponibles avant ré-échantillonnage, impactant ainsi la qualité de l'information apportée par les profils temporels. La grande diversité observée sur l'ensemble d'une région éco-climatique constitue le deuxième facteur d'amélioration des performances de la classification naïve. Cette tendance semble être confirmée par les Fscore mesurés. En effet ceux-ci sont relativement stables quelque soit la méthode employée. Les colonnes de la figure 10.3 illustrent, par des couleurs homogènes, cette stabilité. Nous pouvons noter que, contrairement aux expérimentations Formosat-2, les performances pour les classes de cultures demeurent constantes pour toutes les méthodes.

La seule grande exception à ce constat demeure l'approche de mise à jour de la carte de l'année passée (MAJ). En effet, comme attendu, cette approche ne permet d'appréhender la rotation des cultures avec un historique aussi peu profond. Cependant, elle apporte une bonne contribution à la reconnaissance de classes pérennes comme les routes ou les vergers. Nous continuons cette analyse des performances, mais sans prendre en compte cette approche par la suite.

Pour toutes les méthodes proposées, trois classes représentent un défi : les Vergers, les Landes ligneuses et les Plages et dunes. La classe de Bâti dense ne fait pas partie de celles-ci car les performances demeurent similaires pour toutes les méthodes sans exception, avec un Fscore aux alentours de 12 %. La classe de Surfaces routes atteint des performances similaires, cependant elle ne varie pas autant en fonction des différentes méthodes. En effet, pour ces trois classes, nous observons des variations d'environ 10 % dans la mesure de Fscore. Ces variations constituent le facteur majeur de différences entre les performances globales observées dans la table 10.3. Notamment, le classifieur unique possède une PG supérieure à celles des classifications naïves, grâce à la bonne prédiction de la classe Plages et dunes avec un Fscore de 62 %, contre 33 % et 24 % pour les cas naïfs. Ces différences de Fscore semblent indiquer que le classifieur unique ne constitue pas une simple copie d'un des deux cas naïfs mais bien un classifieur différent.

La réduction de l'écart entre les classifications supervisées et naïves réduit également la marge de progression apportée par les systèmes de fusion. Que ce soit le classifieur bayésien ou les votes majoritaires pondérés, ces approches fournissent des performances similaires, que ce soit en termes de PG ou de Fscore. Le VM seul, et dans une moindre mesure le VC, produit de nombreux cas d'indécisions. Ceux-ci concernent essentiellement les classes déjà identifiées comme étant difficiles à classer (Plages et dunes, Landes ligneuses, Vergers). Nous constatons que l'apport de la matrice de transition résout ces cas d'indécisions, tout en n'offrant qu'une légère amélioration des performances par rapport aux classifications naïves.

La faible différence entre les classifications naïves et le classifieur unique, semble indiquer que la diversité spatiale fournit autant d'informations que la diversité temporelle. Cette hypothèse requiert cependant un historique plus profond (plus de domaines Source) pour être confirmée.

L'écart de performances des classifications naïves constitue un résultat intéressant. En effet, la paire d'années (2016, 2018) produit de meilleures performances que la paire (2017, 2018). Nous savons que 2016 et 2018 sont des années très similaires et qu'il y a des rotations de cultures alternant tous les deux ans comme le

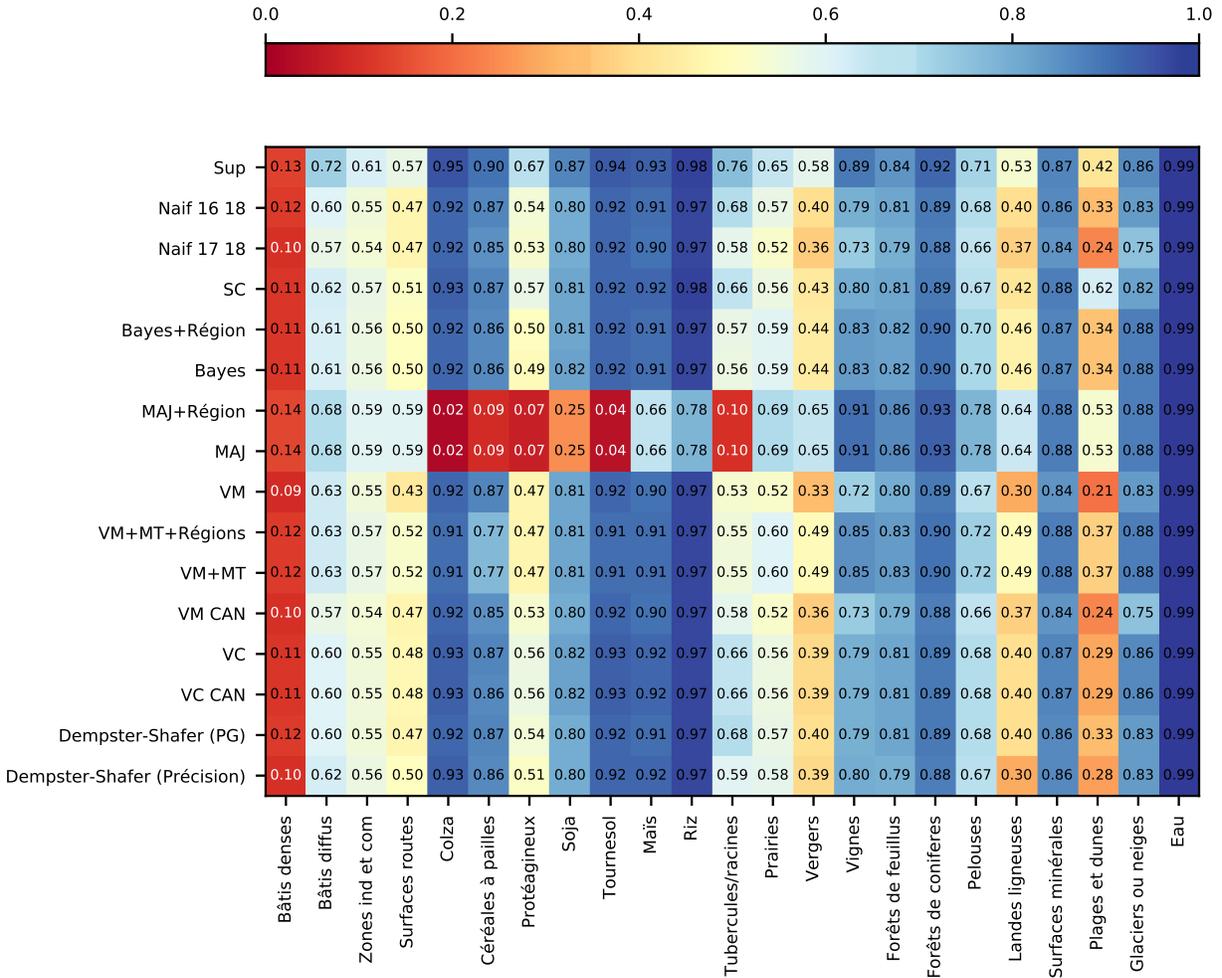


FIGURE 10.3 : FScore obtenus par les différentes expérimentations.

TABLE 10.3 : Performances obtenues par les différentes méthodes proposées sur la zone Grand Sud. Cette table recense la PG, le kappa, le FScore minimal et la classe associée.

Méthode	PG	kappa	FScore min	Classe FScore min
Supervisée	0.835	0.814	0.131	Bâti dense
Naïve 2016	0.804	0.778	0.119	Bâti dense
Naïve 2017	0.788	0.76	0.105	Bâti dense
VM	0.75	0.72	0.088	Bâti dense
VM + MT	0.82	0.796	0.120	Bâti dense
VM + MT + régions	0.82	0.79	0.120	Bâti dense
VM et transfert d'étiquettes	0.788	0.76	0.105	Bâti dense
VC	0.80	0.77	0.111	Bâti dense
VC et transfert d'étiquettes	0.805	0.779	0.112	Bâti dense
Dempster-Shafer (PG)	0.80	0.77	0.119	Bâti dense
Dempster-Shafer (précision)	0.79	0.77	0.105	Bâti dense
Classifieur Unique	0.81	0.787	0.109	Bâti dense
Classifieur Bayésien	0.817	0.792	0.114	Bâti dense
Classifieur Bayésien + régions	0.817	0.792	0.114	Bâti dense
Mise à jour	0.759	0.728	0.015	Colza
Mise à jour + régions	0.75	0.72	0.015	Colza

Tournesol et le Blé. Ces deux raisons permettent d'expliquer cette différence mais il s'agit d'une coïncidence. Celle-ci permet néanmoins de mettre en évidence l'apport des méthodes de fusion qui améliorent la précision des cartes produites, par rapport à la classification naïve 2016, malgré les fortes différences climatiques entre 2017 et 2018.

Nous notons le manque d'influence, lors de l'utilisation de la matrice de transition, du découpage en zone éco-climatique, qui pourtant devrait avoir un fort impact. Néanmoins, la classification naïve prend déjà en compte cette stratification, ce qui peut expliquer ce manque d'impact. Une deuxième explication possible concerne la clef de transition qui ne contient qu'une classe. Un historique plus profond pourrait augmenter l'impact de la stratification lors de l'utilisation de la MT.

La stabilité des performances pour les classes annuelles constitue un résultat très intéressant. En effet, contrairement à ce que nous avons observé lors des précédentes expérimentations, les Fscore associées à celles-ci demeurent supérieurs à 80 %. Là encore, nous pouvons mettre en avant la forte diversité offerte par la couverture de la zone traitée, qui permet aux classificateurs de mieux appréhender les distorsions entre les années.

10.4.2 Rendu visuel des approches de classification

La table 10.3 et la figure 10.3 présentent des performances similaires pour la majorité des méthodes proposées. Un facteur important reste l'absence de différence dans ces métriques lors du découpage en région éco-climatique, notamment pour les approches exploitant les MT. Or, cette exacte correspondance des performances entre l'utilisation ou non des zones éco-climatiques, provient de l'effet de masse associé à ces

métriques. En effet, dans les matrices de confusion, nous distinguons des variations du nombre d'échantillons dans les différentes cases. Cependant, ces variations ne concernent pas une quantité suffisante d'échantillons pour être observables dans les métriques (du moins avec une précision de 10^{-3}).

Les figures 10.5 et 10.6 regroupent des illustrations des cartes d'OCS produites à partir des 16 cas différents présentés dans ce chapitre. Les couleurs correspondent à la nomenclature présentée dans la figure 10.4. La figure 10.5 illustre les résultats obtenus après une étape de classification ou de mise à jour, alors que la figure 10.6 présente les approches de fusion de classifieurs. La zone présentée correspond au sud ouest de Toulouse et contient à la fois des zones agricoles et urbaines, ainsi que diverses surfaces, comme des plans d'eau, des routes et zones industrielles.

Ces vignettes permettent d'illustrer les différences de production entre les méthodes. Ainsi, nous pouvons observer la forte confusion entre les classes de Bâti diffus et Zones industrielles (il n'y a pas de pixels étiquetés Bâti dense dans cette zone). Un second constat porte sur la perte des contours distincts des parcelles agricoles, visibles sur la classification supervisée, notamment dans le centre Ouest de la zone pour les classifications naïves. Le classifieur unique produit une carte de qualité visuelle proche de la classification supervisée, cependant plusieurs petits éléments (mais aussi des parcelles entières) disparaissent.

Nous ne passerons pas en revue toutes les méthodes et leurs différences pixels à pixels. Cependant, nous pouvons constater les erreurs flagrantes commises par l'approche de mise à jour (MAJ), notamment sur les cultures circulaires dans le centre Ouest de l'image. Inversement, nous retrouvons une partie du réseau routier, classé en zone industrielle dans la partie Nord de l'image.

La figure 10.6 présente le cas intéressant du VM et des non-décisions, qui apparaissent en blanc dans ces illustrations. Nous pouvons voir, grâce au VM que les classifications naïves demeurent cohérentes entre elles sur les grands objets (fleuve, lac, parcelles agricoles) mais divergent sur les bordures ainsi que la classification d'éléments urbains (confusions entre les bâtis). Ce résultat renouvelle le fort intérêt du VM, dans la mesure où il exploite la classification naïve et constitue une étape de traitement très légère en comparaison à d'autres approches. Pour le VC, les non-décisions restent très éparées, rendant difficile leur visualisation.

Ces visualisations permettent de mettre en évidence les différences entre les méthodes de résolutions des cas de non-décisions. Cet effet est particulièrement visible sur les parcelles agricoles aux limites Nord du centre de l'image. En effet, l'utilisation de la MT permet de détecter une partie de la parcelle agricole, alors que les classifications naïves fournissent des étiquettes de bâtis, vignes et cultures. Nous pourrions tirer avantage de ce genre d'amélioration en la combinant avec une approche de segmentation pour déterminer la taille de l'objet, ou en appliquant une régularisation spatiale sur les cartes ainsi produites.

Nous pouvons voir, grâce à ces illustrations, que le rendu visuel reste cohérent avec les métriques calculées. Toutes ces cartes se ressemblent fortement, malgré de légères différences. Pour des PG et Fscore, ces différences restent difficilement visibles. L'annexe F.1 représente une carte de différences sur cette zone entre les classifications produites par le classifieur bayésien et le classifieur unique. Les cas de divergences apparaissent en blanc, en cas de cohérence l'OCS est représentée. Nous observons une concentration de ces cas de différences sur les bordures de parcelles, ainsi que sur les petits éléments de la carte.

	Bâti denses		Tubercules/Racines
	Bâti diffus		Prairies
	Zones industrielles et commerciales		Vergers
	Surfaces routes		Vignes
	Colza		Forêts de feuillus
	Céréales à pailles		Forêts de conifères
	Protéagineux		Pelouses
	Soja		Landes ligneuses
	Tournesol		Surfaces minérales
	Maïs		Plages et dunes
	Riz		Glaciers et neiges
	Eau		

FIGURE 10.4 : Nomenclature à 23 classes pour la classification des images Sentinel-2. Les couleurs correspondent à celles utilisées dans les cartes.

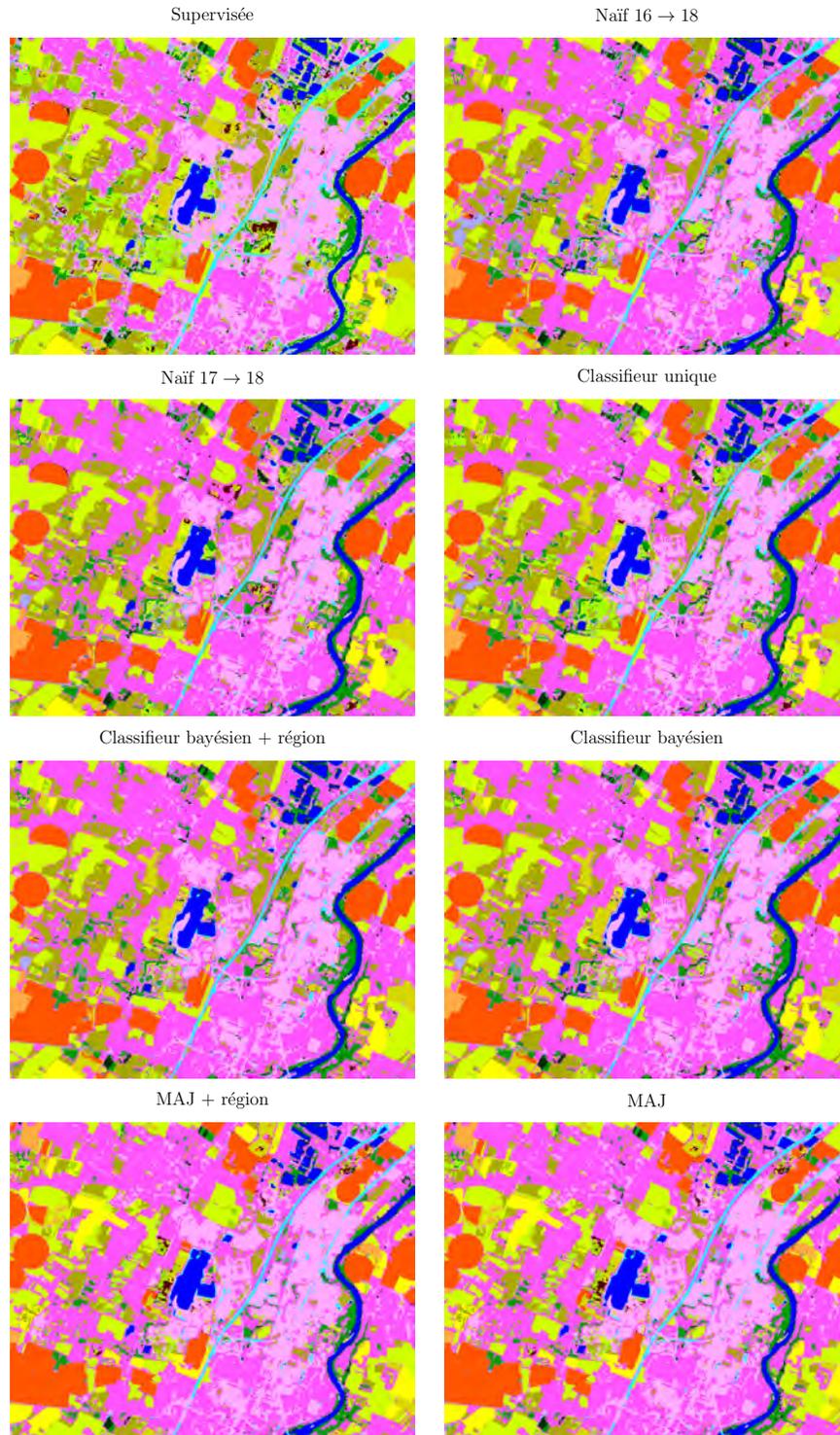


FIGURE 10.5 : Illustration des cartes générées à partir des données Sentinel-2 et les différentes méthodes proposées (Classifications).

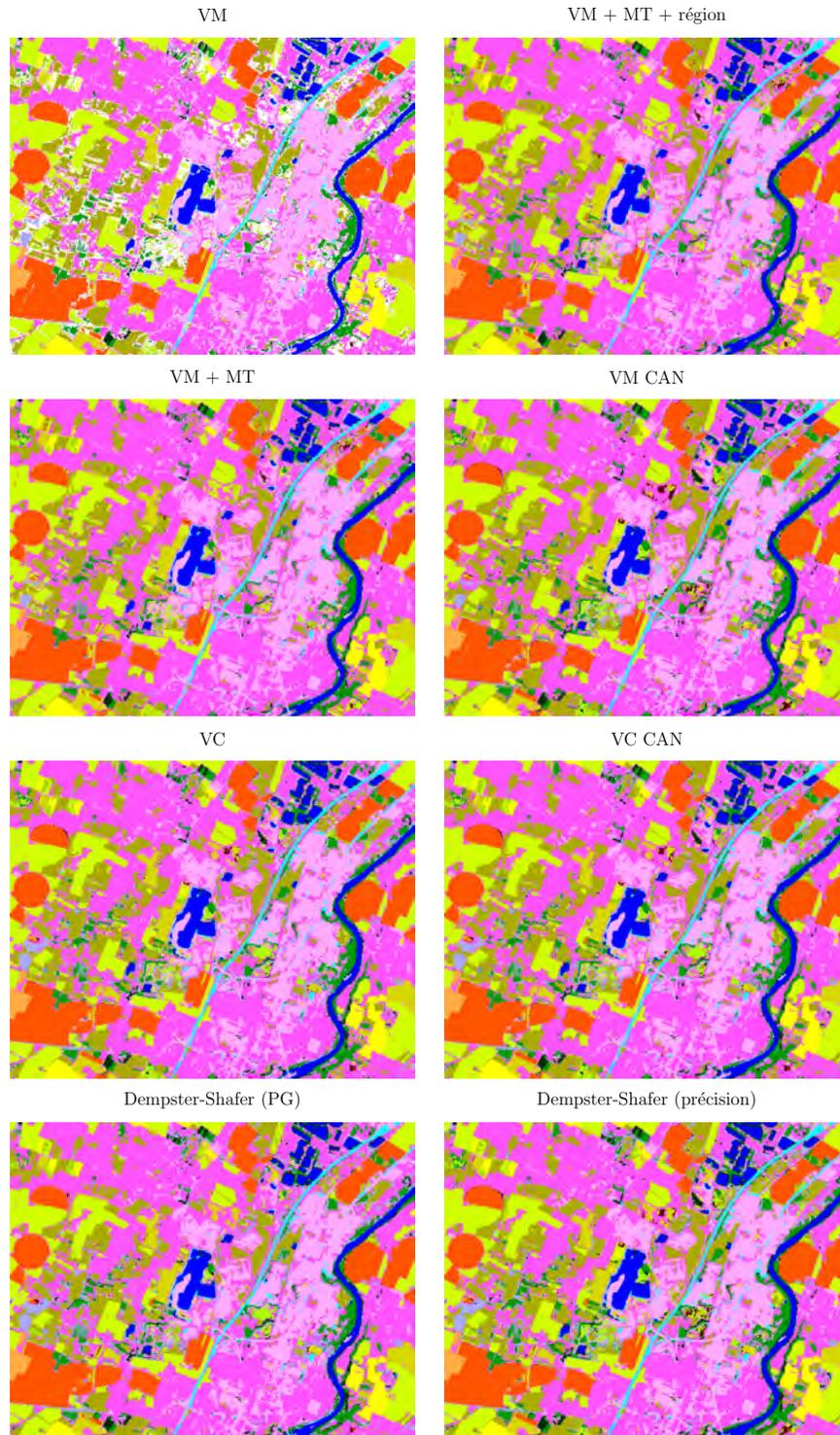


FIGURE 10.6 : Illustration des cartes générées à partir des données Sentinel-2 et les différentes méthodes proposées (Fusion de classifieurs).

10.4.3 Évaluation des coûts

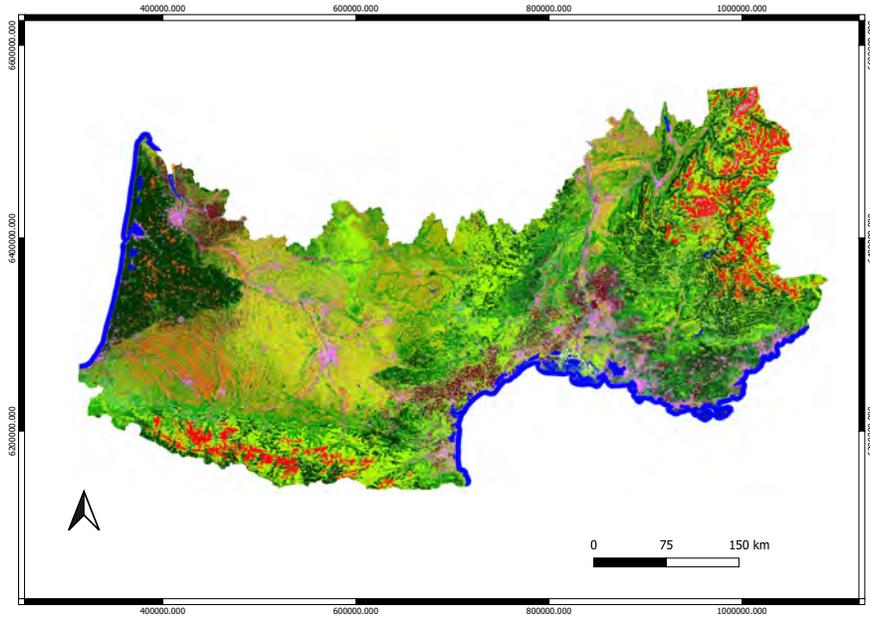


FIGURE 10.7 : Carte d'occupation des sols de 2018 couvrant le sud de la France métropolitaine continentale.

Similairement au jeu de données Formosat-2, nous pouvons mettre en relation les coûts de mise en oeuvre et la performance globale obtenue. Une fois encore cette évaluation dépend des données utilisées en entrée des différentes étapes. Les temps mesurés correspondent aux différentes étapes de la chaîne *iota*², chacune de ces étapes contient tous les pré-requis nécessaires. Par exemple, avant de réaliser la classification, toutes les bandes spectrales de toutes les images sont ré-échantillonnées pour correspondre entre elles, en termes de résolution spatiale et temporelle. Néanmoins, nous avons déterminé les relations suivantes :

- $c = 20 \times a$, le temps de classification (1 tuile)
- $p = \frac{1}{9}a$, le temps de post-traitement (1 tuile)

Pour générer une carte de l'étendue présentée dans la figure 10.7, les relations sont légèrement différentes. En effet, le découpage en région éco-climatique, impacte les temps d'apprentissage, proportionnellement au nombre d'échantillons, que l'on peut corréliser au nombre de tuiles composant la région. Nous notons R le nombre de régions éco-climatiques considérées. Le processus de classification est plus complexe à déterminer, nous utilisons la variable C pour indiquer le nombre de classifications opérées. En effet, le nombre de classifications d'une tuile équivaut au nombre de régions éco-climatiques couvertes par celle-ci. Une tuile étant alors classée par chaque classifieur éligible.

Le temps de post-traitement est calculé directement sur la mosaïque d'image. Nous différencions les méthodes exploitant la MT qui utilise le découpage en zone éco-climatique et les autres exploitant l'emprise totale. Ainsi, nous obtenons de nouvelles équations, présentées dans la table 10.4. Nous avons déterminé les relations suivantes :

- $c = 10a$, le temps de classification (1 tuile)
- $p = a$, le temps de post-traitement (emprise totale)

TABLE 10.4 : Évaluation de l'effort de production des différentes méthodes

Méthode	Apprentissage	Classification	Post-Traitement	Temps (en fonction de a)
Supervisé	Ra	Cc	0	$(R + 10C)a$
Naïf	Ra	Cc	0	$(R + 10C)a$
MAJ	RNa	CNc	p	$((R + 10C)N + 1)a$
VM	RNa	CNc	p	$((R + 10C)N + 1)a$
VC	RNa	$2CNc$	$2p$	$((R + 20C)N + 2)a$
Correction non-décisions	RNa	CNc	$2Rp$	$((R + 10C)N + 2R)a$
Dempster-Shafer	RNa	$C(N^2 - N)c$	$((N^2 - N) + 1)p$	$(RN + (10C + 1)(N^2 - N) + 1)a$
Classifieur Unique	RNa	Cc	0	$(RN + 10C)a$
Classifieur Bayésien	RNa	$C \times 2Nc$	$2Rp$	$((R + 20C)N + 2R)a$

À partir de ces équations, nous pouvons réaliser un graphe de performances, présenté dans la figure 10.8. Pour qu'il soit comparable à celui précédemment présenté pour Formosat-2, nous avons estimé le temps total maximal, pour la méthode la plus coûteuse (la fusion par Dempster-Shafer) avec 6 années d'historique. Comme observé dans les tables de performances, l'écart entre les classification naïves et supervisées est réduit. Les approches de fusion par Dempster-Shafer et VC, exploitant deux années d'historique, offrent des performances et un coût de production similaires. Le VM combiné avec la MT, ainsi que le classifieur unique, offrent les meilleures performances pour des coûts raisonnables.

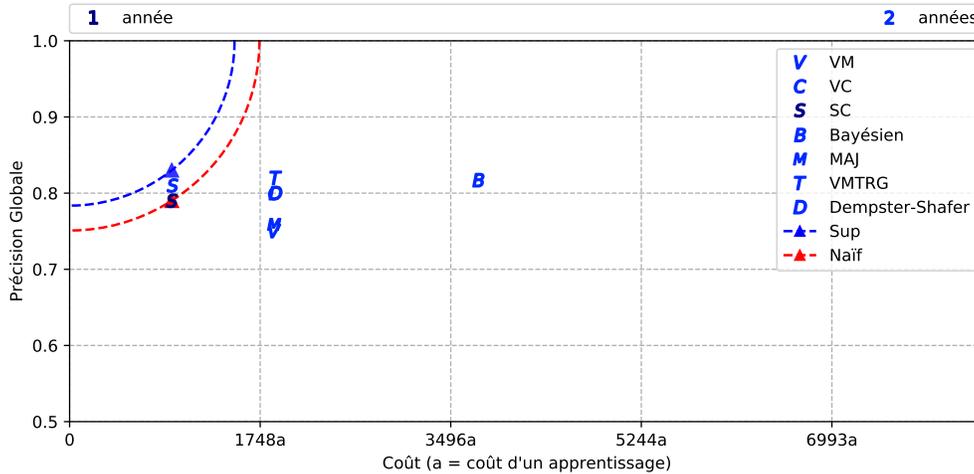


FIGURE 10.8 : Évaluation des coûts pour différentes méthodes en fonction du temps d'un apprentissage et différents nombres d'années utilisées.

10.5 Un début prometteur

Pressés par les échéances, nous n'avons pas pu réaliser plus d'expérimentations à partir des données Sentinel-2.

Cependant, les premiers résultats obtenus s'annoncent prometteurs, car ils suivent les mêmes tendances que les résultats analysés à partir des données Formosat-2.

Ainsi, nous retrouvons les situations favorables à l'application des méthodes proposées comme la très bonne reconnaissance des classes majoritairement représentées dans les données de référence. Nous retrouvons également la bonne prédiction des classes caractéristiques comme l'eau par exemple.

Nous retrouvons également les cas défavorables, comme la séparation de classes très similaires, ici le bâti (dense et diffus, ainsi que les zones industrielles). Les classes peu représentées dans la donnée de référence, représentent également une situation complexe à gérer avec une moins bonne qualité de prédiction (plages et dunes, landes ligneuses, vergers).

Grâce à la confirmation de ces tendances, nous pouvons donc estimer qu'avec un historique plus profond, l'écart entre les méthodes proposées et la classification supervisée sera réduit par certaines approches, comme le classifieur unique ou le **VM** couplée avec une solution efficace du problème des non-décisions.

Ces expérimentations permettent de mettre en lumière le problème de la gestion de données. En effet, l'intégralité de l'historique occupe un volume conséquent.

La génération de l'historique constitue le premier obstacle rencontré. En effet, pour cela nous devons accéder aux séries temporelles, ou à défaut aux images des domaines Source pour générer les échantillons et entraîner les classifieurs. Sur la trentaine de tuiles exploitées, le volume considéré nécessite une dizaine de teraoctets de stockage.

La possibilité d'appliquer une méthode à un sous-ensemble des données a représenté un second défi. En effet, appliquer à l'intégralité des séries temporelles, composant l'intégralité de la zone d'étude, une méthode s'avère complexe. Ainsi nous n'avons pas pu exploiter l'**ACC** par exemple. De même, l'application du **TO** a requis trop de temps pour être envisageable dans les délais impartis. Néanmoins, cette situation a permis de mettre en avant la simplicité d'utilisation et les bonnes performances des systèmes de fusion. En effet, la plupart de ces méthodes interviennent seulement au niveau du pixel, favorisant ainsi l'utilisation des approches de programmation parallèle telles que le streaming et le multi-threading.

Parmi les méthodes de fusion, les plus simples à exploiter restent les **VM** et **VC** (sous réserve de pouvoir régler les non-décisions) ainsi que le classifieur unique. Ces trois approches permettent de réduire également la quantité de données à conserver obligatoirement. En effet, les votes majoritaires pondérés requièrent uniquement la conservation des classifieurs (quelques mégaoctets de données dans le cadre des **RF**). Cependant pour que les classifieurs demeurent exploitables, nous devons prêter une attention particulière à la préparation des données, en imposant par exemple une grille temporelle fixe qui restera identique pour chaque nouvelle acquisition de domaine.

Le classifieur unique lui, requiert uniquement la conservation des échantillons d'apprentissage. De plus, nous pouvons envisager de stocker les données originales (sans ré-échantillonnage) et ainsi d'adapter ces échantillons au moment de l'apprentissage en spécifiant la grille temporelle souhaitée.

Ainsi, le classifieur unique représente la méthode permettant la plus grande liberté d'utilisation, tout en produisant des cartes associés de très bonne qualité.

Sixième partie

Conclusions

11.1 Bilan

L'objectif général de la thèse consiste à produire des cartes d'**OCcupation des Sols (OCS)** automatiquement, sans utiliser les données de référence de la période correspondante. En effet, l'exploitation des approches de classification supervisée excellent dans ce domaine, mais ces performances requièrent des données de référence. Les délais très longs de mise à disposition des données de référence posent un problème majeur que nous avons identifié.

Heureusement, l'étude de l'**OCS** requiert une acquisition régulière de données (images et références) sur de nombreuses années, permettant ainsi la constitution d'un historique sur la zone voulue. Cet historique permet d'observer l'évolution de l'**OCS** au fil des années. Il permet donc de déterminer les caractéristiques et les habitudes d'exploitation pour la zone concernée. Cependant, les données de l'historique ne possèdent qu'une période de validité limitée à celle d'acquisition. En effet, les données de référence représentent un relevé ponctuel de l'**OCS**. De nombreuses classes d'**OCS** changent annuellement (classes annuelles), notamment celles de cultures, rendant ainsi les données de référence obsolètes. De plus, différents facteurs, comme le nombre d'acquisitions et la couverture nuageuse, impactent la qualité et le contenu de la série temporelle annuelle, causant une première distorsion à prendre en compte pour exploiter les images de l'historique. Le climat représente un second facteur de distorsion. En effet, celui-ci gouverne le développement de la végétation, impliquant un possible décalage dans le profil temporel de ces classes. Ces deux distorsions constituent les paramètres les plus importants à considérer pour l'utilisation de données couvrant plusieurs années.

Les travaux menés dans cette thèse proposent d'exploiter cet historique afin de produire une carte d'**OCS** du millésime, sans exploiter les données de référence de la période correspondante.

Ce sujet se situe à l'intersection de deux grandes problématiques : la production opérationnelle de cartes annuelles sur de grandes étendues pour lesquelles les données de référence ne sont pas encore disponibles et la production de cartes sur des années passées, pour lesquelles les données de référence ne seront jamais disponibles.

Grâce à un riche jeu de données de 7 années de séries temporelles d’images Formosat-2 et des données de référence correspondantes, nous avons pu proposer différentes approches permettant de produire des cartes d’OCS, tenant compte des deux configurations du problème. En effet, en considérant une approche un contre tous, une des années constitue l’année à classer (sans données de référence) et toutes les autres représentent l’historique. Dès lors, les méthodes proposées doivent tenir compte de la taille de l’historique, afin d’étudier les différentes configurations et répondre au mieux aux diverses situations possibles. Ainsi, nous avons divisé les approches des travaux de thèse selon deux axes, l’exploitation d’une seule année passée pour produire la carte, et l’exploitation de plusieurs années.

Dans un premier temps, nous avons exploité un classifieur déjà entraîné sur une année de l’historique. Nous avons baptisé cette approche classification naïve, car elle ne prend en compte aucune distorsion possible entre les années. Assurer la cohérence des canaux entre les séries temporelles constitue la seule contrainte requise par la classification naïve. Nous exploitons donc un ré-échantillonnage temporel afin que toutes les séries temporelles soient alignées sur une grille régulière identique à toutes les années. Cette approche simple a permis de définir une marge de progression de l’ordre de 20 % entre les classifications naïve et supervisée.

L’adaptation de domaine constitue un champ d’expertise offrant des perspectives de solutions, car elle est très exploitée sur des problématiques similaires. En effet, en adaptation de domaine, on considère un domaine Source, qui possède des données de référence, et un domaine Cible, qui n’en possède pas. Le principe général vise alors à minimiser les différences entre les domaines afin d’exploiter le domaine Source pour classer le domaine Cible. Cependant, un rapide constat met en évidence des différences importantes entre les cas présentés dans les travaux existants et notre problème, notamment en termes de la dimension des données (séries temporelles annuelles) et le nombre de classes. Nous avons exploité deux approches d’adaptation de domaine : la projection via l’Analyse Canonique des Corrélations (ACC) et le Transport Optimal (TO).

La projection ACC permet de manipuler les séries temporelles originales (sans ré-échantillonnage temporel), afin de déterminer un espace dans lequel on minimise les distorsions entre les années. L’ACC permet de déterminer une matrice de projection pour chaque série temporelle. Ainsi, en entraînant un classifieur à partir des données du domaine Source nous pouvons l’appliquer au domaine Cible. Nous avons montré que l’exploitation de l’ACC sur des séries temporelles annuelles, n’apporte pas de meilleures performances qu’une classification naïve. Des études complémentaires ont mis en évidence, que les changements d’OCS géo-localisés provoquent des dégradations dans la qualité de la projection. En effet, en estimant la projection entre deux séries temporelles représentant la même scène, la qualité de la classification après projection s’améliore. Un problème majeur reste la grande quantité de changements entre deux années, ainsi que le volume à traiter dans le cas de grandes étendues.

Le TO est une approche permettant d’estimer une transformation (le transport) réalisant une mise en correspondance entre les deux domaines. À l’aide d’une matrice de coût et d’algorithmes de minimisation, la transformation obtenue permet de projeter les données du domaine Source dans le domaine Cible à moindre coût. Loin des cas théoriques observés dans la revue bibliographique, nous avons mis en lumière l’importance de la sélection des échantillons utilisés pour l’estimation du TO. En effet, les performances obtenues lors de la validation des cartes produites après TO varient de plus de 70 % en moyenne, en fonction de l’approche de sélection. Dans un contexte d’application concret, nous obtenons les meilleures performances du TO lorsque celui-ci est estimé pour chaque classe d’OCS individuellement.

Nous avons mis en évidence, que l’apport du TO demeure minime par rapport à une classification naïve. En effet, sa contribution majeure consiste à mieux appréhender les artefacts laissés par le ré-échantillonnage

temporel dans les cas de longues absences de données. Le coût requis pour l’exploitation du **TO** dans une chaîne de production opérationnelle demeure considérable, surtout pour une exploitation à grande échelle, mais également au vu du faible impact sur la qualité de la carte.

Dans un second temps, nous avons proposé des méthodes exploitant plusieurs années de données de l’historique. Une première approche consiste à modéliser la succession des classes d’**OCS** à partir des cartes supervisées. Nous proposons l’exploitation d’une **Matrice de Transition (MT)** permettant d’associer une probabilité de transition entre une classe (la prédiction) et une succession de classes (l’historique). Cette approche permet de mettre à jour la carte d’**OCS** la plus récente, par une prédiction sur le maximum de probabilités, sans utiliser les images du millésime. Les performances en demi-teinte mettent en évidence un système de prédiction trop faible pour appréhender convenablement la forte complexité des changements d’**OCS** entre les années. Cependant, elle permet, avec un effort minimal, de prédire avec succès les transitions induites artificiellement comme la rotation des cultures. Néanmoins, la qualité de l’exploitation de la **MT** dépend fortement de la composition de l’historique et des habitudes de transition des classes. En effet, pour notre jeu de données, une configuration optimale de l’historique permet de modéliser des rotations de cultures tri-annuelles, qui sont les plus fréquentes dans la zone étudiée.

Une seconde approche consiste à entraîner un unique classifieur, à partir des échantillons d’apprentissage des périodes disponibles dans l’historique. Le principe général est alors de permettre au classifieur de déterminer les primitives les plus pertinentes pour prendre en compte les distorsions entre les années. Cette approche de classifieur “multi-naïf” produit des cartes de bonne qualité, pour un coût d’apprentissage proportionnel au nombre d’années utilisées. Le résultat le plus notable obtenu par le classifieur unique consiste à améliorer le taux de bonne classification pour les classes les moins représentées dans les données de référence. Cet effet provient en grande partie de l’augmentation de la représentation de cette classe grâce à l’historique.

Une troisième approche consiste à exploiter les forces de la classification naïve. En effet, la comparaison entre toutes les combinaisons possibles d’années montre que les performances restent variables, et sous certaines conditions, la classification naïve permet une très bonne prédiction de classes minoritaires. Ainsi, nous avons mis en place des systèmes de vote majoritaire pondéré, permettant une combinaison optimale des classifications naïves (et issues du **TO** le cas échéant). Nous avons pu comparer trois votes pondérés : le **Vote Majoritaire (VM)**, le **Vote par Confiance (VC)** et le **Vote par Probabilités (VP)**, exploitant les informations d’étiquettes, de la confiance du classifieur dans sa propre décision, et enfin un tri ordonné des étiquettes par le classifieur associé à une valeur de probabilité par classe.

Nous avons également exploité la fusion de Dempster-Shafer qui permet d’introduire une notion de croyance pour chaque classification naïve utilisée dans le vote. Cette pondération permet d’appuyer la décision des classifieurs les plus pertinents. Pour obtenir cette pondération, nous réalisons la classification naïve de toutes les périodes disponibles dans l’historique. Cette étape nous permet alors d’associer une matrice de confusion moyenne à chaque classifieur et donc à chaque votant.

Enfin la dernière approche proposée consiste à réaliser une fusion exploitant une règle de décision bayésienne. Le principe de cette approche vise à retarder la décision de chaque élément proposé par les autres méthodes afin de prendre une unique décision dans un modèle intégré. Ainsi, nous pouvons fusionner les différentes classifications naïves, la **MT**, et les proportions a priori afin de prédire la carte d’**OCS** du millésime.

Toutes ces approches offrent des performances similaires, exception faite de la mise à jour de la carte, avec de légères variations observées avec l’analyse statistique des performances par classe (**FScore**). Ainsi, élire une unique approche “souveraine” s’avère complexe, d’autant plus que les performances mesurées proviennent

d'un unique jeu de données et que la composition de l'historique influence les coûts de production.

Nous avons pu bénéficier des données Sentinel-2 pour réaliser un premier passage à l'échelle de la plupart des méthodes proposées. Grâce à cela nous avons pu constater que les mêmes tendances et comportements des méthodes observés sur le jeu de données Formosat-2 se retrouvent lors de l'exploitation des données Sentinel-2. Cette comparaison ne peut pas être exhaustive, dans la mesure où seulement trois années de données Sentinel-2 sont actuellement disponibles. Néanmoins la grande qualité du jeu de données, ainsi que la diversité spatiale offerte par Sentinel-2, fournissent une information équivalente à celle offerte par la diversité temporelle de Formosat-2.

Ainsi, nous pouvons affirmer que la diversité au sein de l'historique, qu'elle soit spatiale ou temporelle, représente un atout considérable pour la production de cartes d'OCS sans utilisation des données de référence du millésime. Nous avons montré que les méthodes d'adaptation de domaine [Tuia et al., 2016], souvent très complexes, s'avèrent moins performantes dans ce cadre applicatif, à moins de disposer d'un historique diversifié. Ainsi, un simple VM associé à l'exploitation d'une MT permet de surpasser, en termes de métriques statistiques, les autres méthodes. Le classifieur unique constitue également une approche viable pour la production de cartes d'OCS.

Nous avons identifié deux paramètres primordiaux à considérer lors de l'exploitation de nos méthodes. Le premier concerne la définition des besoins d'utilisation de la carte d'OCS et notamment de la nomenclature. Nous avons montré que la fusion de classes post-classification à partir d'une nomenclature plus détaillée que celle visée, pouvait permettre d'obtenir de bonnes performances, si les erreurs commises par le classifieur sont cohérentes. Le second paramètre concerne le volume de données, et notamment la gestion de l'historique. En effet, en manipulant les données Sentinel-2 sur un tiers de la France, nous avons été confrontés à de sérieux problèmes de stockages, notamment pour la préparation des trois séries temporelles annuelles. Ce paramètre impacte le choix de la méthode exploitée. En effet, pour réaliser l'approche de classification naïves, une grille temporelle commune, et donc définie à l'avance, reste obligatoire afin de ne conserver dans l'historique que les classifieurs déjà entraînés et les cartes supervisées. À l'inverse, le classifieur unique requiert de conserver uniquement les échantillons d'apprentissage issus des séries temporelles de l'historique. Cette configuration permet de modifier à la volée la grille temporelle de ré-échantillonnage, mais contraint à une nomenclature figée dans le temps, sous peine de devoir générer à nouveau l'historique dans son intégralité.

La conclusion générale demeure que le choix d'une approche permettant de produire une carte d'OCS sans utilisation des données de référence de la période correspondante appartient à l'utilisateur en fonction des objectifs d'utilisation de la carte et des ressources disponibles, à cette fin nous avons fourni des critères permettant de guider la sélection d'une méthode adaptée.

11.2 Perspectives

La littérature offre un grand nombre d'approches différentes proposant une ébauche de solution à notre problème. Nous avons donc du réaliser des choix, afin de ne conserver qu'une ou deux méthodes par catégorie d'algorithme (adaptation de domaine, fusion de classifieur). Toutes les approches non-considérées constituent alors des perspectives méthodologiques. Néanmoins, cette section se concentre sur les améliorations et les applications des méthodes expérimentées dans ces travaux.

La classification naïve constitue le premier maillon de la plupart des approches proposées. Ainsi, l'amélioration de la qualité de celle-ci impacte obligatoirement les performances de toutes les autres méthodes.

L'arrivée de Sentinel-2 et sa grande qualité ont mis en avant cet aspect, qui représente une source d'amélioration à faible coût. L'exploitation de la procédure **SAP** constitue une des perspectives les plus intéressantes pour l'amélioration des classifications naïves. Une autre approche, non abordée faute de temps, consiste à exploiter des méthodes d'augmentation de données. Celles-ci, dans le cadre de la classification supervisée, améliorent les performances obtenues en réduisant le déséquilibre entre les classes lors de l'apprentissage. Le meilleur équilibre des classes observé dans le jeu de données Sentinel-2 permet d'appuyer également cette piste.

Les approches d'adaptation de domaine sont intéressantes dans le cas d'un historique pauvre (en diversité) ou subissant de très fortes distorsions.

Elles représentent également une bonne alternative pour des situations de changements de capteurs. En effet, nous avons écarté cette problématique au profit d'une exploitation continue d'un seul capteur, néanmoins cette situation se pose dès que le satellite arrive en fin de vie, ou que l'on cherche à traiter des images d'archives.

En effet, les utilisateurs optent de plus en plus pour la classification d'images passées afin d'étudier l'évolution de l'**OCS** sur le long terme. Ces problématiques vont demander la mise en correspondance entre des données d'anciens et de récents satellites. Les approches d'adaptation de domaine et de transfert d'apprentissage constituent une des meilleures alternatives potentielles à ces problématiques.

Les systèmes de fusion de classifieurs représentent une grande part de ces travaux de thèse. Néanmoins, nous avons choisi d'interrompre la prospection de certaines méthodes, tant la diversité des configurations possibles est vaste. En effet, sur le principe des votes majoritaires pondérés, le choix de la pondération à lui seul regroupe de nombreuses alternatives. Nous pouvons par exemple, ajouter une pondération par le **FScore** par classe, à l'image de la fusion Dempster-Shafer. Un autre exemple serait d'exploiter une mesure de similarité entre les domaines Source et Cible, qui devient alors une pondération de la décision de chaque classifieur.

L'utilisation de l'inférence bayésienne ouvre la porte à toutes ces possibilités. En effet, toutes les informations pouvant être traduites sous formes de probabilités devient un facteur potentiellement exploitable.

Enfin, de nombreuses pistes d'améliorations concernent la **MT**. En effet, l'utilisation de contraintes plus fortes, associées à la capacité d'appréhender des systèmes complexes comme la rotation des cultures permettraient d'améliorer l'approche de mise à jour de la carte. Ainsi, cela rendrait alors la production d'une première version de la carte du millésime dès les premiers mois de l'année courante.

Dans un contexte plus général, les travaux de thèse bénéficieraient des améliorations des méthodes de stockage et de manipulation de données. En effet, de telles innovations permettraient de conserver plus d'informations dans l'historique, et donc une amélioration des performances.

- [Alaibakhsh et al., 2015] Alaibakhsh, M., Emelyanova, I., Barron, O., Mohyeddin, A., and Khiadani, M. (2015). Multivariate detection and attribution of land-cover changes in the central pilbara, western australia. *International Journal of Remote Sensing*, 36(10) :2599–2621.
- [Alajlan et al., 2014] Alajlan, N., Pasolli, E., Melgani, F., and Franzoso, A. (2014). Large-scale image classification using active learning. *IEEE Geoscience and Remote Sensing Letters*, 11(1) :259–263.
- [Argyriou et al., 2007] Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. In *Advances in neural information processing systems*, pages 41–48.
- [Bahirat et al., 2012] Bahirat, K., Bovolo, F., Bruzzone, L., and Chaudhuri, S. (2012). A novel domain adaptation bayesian classifier for updating land-cover maps with class differences in source and target domains. *IEEE Transactions on Geoscience and Remote Sensing*, 50(7) :2810–2826.
- [Benamou and Brenier, 2000] Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3) :375–393.
- [Benzecri, 1977] Benzecri, J.-P. (1977). Analyse discriminante et analyse factorielle. *Cahiers de l'analyse des données*, 2(4) :369–406.
- [Biggio et al., 2007] Biggio, B., Fumera, G., and Roli, F. (2007). Bayesian analysis of linear combiners. In *International Workshop on Multiple Classifier Systems*, pages 292–301. Springer.
- [Bonissone et al., 2010] Bonissone, P., Cadenas, J. M., Garrido, M. C., and Díaz-Valladares, R. A. (2010). A fuzzy random forest. *International Journal of Approximate Reasoning*, 51(7) :729–747.
- [Bradski, 2000] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- [Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1) :5–32.
- [Brooks et al., 2014] Brooks, E. B., Wynne, R. H., Thomas, V. A., Blinn, C. E., and Coulston, J. W. (2014). On-the-fly massively multitemporal change detection using statistical quality control charts and landsat data. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6) :3316–3332.

- [Bruzzone et al., 2006] Bruzzone, L., Chi, M., and Marconcini, M. (2006). A novel transductive svm for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11) :3363–3373.
- [Bruzzone and Cossu, 2002] Bruzzone, L. and Cossu, R. (2002). A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. *Geoscience and Remote Sensing, IEEE Transactions on*, 40(9) :1984–1996.
- [Bruzzone et al., 2004] Bruzzone, L., Cossu, R., and Vernazza, G. (2004). Detection of land-cover transitions by combining multirate classifiers. *Pattern Recognition Letters*, 25(13) :1491–1500.
- [Bruzzone and Marconcini, 2009] Bruzzone, L. and Marconcini, M. (2009). Toward the automatic updating of land-cover maps by a domain-adaptation svm classifier and a circular validation strategy. *IEEE Transactions on Geoscience and Remote Sensing*, 47(4) :1108–1122.
- [Bruzzone and Persello, 2009] Bruzzone, L. and Persello, C. (2009). A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE Transactions on Geoscience and Remote Sensing*, 47(9) :3180–3191.
- [Bruzzone and Prieto, 2001] Bruzzone, L. and Prieto, D. (2001). Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 39(2) :456–460.
- [Bruzzone and Serpico, 1997] Bruzzone, L. and Serpico, S. (1997). An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4) :858–867.
- [Card, 1982] Card, D. (1982). Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering and Remote Sensing*, 48(3) :431–439.
- [Castellazzi et al., 2008] Castellazzi, M., Wood, G., Burgess, P. J., Morris, J., Conrad, K., and Perry, J. (2008). A systematic representation of crop rotations. *Agricultural Systems*, 97(1-2) :26–33.
- [Charlet et al., 2004] Charlet, J., Bachimont, B., and Troncy, R. (2004). Ontologies pour le web sémantique. *Revue I3, numéro Hors Série « Web sémantique*, pages 43–63.
- [Chen et al., 2012a] Chen, X., Chen, J., Shi, Y., and Yamaguchi, Y. (2012a). An automated approach for updating land cover maps based on integrated change detection and classification methods. *ISPRS Journal of Photogrammetry and Remote Sensing*, 71(nil) :86–95.
- [Chen et al., 2012b] Chen, X., Vogelmann, J. E., Chander, G., Ji, L., Tolk, B., Huang, C., and Rollins, M. (2012b). Cross-sensor comparisons between landsat 5 tm and irs-p6 awifs and disturbance detection using integrated landsat and awifs time-series images. *International Journal of Remote Sensing*, 34(7) :2432–2453.
- [Chi and Bruzzone, 2007] Chi, M. and Bruzzone, L. (2007). Semisupervised classification of hyperspectral images by SVMs optimized in the primal. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6) :1870–1880.

- [Claverie et al., 2012] Claverie, M., Demarez, V., Duchemin, B., Hagolle, O., Ducrot, D., Marais-Sicre, C., Dejoux, J.-F., Huc, M., Keravec, P., Béziat, P., Fieuzal, R., Ceschia, E., and Dedieu, G. (2012). Maize and sunflower biomass estimation in southwest france using high spatial and temporal resolution remote sensing data. *Remote Sensing of Environment*, 124(nil) :844–857.
- [Clemen, 1989] Clemen, R. T. (1989). Combining forecasts : A review and annotated bibliography. *International journal of forecasting*, 5(4) :559–583.
- [Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1) :37–46.
- [Congalton and Green, 2002] Congalton, R. G. and Green, K. (2002). *Assessing the accuracy of remotely sensed data : principles and practices*. CRC press.
- [Congalton and Mead, 1983] Congalton, R. G. and Mead, R. A. (1983). A quantitative method to test for consistency and correctness in photointerpretation. *Photogrammetric Engineering and Remote Sensing*, 49(1) :69–74.
- [Coppin et al., 2004] Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., and Lambin, E. (2004). Review Article Digital change detection methods in ecosystem monitoring : a review. *International Journal of Remote Sensing*, 25(9) :1565–1596.
- [Courty et al., 2014] Courty, N., Flamary, R., and Tuia, D. (2014). Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer.
- [Courty et al., 2017] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9) :1853–1865.
- [Dai et al., 2008] Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2008). Self-taught clustering. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, page nil.
- [Dantzig et al., 1955] Dantzig, G. B., Orden, A., Wolfe, P., et al. (1955). The generalized simplex method for minimizing a linear form under linear inequality restraints. *Pacific Journal of Mathematics*, 5(2) :183–195.
- [de Carvalho et al., 2013] de Carvalho, O., Guimarães, R., Silva, N., Gillespie, A., Gomes, R., Silva, C., and de Carvalho, A. (2013). Radiometric normalization of temporal images combining automatic detection of pseudo-invariant features from the distance and similarity spectral measures, density scatterplot analysis, and robust regression. *Remote Sensing*, 5(6) :2763–2794.
- [Demir et al., 2010] Demir, B., Persello, C., and Bruzzone, L. (2010). Batch-mode active-learning methods for the interactive classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(3) :1014–1031.
- [Derksen et al., 2019] Derksen, D., Inglada, J., and Michel, J. (2019). A metric for evaluating the geometric quality of land cover maps generated with contextual features from high-dimensional satellite image time series without dense reference data. *Remote Sensing*, 11(16) :1929.

- [Dewan and Yamaguchi, 2009] Dewan, A. M. and Yamaguchi, Y. (2009). Land use and land cover change in greater dhaka, bangladesh : Using remote sensing to promote sustainable urbanization. *Applied Geography*, 29(3) :390–401.
- [Di and Crawford, 2011] Di, W. and Crawford, M. M. (2011). View generation for multiview maximum disagreement based active learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5) :1942–1954.
- [Dogliotti et al., 2003] Dogliotti, S., Rossing, W., and Van Ittersum, M. (2003). Rotat, a tool for systematically generating crop rotations. *European Journal of Agronomy*, 19(2) :239–250.
- [Dutrieux et al., 2015] Dutrieux, L. P., Verbesselt, J., Kooistra, L., and Herold, M. (2015). Monitoring forest cover loss using multiple data streams, a case study of a tropical dry forest in bolivia. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107(nil) :112–125.
- [Escofier-Cordier, 1969] Escofier-Cordier, B. (1969). L’analyse factorielle des correspondances. *Cahiers du Bureau universitaire de recherche opérationnelle Série Recherche*, 13 :25–59.
- [Fan et al., 2005] Fan, W., Davidson, I., Zadrozny, B., and Philip, S. Y. (2005). An improved categorization of classifier’s sensitivity on sample selection bias. In *ICDM*, volume 5, pages 605–608.
- [Flamary and Courty, 2017] Flamary, R. and Courty, N. (2017). Pot python optimal transport library. <https://github.com/rflamary/POT>.
- [Flamary et al., 2015] Flamary, R., Fauvel, M., Dalla Mura, M., and Valero, S. (2015). Analysis of multitemporal classification techniques for forecasting image time series. *Geoscience and Remote Sensing Letters, IEEE*, 12(5) :953–957.
- [Foody, 1992] Foody, G. M. (1992). On the compensation for chance agreement in image classification accuracy assessment, photogram. *Eng. Remote Sens.*, 58 :1459–1460.
- [Franklin et al., 2015] Franklin, S. E., Ahmed, O. S., Wulder, M. A., White, J. C., Hermosilla, T., and Coops, N. C. (2015). Large area mapping of annual land cover dynamics using multitemporal change detection and classification of landsat time series data. *Canadian Journal of Remote Sensing*, 41(4) :293–314.
- [Girres and Touya, 2010] Girres, J.-F. and Touya, G. (2010). Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14(4) :435–459.
- [Gomez-Chova et al., 2008] Gomez-Chova, L., Camps-Valls, G., Munoz-Mari, J., and Calpe, J. (2008). Semisupervised image classification with laplacian support vector machines. *IEEE Geoscience and Remote Sensing Letters*, 5(3) :336–340.
- [Gonzalez et al., 2015] Gonzalez, D. M., Camps-Valls, G., and Tuia, D. (2015). Weakly supervised alignment of multisensor images. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, page nil.
- [Hagolle et al., 2010] Hagolle, O., Huc, M., Pascual, D. V., and Dedieu, G. (2010). A multi-temporal method for cloud detection, applied to formosat-2, venps, landsat and sentinel-2 images. *Remote Sensing of Environment*, 114(8) :1747–1755.

- [Hansen et al., 2013] Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160) :850–853.
- [Hartigan and Wong, 1979] Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136 : A k-means clustering algorithm. *Applied Statistics*, 28(1) :100.
- [Hashem, 1997] Hashem, S. (1997). Optimal linear combinations of neural networks. *Neural networks*, 10(4) :599–614.
- [Hemissi et al., 2013] Hemissi, S., Farah, I. R., Ettabaa, K. S., and Solaiman, B. (2013). Multi-spectro-temporal analysis of hyperspectral imagery based on 3-d spectral modeling and multilinear algebra. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1) :199–216.
- [Hilker et al., 2009] Hilker, T., Wulder, M. A., Coops, N. C., Linke, J., McDermid, G., Masek, J. G., Gao, F., and White, J. C. (2009). A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on landsat and modis. *Remote Sensing of Environment*, 113(8) :1613–1627.
- [Ho et al., 1994] Ho, T. K., Hull, J. J., and Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1) :66–75.
- [Hotelling, 1936] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3-4) :321–377.
- [Huang et al., 2007] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608.
- [Igel et al., 2008] Igel, C., Heidrich-Meisner, V., and Glasmachers, T. (2008). Shark. *Journal of Machine Learning Research*, 9 :993–996.
- [Inamdar et al., 2008] Inamdar, S., Bovolo, F., Bruzzone, L., and Chaudhuri, S. (2008). Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images. *Geoscience and Remote Sensing, IEEE Transactions on*, 46(4) :1243–1252.
- [Inglada et al., 2015] Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., and Koetz, B. (2015). Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. *Remote Sensing*, 7(9) :12356.
- [Inglada and Christophe, 2009] Inglada, J. and Christophe, E. (2009). The orfeo toolbox remote sensing image processing software. In *2009 IEEE International Geoscience and Remote Sensing Symposium*, volume 4, pages IV–733. IEEE.
- [Inglada et al., 2017] Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., and Rodes, I. (2017). Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1) :95.

- [Inoue and Narihisa, 2002] Inoue, H. and Narihisa, H. (2002). Optimizing a multiple classifier system. In *Pacific Rim International Conference on Artificial Intelligence*, pages 285–294. Springer.
- [Izquierdo-Verdiguier et al., 2013] Izquierdo-Verdiguier, E., Laparra, V., Gomez-Chova, L., and Camps-Valls, G. (2013). Encoding invariances in remote sensing image classification with svm. *IEEE Geoscience and Remote Sensing Letters*, 10(5) :981–985.
- [Jebara, 2004] Jebara, T. (2004). Multi-task feature and kernel selection for svms. In *Proceedings of the twenty-first international conference on Machine learning*, page 55. ACM.
- [Jouy and Wissocq, 2011] Jouy, L. and Wissocq, A. (2011). Observatoire des pratiques - 34 types de successions culturelles en france. Accessed : 2019-08-30.
- [Jun and Ghosh, 2008] Jun, G. and Ghosh, J. (2008). An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis. In *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, page nil.
- [Jung et al., 2006] Jung, M., Henkel, K., Herold, M., and Churkina, G. (2006). Exploiting synergies of global land cover products for carbon cycle modeling. *Remote Sensing of Environment*, 101(4) :534–553.
- [Kosztra et al., 2019] Kosztra, B., Buttner, G., Hazeu, G., and Arnold, S. (2019). Updated clc illustrated nomenclature guidelines. Accessed : 2019-08-30.
- [Lam and Suen, 1997] Lam, L. and Suen, S. (1997). Application of majority voting to pattern recognition : an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A : Systems and Humans*, 27(5) :553–568.
- [Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- [Le Ber et al., 2006a] Le Ber, F., Benoît, M., Schott, C., Mari, J.-F., and Mignolet, C. (2006a). Studying crop sequences with carrotage, a hmm-based data mining software. *Ecological modelling*, 191(1) :170–185.
- [Le Ber et al., 2006b] Le Ber, F., Lieber, J., and Napoli, A. (2006b). Les systèmes à base de connaissances.
- [Li and Sethi, 2006] Li, M. and Sethi, I. K. (2006). Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(8) :1251–1261.
- [Li et al., 2015] Li, X., Gong, P., and Liang, L. (2015). A 30-year (1984-2013) record of annual urban dynamics of beijing city derived from landsat data. *Remote Sensing of Environment*, 166(nil) :78–90.
- [Liu et al., 2007] Liu, C., Frazier, P., and Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, 107(4) :606–616.
- [Liu et al., 2019] Liu, S., Marinelli, D., Bruzzone, L., and Bovolo, F. (2019). A review of change detection in multitemporal hyperspectral images : Current techniques, applications, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 7(2) :140–158.
- [Maclure and Willett, 1987] Maclure, M. and Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126(2) :161–169.

- [Marcialis and Roli, 2003] Marcialis, G. L. and Roli, F. (2003). Fusion of face recognition algorithms for video-based surveillance systems. In *Multisensor Surveillance Systems*, pages 235–249. Springer.
- [Margono et al., 2012] Margono, B. A., Turubanova, S., Zhuravleva, I., Potapov, P., Tyukavina, A., Baccini, A., Goetz, S., and Hansen, M. C. (2012). Mapping and monitoring deforestation and forest degradation in sumatra (indonesia) using landsat time series data sets from 1990 to 2010. *Environmental Research Letters*, 7(3) :034010.
- [Matasci et al., 2012] Matasci, G., Tuia, D., and Kanevski, M. (2012). Svm-based boosting of active learning strategies for efficient domain adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(5) :1335–1343.
- [Meigs et al., 2015] Meigs, G. W., Kennedy, R. E., Gray, A. N., and Gregory, M. J. (2015). Spatiotemporal dynamics of recent mountain pine beetle and western spruce budworm outbreaks across the pacific northwest region, usa. *Forest Ecology and Management*, 339(nil) :71–86.
- [Mihai et al., 2015] Mihai, B., Nistor, C., and Simion, G. (2015). Post-socialist urban growth of bucharest, romania - a change detection analysis on landsat imagery (1984-2010). *Acta geographica Slovenica*, 55(2) :nil.
- [Mitra et al., 2004] Mitra, P., Shankar, B. U., and Pal, S. K. (2004). Segmentation of multispectral remote sensing images using active support vector machines. *Pattern recognition letters*, 25(9) :1067–1074.
- [Montoya-Zegarra et al., 2013] Montoya-Zegarra, J. A., Leistner, C., and Schindler, K. (2013). Semantic tie points. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, page nil.
- [Navigli and Velardi, 2004] Navigli, R. and Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2) :151–179.
- [Neigh et al., 2014] Neigh, C., Bolton, D., Williams, J., and Diabate, M. (2014). Evaluating an automated approach for monitoring forest disturbances in the pacific northwest from logging, fire and insect outbreaks with landsat time series data. *Forests*, 5(12) :3169–3198.
- [Nielsen, 2007] Nielsen, A. A. (2007). The regularized iteratively reweighted mad method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2) :463–478.
- [Nielsen and Conradsen, 1997] Nielsen, A. A. and Conradsen, K. (1997). Multivariate alteration detection (mad) in multispectral, bi-temporal image data : a new approach to change detection studies. *Dept. Mathematical Modelling, Tech. Univ. Denmark, Lyngby, Denmark, Tech. Rep.*, 11 :1997.
- [Nielsen et al., 1998] Nielsen, A. A., Conradsen, K., and Simpson, J. J. (1998). Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data : New approaches to change detection studies. *Remote Sensing of Environment*, 64(1) :1–19.
- [Olofsson et al., 2014] Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148 :42–57.

- [Opitz and Shavlik, 1996] Opitz, D. W. and Shavlik, J. W. (1996). Generating accurate and diverse members of a neural-network ensemble. In *Advances in neural information processing systems*, pages 535–541.
- [Osman, 2015] Osman, J. (2015). *Connaissances expertes et modélisation pour l’exploitation d’images d’observation de la Terre à hautes résolutions spatiale, spectrale et temporelle*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier.
- [Osman et al., 2012] Osman, J., Inglada, J., Dejoux, J.-F., Hagolle, O., and Dedieu, G. (2012). Fusion of multi-temporal high resolution optical image series and crop rotation information for land-cover map production. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 6785–6788. IEEE.
- [Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10) :1345–1359.
- [Pardo-Pascual et al., 2014] Pardo-Pascual, J. E., Almonacid-Caballer, J., Ruiz, L. A., Palomar-Vázquez, J., and Rodrigo-Aleman, R. (2014). Evaluation of storm impact on sandy beaches of the gulf of valencia using landsat imagery series. *Geomorphology*, 214(nil) :388–401.
- [Paris et al., 2019] Paris, C., Bruzzone, L., and Fernandez-Prieto, D. (2019). A novel approach to the unsupervised update of land-cover maps by classification of time series of multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(7) :4259–4277.
- [Parker et al., 2015] Parker, B. M., Lewis, T., and Srivastava, S. K. (2015). Estimation and evaluation of multi-decadal fire severity patterns using landsat sensors. *Remote Sensing of Environment*, 170(nil) :340–349.
- [Patel et al., 2015] Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation : A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3) :53–69.
- [Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.
- [Pelletier et al., 2017] Pelletier, C., Valero, S., Inglada, J., Champion, N., Sicre, C. M., and Dedieu, G. (2017). Effect of training class label noise on classification performances for land cover mapping with satellite image time series. *Remote Sensing*, 9(2) :173.
- [Persello et al., 2014] Persello, C., Boularias, A., Dalponte, M., Gobakken, T., Naesset, E., and Scholkopf, B. (2014). Cost-sensitive active learning with lookahead : Optimizing field surveys for remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(10) :6652–6664.
- [Petitjean et al., 2011] Petitjean, F., Ketterlin, A., and Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3) :678–693.
- [Peyré, 2019a] Peyré, G. (2019a). Le transport optimal numérique et ses applications - partie 1. Accessed : 2019-08-30.

- [Peyré, 2019b] Peyré, G. (2019b). Le transport optimal numérique et ses applications - partie 2. Accessed : 2019-08-30.
- [Polikar, 2012] Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning*, pages 1–34. Springer.
- [Pontius Jr and Millones, 2011] Pontius Jr, R. G. and Millones, M. (2011). Death to kappa : birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15) :4407–4429.
- [Postadjian et al., 2018] Postadjian, T., Bris, A. L., Sahbi, H., and Mallet, C. (2018). Domain adaptation for large scale classification of very high resolution satellite images with deep convolutional neural networks. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, page nil.
- [Powers, 2012] Powers, D. M. W. (2012). The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 345–355, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Quinlan, 1993] Quinlan, J. (1993). C4. 5 : Programs for machine learning. morgan kaufmann, san francisco. *C4. 5 : Programs for machine learning. Morgan Kaufmann, San Francisco.*
- [Quionero-Candela et al., 2009] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.
- [Rachev, 1985] Rachev, S. T. (1985). The monge-kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4) :647–676.
- [Raina et al., 2007] Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self-taught learning : transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM.
- [Rajan et al., 2006] Rajan, S., Ghosh, J., and Crawford, M. (2006). Exploiting class hierarchies for knowledge transfer in hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11) :3408–3417.
- [Rajan et al., 2008] Rajan, S., Ghosh, J., and Crawford, M. (2008). An active learning approach to hyperspectral data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4) :1231–1242.
- [Rakotomamonjy et al., 2015] Rakotomamonjy, A., Flamary, R., and Courty, N. (2015). Generalized conditional gradient : analysis of convergence and applications. *arXiv preprint arXiv :1510.06567*.
- [Ramadan et al., 2004] Ramadan, E., zhi Feng, X., and Cheng, Z. (2004). Satellite remote sensing for urban growth assessment in shaoxing city, zhejiang province. *Journal of Zhejiang University-SCIENCE A*, 5(9) :1095–1101.
- [Rastrigin and Erenstein, 1981] Rastrigin, L. and Erenstein, R. (1981). Method of collective recognition. *Energoizdat, Moscow*, 595.
- [Redko et al., 2018] Redko, I., Courty, N., Flamary, R., and Tuia, D. (2018). Optimal transport for multi-source domain adaptation under target shift. *arXiv preprint arXiv :1803.04899*.

- [Reiter and Dale, 2000] Reiter, E. and Dale, R. (2000). *Building natural language generation systems*. Cambridge university press.
- [Richardson and Domingos, 2006] Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine learning*, 62(1-2) :107–136.
- [Rückert and Kramer, 2008] Rückert, U. and Kramer, S. (2008). Kernel-based inductive transfer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 220–233. Springer.
- [Ruta and Gabrys, 2005] Ruta, D. and Gabrys, B. (2005). Classifier selection for majority voting. *Information fusion*, 6(1) :63–81.
- [Schönhart et al., 2011] Schönhart, M., Schmid, E., and Schneider, U. A. (2011). Croprota—a crop rotation model to support integrated land use assessments. *European Journal of Agronomy*, 34(4) :263–277.
- [Senf et al., 2015] Senf, C., Pflugmacher, D., Wulder, M. A., and Hostert, P. (2015). Characterizing spectral-temporal patterns of defoliator and bark beetle disturbances using landsat time series. *Remote Sensing of Environment*, 170(nil) :166–177.
- [Shlien, 1990] Shlien, S. (1990). Multiple binary decision tree classifiers. *Pattern Recognition*, 23(7) :757–763.
- [Smith and Shafer, 1976] Smith, A. F. M. and Shafer, G. (1976). A mathematical theory of evidence. *Biometrics*, 32(3) :703.
- [Stehman, 1997] Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1) :77–89.
- [Sterling et al., 2012] Sterling, S. M., Ducharne, A., and Polcher, J. (2012). The impact of global land-cover change on the terrestrial water cycle. *Nature Climate Change*, 3(4) :385–390.
- [Stumpf et al., 2012] Stumpf, A., Lachiche, N., Kerle, N., Malet, J.-P., and Puissant, A. (2012). Adaptive spatial sampling with active random forest for object-oriented landslide mapping. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, page nil.
- [Sun et al., 2013] Sun, Z., Wang, C., Wang, H., and Li, J. (2013). Learn multiple-kernel svms for domain adaptation in hyperspectral data. *IEEE Geoscience and Remote Sensing Letters*, 10(5) :1224–1228.
- [Tardy et al., 2017] Tardy, B., Inglada, J., and Michel, J. (2017). Fusion approaches for land cover map production using high resolution image time series without reference data of the corresponding period. *Remote Sensing*, 9(11) :1151.
- [Tardy et al., 2018] Tardy, B., Inglada, J., and Michel, J. (2018). Class selection methods for land cover mapping without reference data of the corresponding period. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6203–6206.
- [Tardy et al., 2019] Tardy, B., Inglada, J., and Michel, J. (2019). Assessment of optimal transport for operational land-cover mapping using high-resolution satellite images time series without reference data of the mapping period. *Remote Sensing*, 11(9) :1047.

- [Tuia and Camps-Valls, 2016] Tuia, D. and Camps-Valls, G. (2016). Kernel manifold alignment for domain adaptation. *PLOS ONE*, 11(2) :e0148655.
- [Tuia et al., 2015] Tuia, D., Flamary, R., Rakotomamonjy, A., and Courty, N. (2015). Multitemporal classification without new labels : a solution with optimal transport. In *Multitemp.*
- [Tuia et al., 2016] Tuia, D., Persello, C., and Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data : An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2) :41–57.
- [Tuia et al., 2009] Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F., and Emery, W. J. (2009). Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7) :2218–2232.
- [Tuia et al., 2014] Tuia, D., Volpi, M., Trolliet, M., and Camps-Valls, G. (2014). Semisupervised manifold alignment of multimodal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(12) :7708–7720.
- [Tumer and Ghosh, 1996] Tumer, K. and Ghosh, J. (1996). Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition*, 29(2) :341–348.
- [Uschold and Gruninger, 1996] Uschold, M. and Gruninger, M. (1996). Ontologies : Principles, methods and applications. *The Knowledge Engineering Review*, 11(2) :93–136.
- [Van Erp et al., 2002] Van Erp, M., Vuurpijl, L., and Schomaker, L. (2002). An overview and comparison of voting methods for pattern recognition. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 195–200. IEEE.
- [Vancutsem et al., 2012] Vancutsem, C., Marinho, E., Kayitakire, F., See, L., and Fritz, S. (2012). Harmonizing and combining existing land cover/land use datasets for cropland area monitoring at the african continental scale. *Remote Sensing*, 5(1) :19–41.
- [Villani, 2008] Villani, C. (2008). *Optimal transport : old and new*, volume 338. Springer Science & Business Media.
- [Volpi et al., 2015] Volpi, M., Camps-Valls, G., and Tuia, D. (2015). Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107(nil) :50–63.
- [Wang et al., 2008] Wang, Z., Song, Y., and Zhang, C. (2008). Transferred dimensionality reduction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 550–565. Springer.
- [Wilson and Sader, 2002] Wilson, E. H. and Sader, S. A. (2002). Detection of forest harvest type using multiple dates of landsat tm imagery. *Remote Sensing of Environment*, 80(3) :385–396.
- [Wolpert, 2002] Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In *Soft computing and industry*, pages 25–42. Springer.

- [Wozniak and Jackowski, 2009] Wozniak, M. and Jackowski, K. (2009). Some remarks on chosen methods of classifier fusion based on weighted voting. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 541–548. Springer.
- [Xu et al., 1992] Xu, L., Krzyzak, A., and Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics*, 22(3) :418–435.
- [Yang and Crawford, 2016] Yang, H. L. and Crawford, M. M. (2016). Domain adaptation with preservation of manifold geometry for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(2) :543–555.
- [Zanotta et al., 2015] Zanotta, D. C., Bruzzone, L., Bovolo, F., and Shimabukuro, Y. E. (2015). An adaptive semisupervised approach to the detection of user-defined recurrent changes in image time series. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7) :3707–3719.
- [Zhu, 2017] Zhu, Z. (2017). Change detection using landsat time series : a review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130(nil) :370–384.
- [Zielstra and Zipf, 2010] Zielstra, D. and Zipf, A. (2010). A comparative study of proprietary geodata and volunteered geographic information for germany. In *13th AGILE international conference on geographic information science*, volume 2010.

AA Apprentissage Automatique. [18](#), [21](#), [34](#), [43](#), [58](#), [134](#)

ACC Analyse Canonique des Corrélations. [63](#), [67](#), [69](#), [76](#), [78](#), [160–162](#), [171](#), [182](#), [185](#)

ACP Analyse en Composantes Principales. [63](#), [67](#), [69](#), [72](#), [81](#)

bagging Bootstrap **aggregating**. [34](#), [38](#), [135](#)

CES OSO Centre d’Expertise Scientifique Occupation des SOs. [53](#), [54](#)

CESBIO Centre d’Études Spatiales de la BIOSphère. [21](#), [22](#), [27](#), [28](#), [30](#), [106](#), [166](#), [169](#)

CLC Corine Land Cover. [16](#), [66](#), [126](#)

CNES Centre National d’Études Spatiales. [170](#)

DTW *Dynamic Time Warping* (ou déformation temporelle dynamique en français). [64](#), [116](#)

EMD *Earth Mover’s Distance*. [95](#), [96](#), [105](#), [137](#)

FScore F_1 Score. [43](#), [47](#), [74](#), [102](#), [125](#), [129](#), [141](#), [154](#), [172](#), [186](#), [188](#)

GIEC Groupe d’experts intergouvernemental sur l’évolution du climat. [14](#)

HPC Calcul à Haute Performance (*High Performance Computer*). [72](#), [170](#)

IGN Institut Géographique National. [21](#)

iota² Infrastructure pour l’Occupation des sols par Traitement Automatique Incorporant les Orfeo Toolbox Applications. [53](#), [166–170](#), [172](#), [180](#)

MAB Programme sur l'homme et la biosphère ou *Man and Biosphere* en anglais. 14

MACCS Multi-sensor Atmospheric Correction and Cloud Screening. 30, 40, 46

MAD Multivariate Alteration Detector. 65, 80, 114

MIR Moyen InfraRouge. 32, 206

MSS *MultiSpectral Scanner*. 15

MT Matrice de Transition. 122, 139, 171, 186–188

NDVI *Normalised Difference Vegetation Index*. 32, 128

NDWI *Normalised Difference Water Index*. 32

OCS OCcupation des Sols. 15, 27, 58, 72, 118, 132, 152, 166, 167, 171, 176, 184–188

OOB Out OfBag. 38

OSM *OpenStreetMap*. 18

OSR Observatoire Spatial Régional. 27

OTB Orfeo ToolBox. 40

PAC Politique Agricole Commune. 21

PG Précision Globale (*Overall Accuracy* en anglais). 42, 45, 66, 73, 101, 124, 129, 138, 147, 152, 172, 173, 176

PIR Proche InfraRouge. 30, 32, 205, 206

POT *Python Optimal Transport*. 93

RBV *Return Beam Vidicon*. 15

RF *Random Forests* ou forêts aléatoires en français. 27, 34, 38, 39, 41, 45, 46, 49, 50, 55, 62, 66, 67, 99, 100, 111–114, 129–131, 134, 135, 138, 161, 182

RPG Registre Parcellaire Graphique. 21, 121, 122, 168

RSI *Remote Sensing Instrument*. 30

SAP Séparation des classes Annuelles et Pérennes. 53, 100, 109, 111–113, 125, 137, 142, 144, 154, 161, 162, 188

SEH Surfaces En Herbe. 29, 47, 49, 76, 78

SIG Système d'Information Géographique. 18

SPOT Satellite pour l'observation de la Terre. 17

SVM Séparateurs à Vaste Marge. 61, 62, 128, 161

TA Transfert d'Apprentissage (*Transfer Learning*). 58

TO Transport Optimal. 88, 125, 137, 152, 171, 182, 185, 186

UTM *Universal Transverse Mercator*. 15, 166

VC Vote par Confiance. 138, 153, 172, 173, 176, 181, 182, 186

VM Vote Majoritaire. 134–136, 138, 154, 172, 173, 176, 181, 182, 186, 187

VP Vote par Probabilités. 138, 153, 172, 186

ANNEXE A

Bandes spectrales

TABLE A.1 : Bandes spectrales du Return Beam Vidicon (RBV)

Bande spectrale	Longueur d'onde (nm)	Résolution (m)
Bleu-Vert	475-575	80
Orange-Rouge	580-680	80
PIR	690-830	80

TABLE A.2 : Bandes spectrales du Multispectral Scanner (MSS)

Bande spectrale	Longueur d'onde (μm)	Résolution (m)
Vert	0.5-0.6	80
Rouge	0.6-0.7	80
PIR	0.7-0.8	80
PIR	0.8-1.1	80

TABLE A.3 : Bandes spectrales du capteur optique Operational Land Imager (OLI)

Bande spectrale	Longueur d'onde (μm)	Résolution (m)
Aérosol	0.43-0.43	30
Bleu	0.45-0.51	30
Vert	0.53-0.59	30
Rouge	0.64-0.67	30
PIR	0.85-0.88	30
MIR 1	1.57-1.65	30
MIR 2	2.11-2.29	30
Panchromatic	0.50-0.68	15
Cirrus	1.36-1.38	30

TABLE A.4 : Bandes spectrales Sentinel-2

Numéro de bande	Sentinel-2A		Sentinel-2B		Résolution
	Longueur d'onde centrale (nm)	Largeur de longueur d'onde (nm)	Longueur d'onde centrale (nm)	Largeur de longueur d'onde (nm)	
1	442.7	21	442.2	21	60
2	492.4	66	492.1	66	10
3	559.8	36	559.0	36	10
4	664.6	31	664.9	31	10
5	704.1	15	703.8	16	20
6	740.5	15	739.1	15	20
7	782.2	20	779.7	20	20
8	832.8	106	832.9	106	10
8a	864.7	21	864.0	22	20
9	945.1	20	943.2	21	60
10	1373.5	31	1376.9	30	60
11	1613.7	91	1610.4	94	20
12	2202.4	175	2185.7	185	20

La classification Naïve

B.1 Illustration de la PG naïve

La figure [B.1](#) est une représentation sous forme de matrice de la table [3.1](#). Elle permet de visualiser les variations de performances pour toutes les paires d'années possibles. Les couleurs sur les colonnes restent généralement proches, alors que des variations plus importantes apparaissent sur les lignes. Un effet visuel intéressant à observer est le carré formé par les années Source et Cible 2010, 2011, 2012.

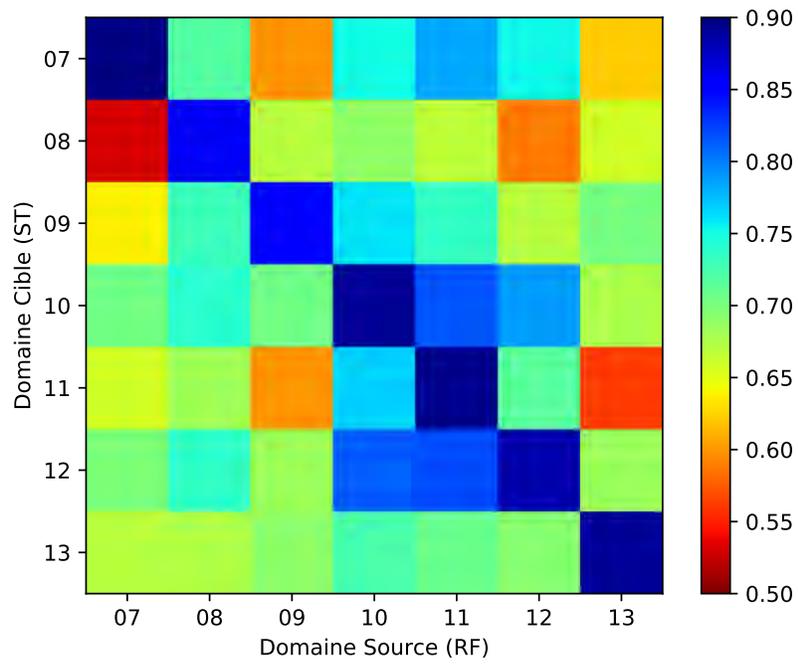


FIGURE B.1 : Matrice représentant les domaines Source (abscisses) et Cible (ordonnées). L'échelle de couleur représente la PG. La diagonale correspond à la classification supervisée.

B.2 Matrice de confusion

Classification naïve du domaine cible 2008 par le domaine Source 2007.

KAPPA : 0.513 PG : 0.567

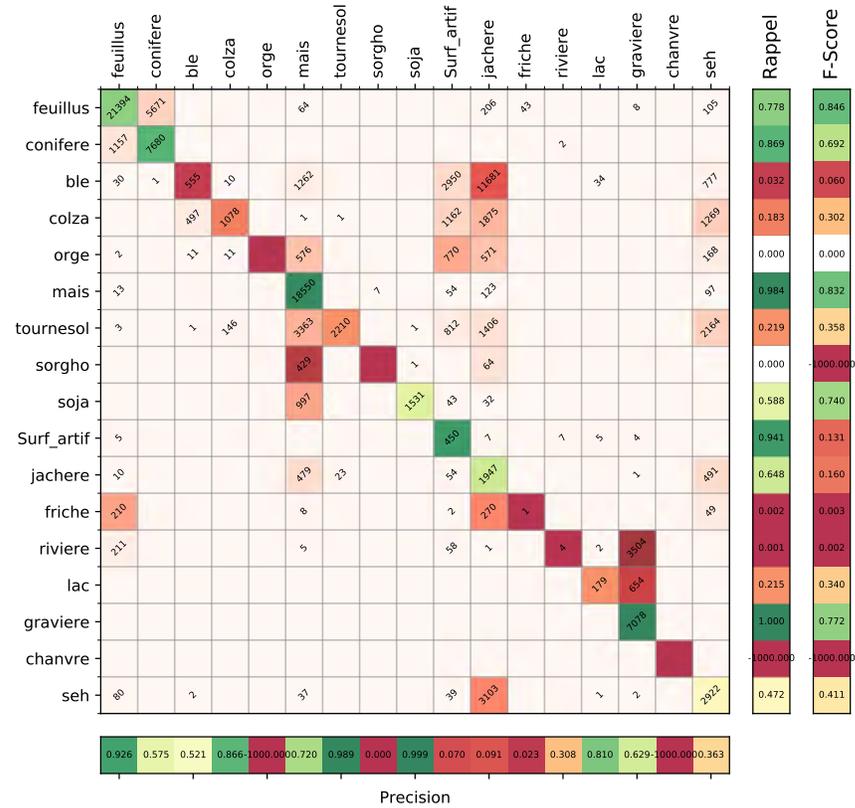


FIGURE B.2 : Matrice de confusion de la classification naïve de 2008 (cible) par 2007 (source).

L'analyse canonique des corrélations

C.1 Résultats

TABLE C.1 : Matrice d'écart entre l'ACC utilisant le masque de pixels altérés et le masque de classe annuelles

Année	07	08	09	10	11	12	13
07		-0.15	0.09	0.04	0.09	0.01	0.13
08	0.02		-0.12	0.13	0.01	0.16	0.06
09	0.05	-0.12		-0.03	0.07	-0.01	0.07
10	0.04	0.04	-0.19		-0.25	0.01	0.04
11	0.02	-0.12	0.09	-0.04		-0.02	0.06
12	0.13	0.02	-0.16	0.05	-0.11		0.06
13	0.15	-0.07	0.13	-0.02	0.12	-0.02	

D.1 Résultats annexes de l'étude de sensibilité

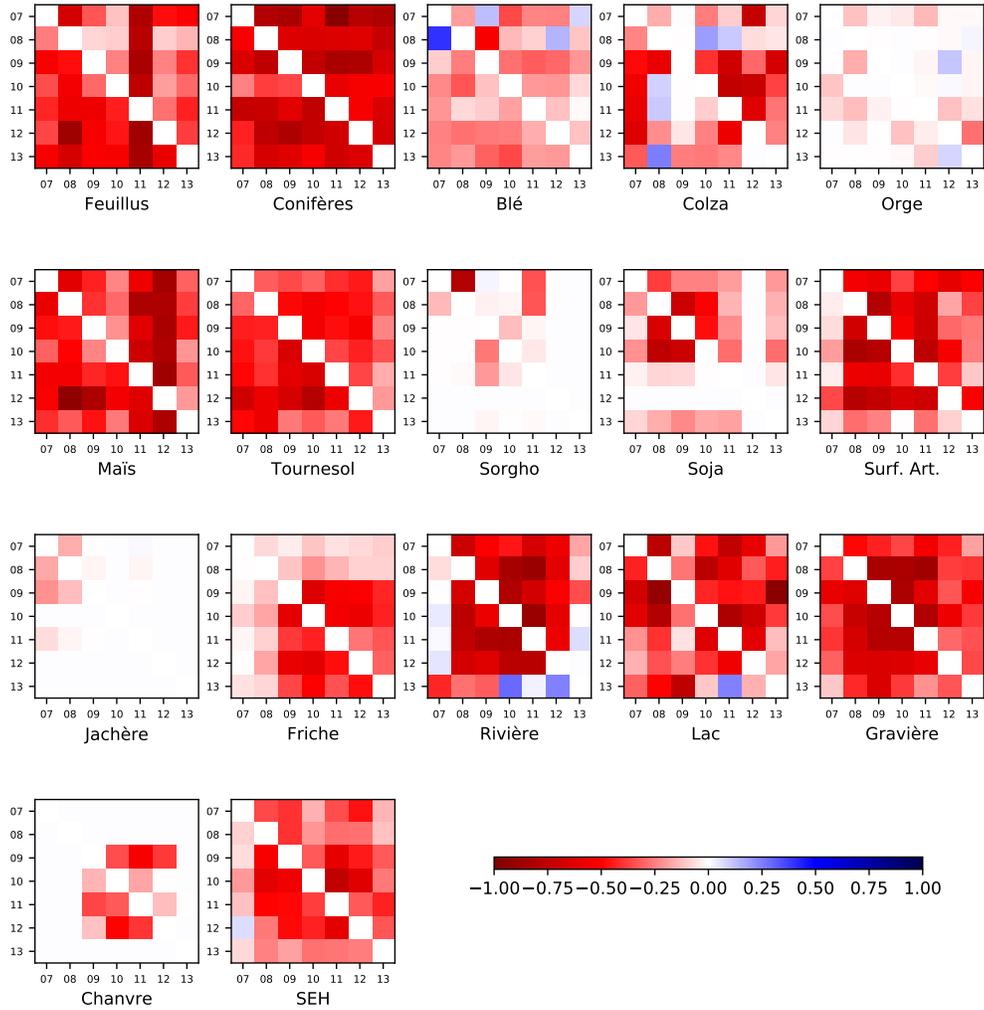


FIGURE D.1 : FScore par classe obtenu par le Sinkhorn avec sélection des échantillons à partir des données de référence des deux domaines.

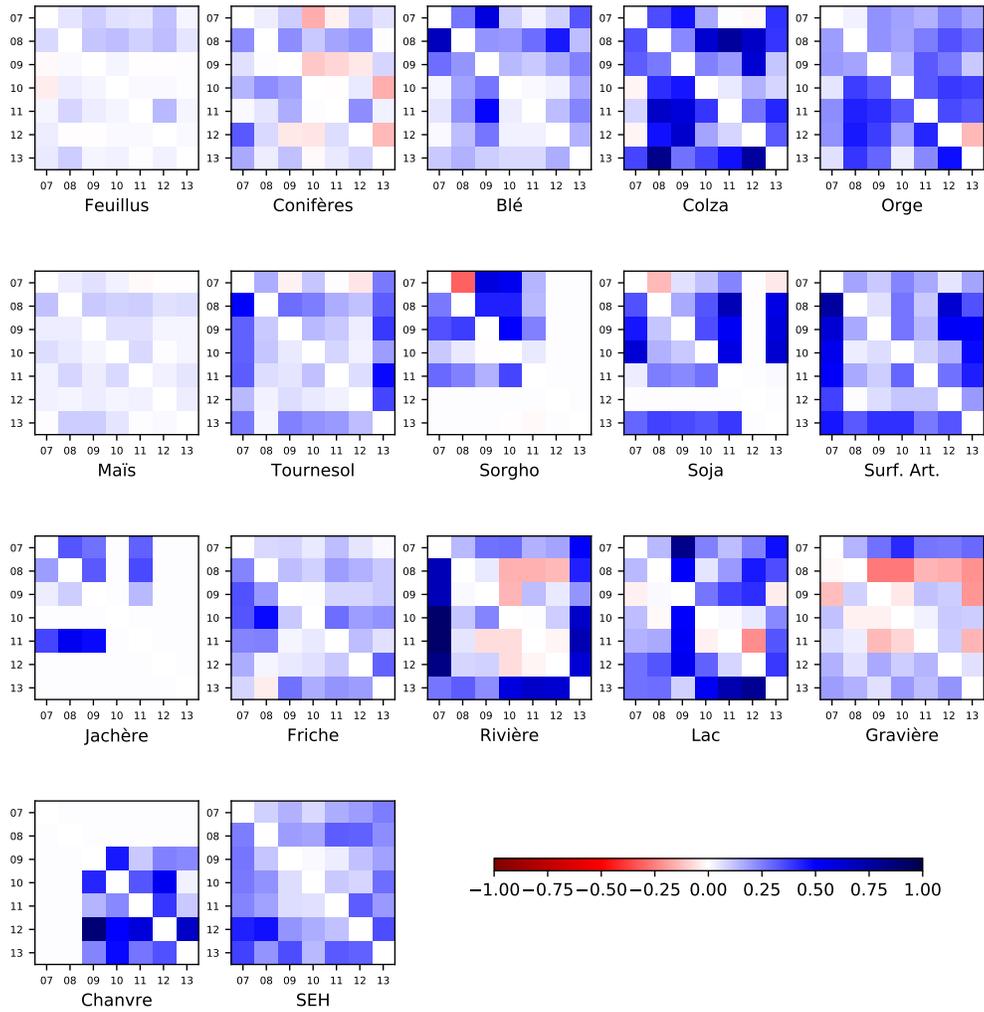


FIGURE D.2 : FScore par classe obtenus par l'EMD avec exploitation des étiquettes des deux domaines.

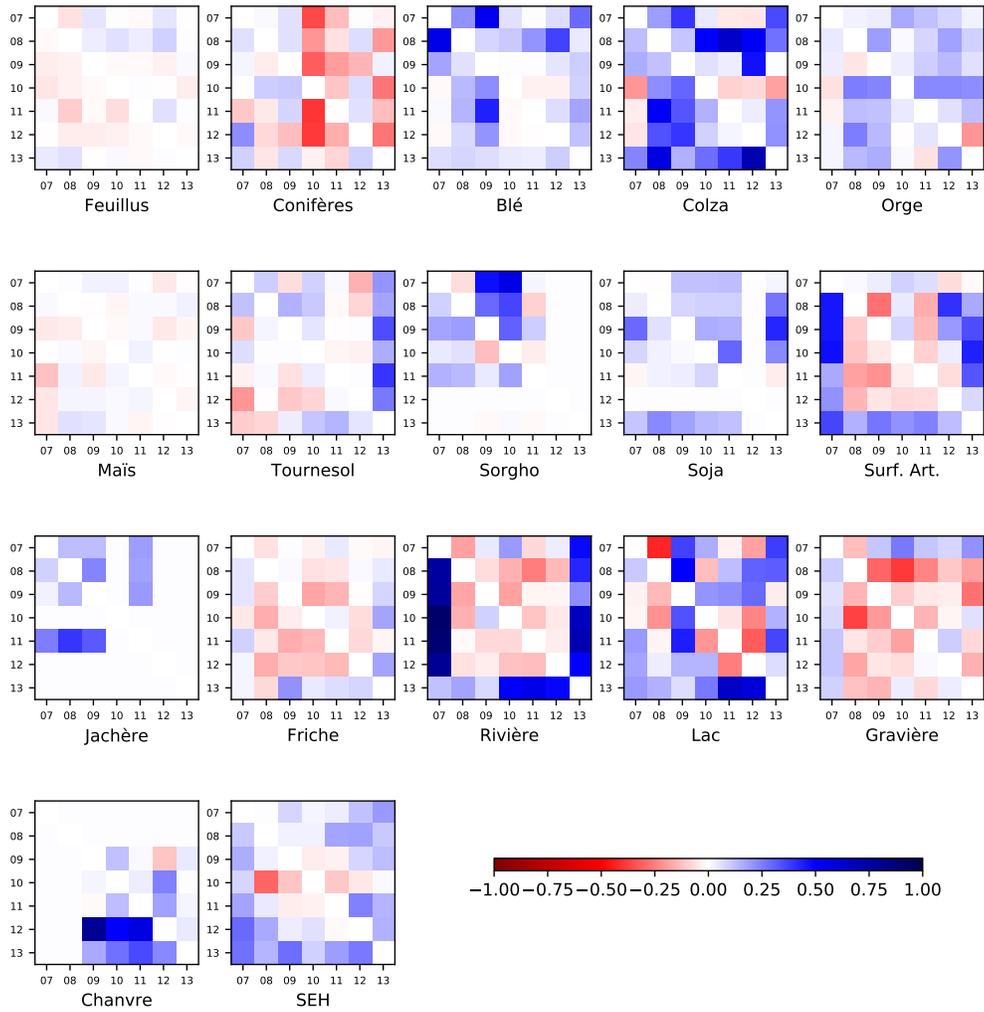


FIGURE D.3 : FScore par classe obtenus par le Sinkhorn avec exploitation des étiquettes des deux domaines.

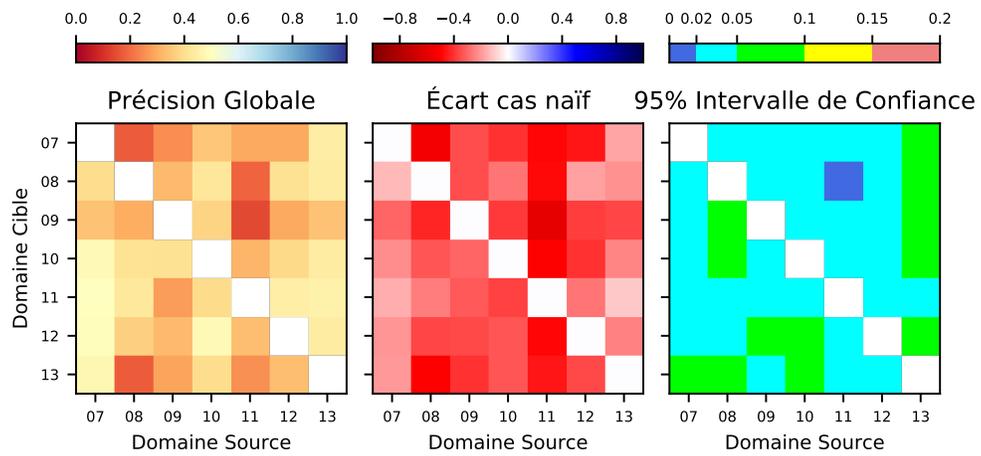


FIGURE D.4 : Matrices de performances pour le Sinkhorn avec la sélection par position.

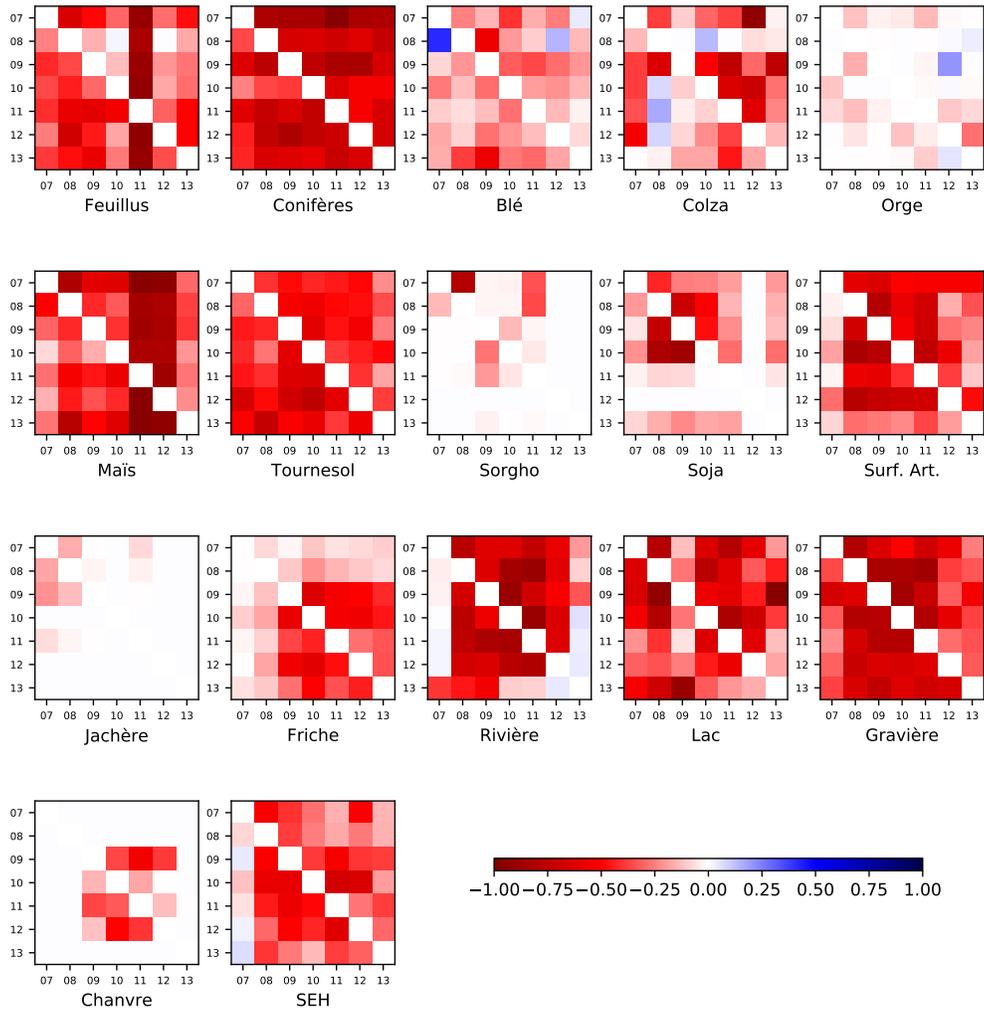


FIGURE D.5 : FScore obtenus par le Sinkhorn avec la sélection par position.

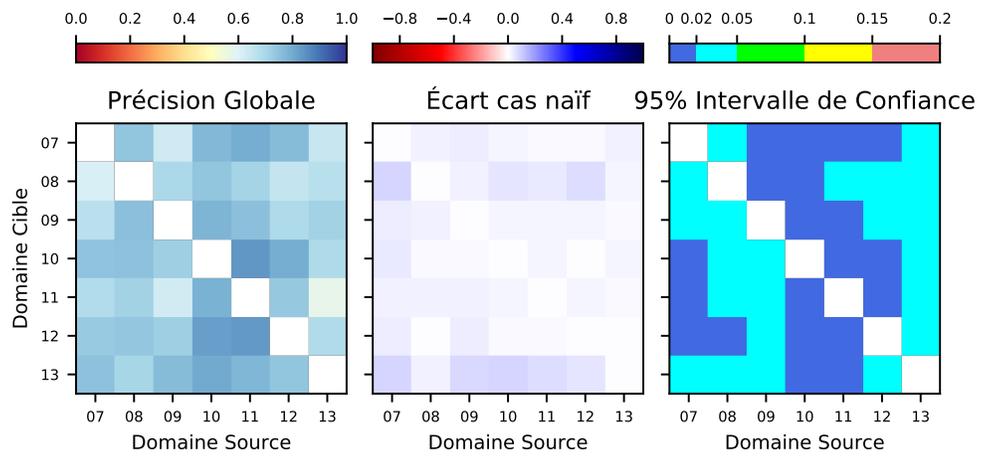


FIGURE D.6 : Matrices de performance pour le TO EMD par classe.

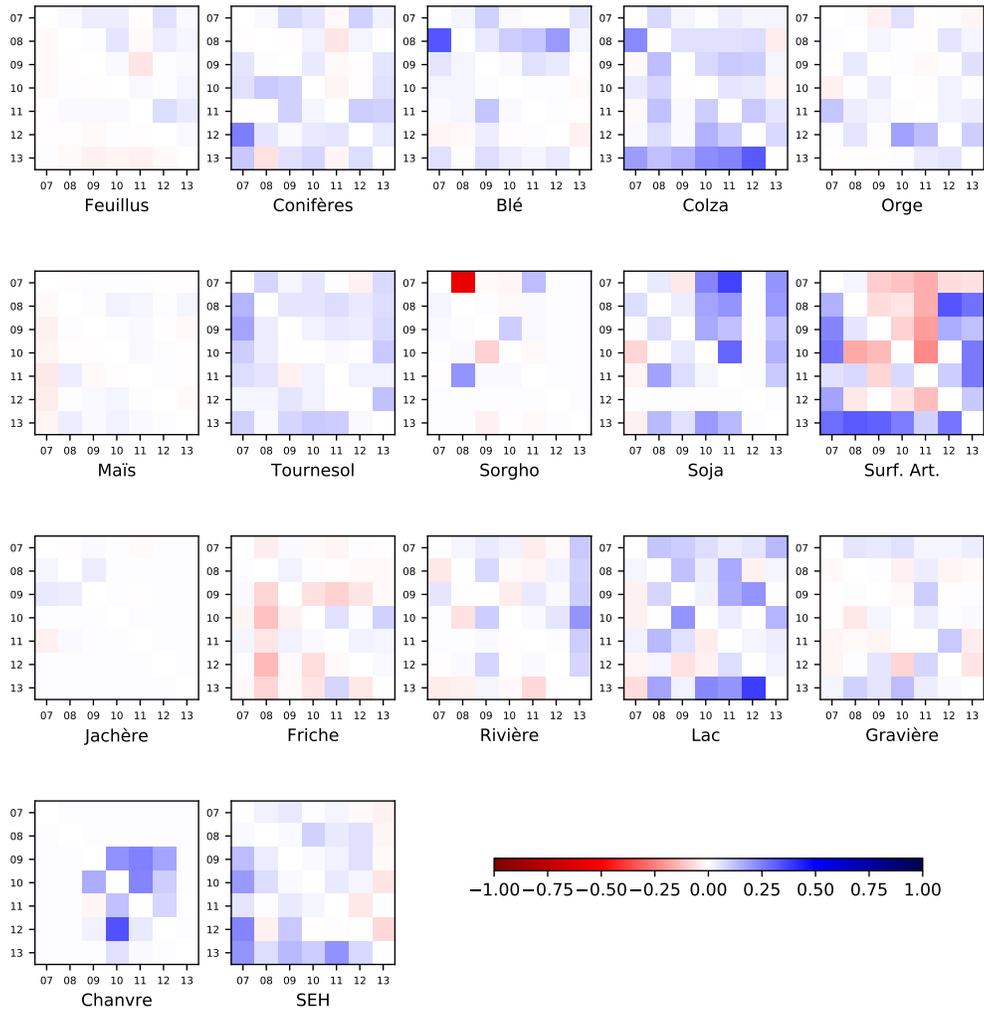


FIGURE D.7 : FScore obtenu par l'EMD par classe. Toutes les classes sont transportées.

ANNEXE E

Synthèse des comparatifs de vote

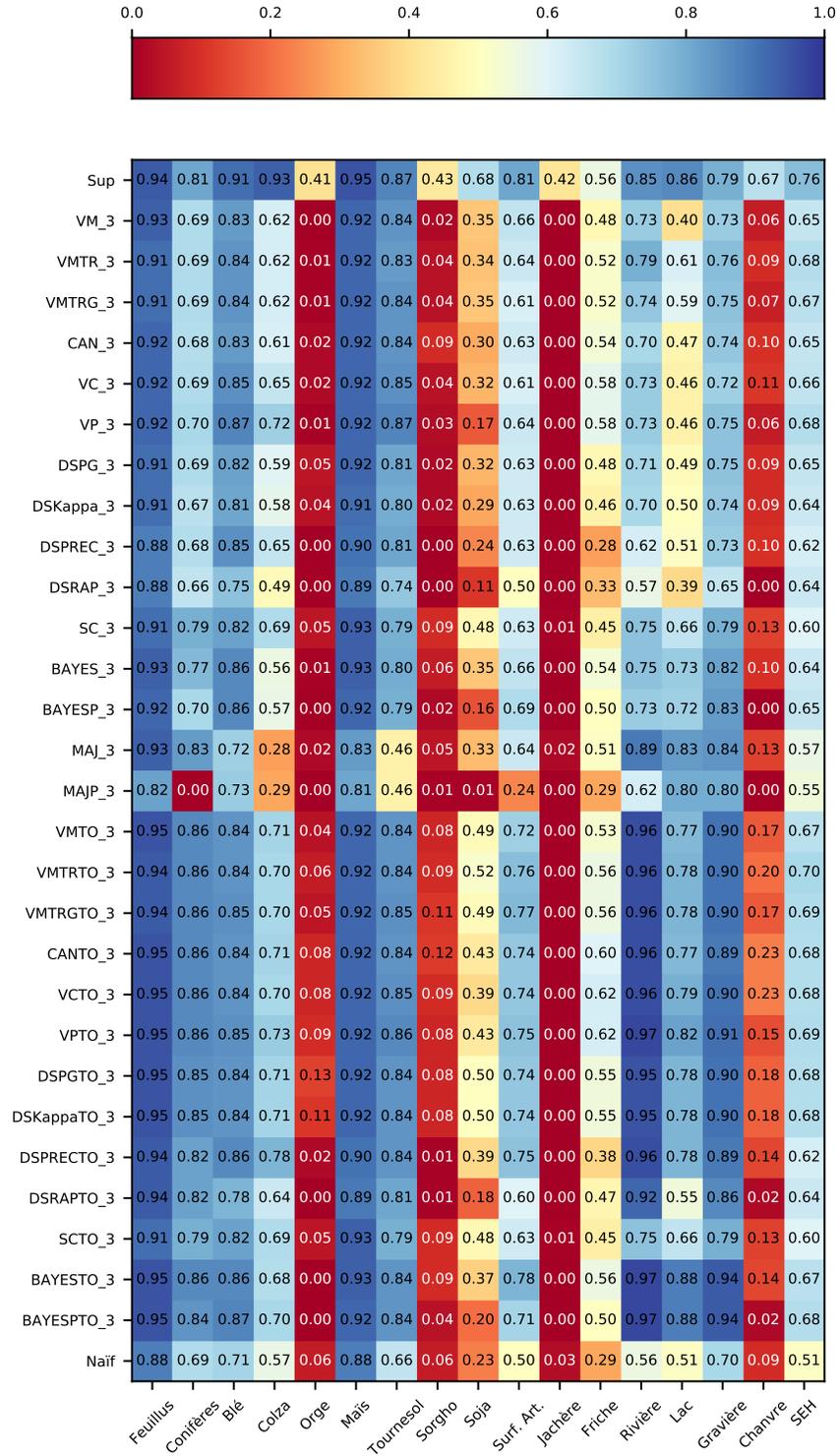


FIGURE E.1 : Comparatif du FScore pour le cas 3.

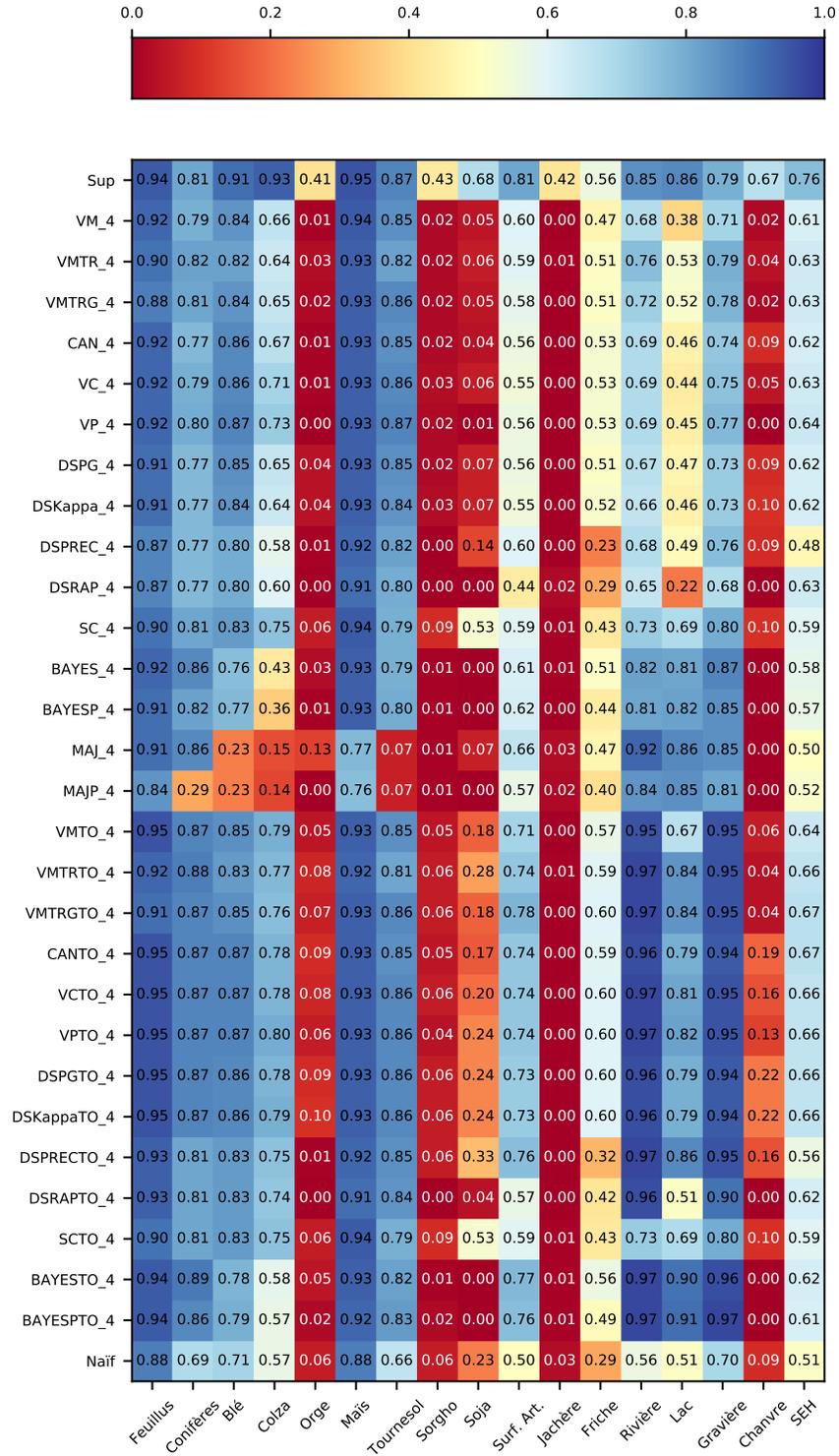


FIGURE E.2 : Comparatif du FScore pour le cas 4.

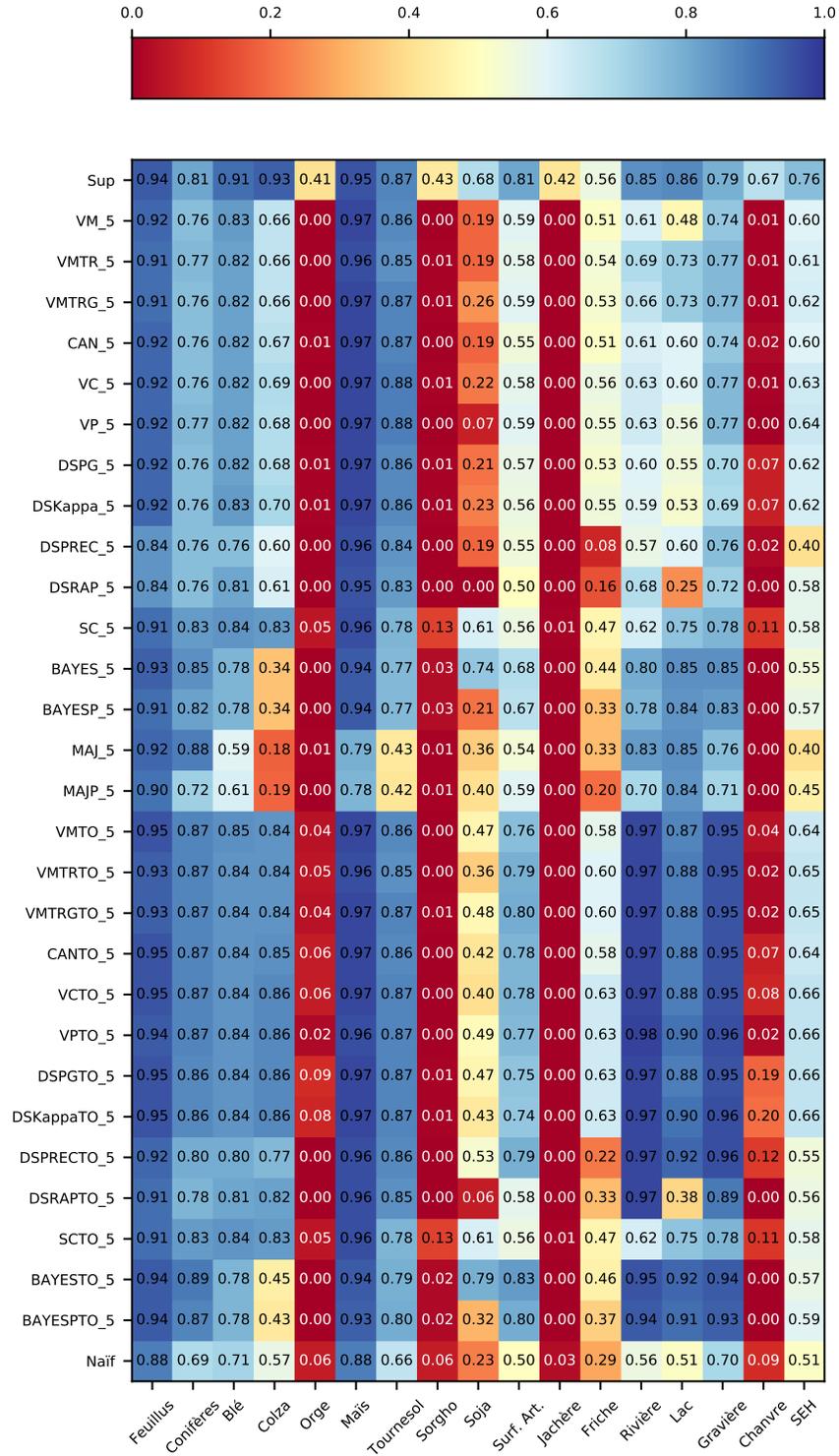


FIGURE E.3 : Comparatif du FScore pour le cas 5.

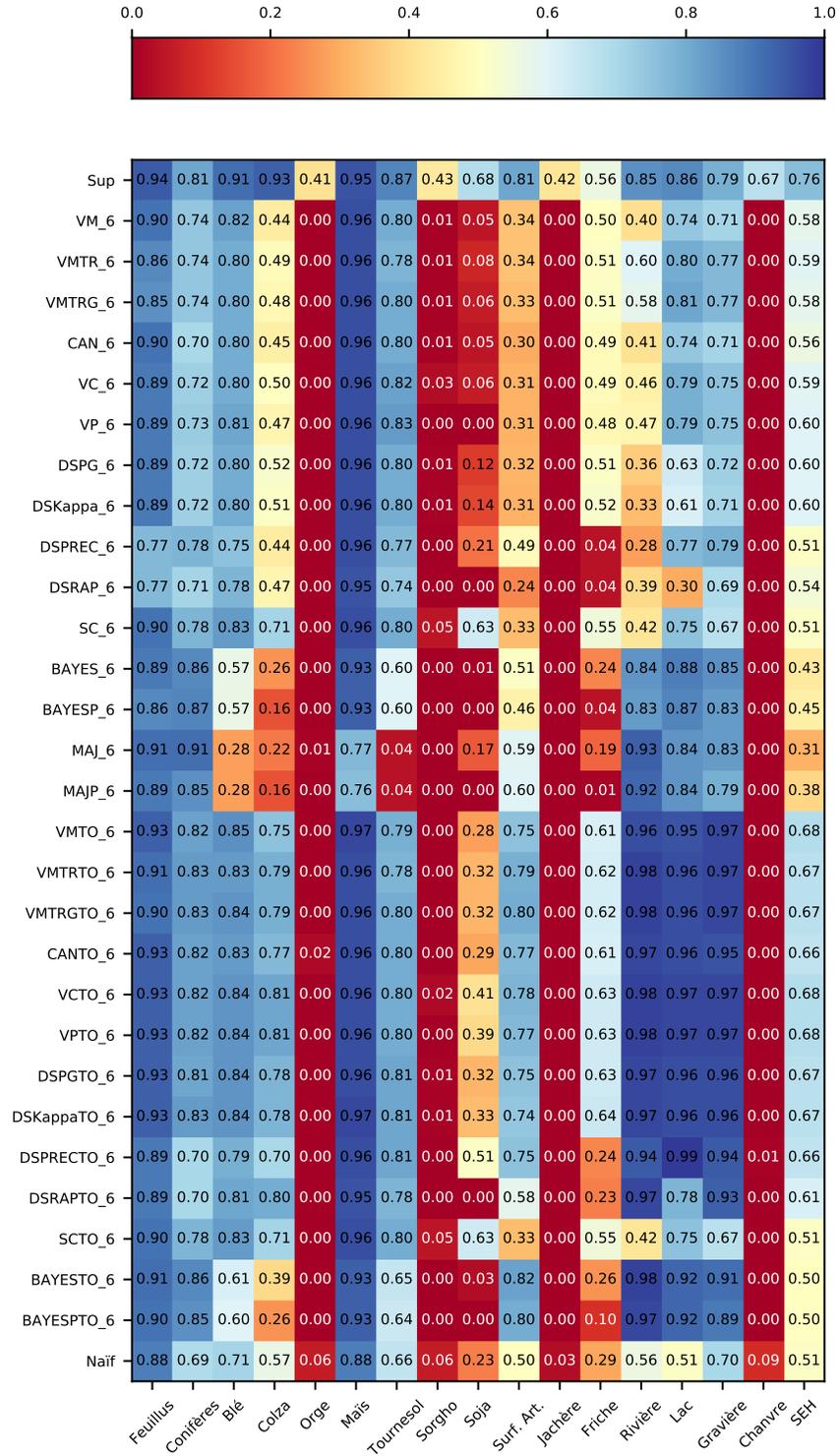


FIGURE E.4 : Comparatif du FScore pour le cas 6.

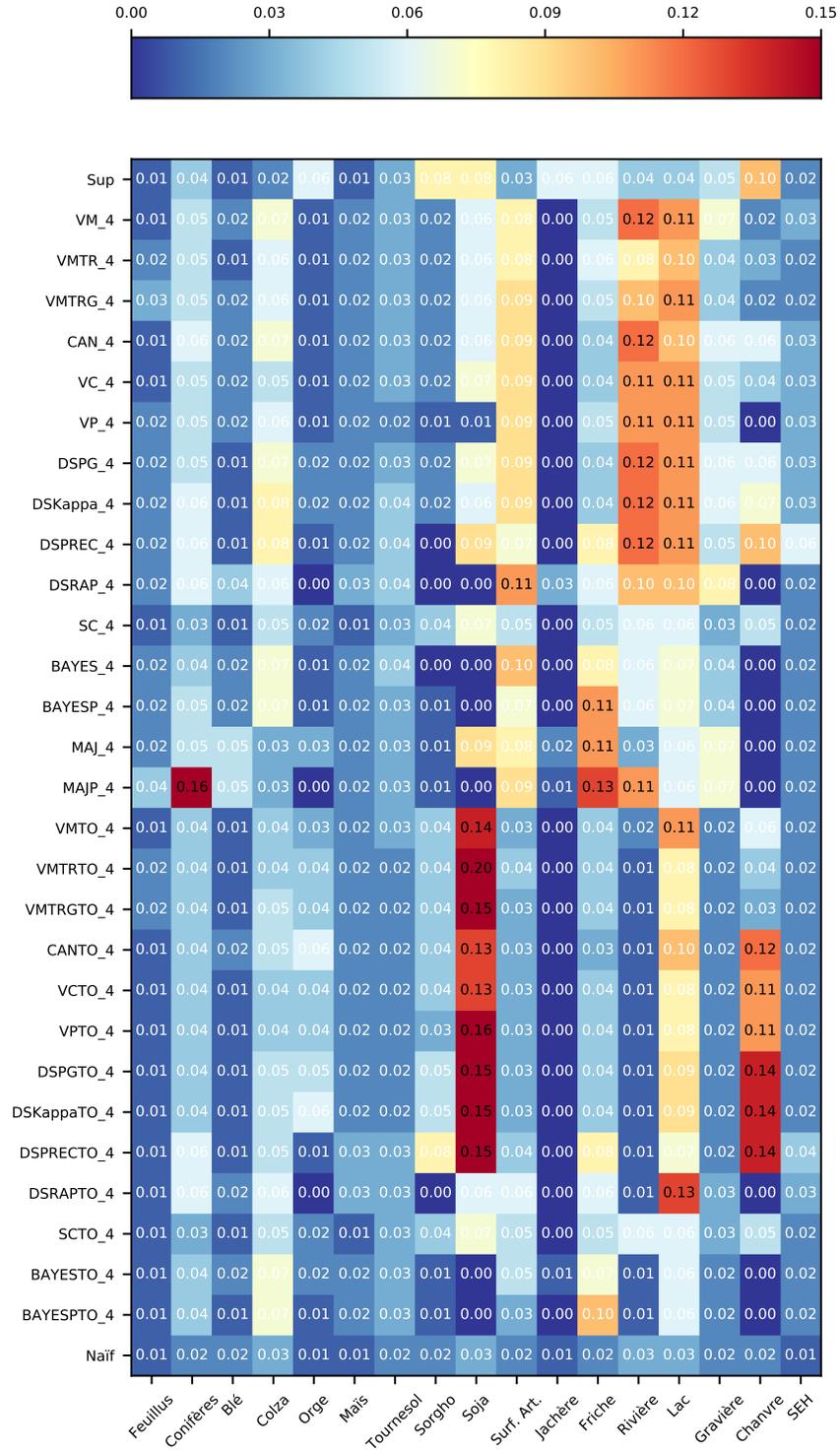


FIGURE E.5 : Intervalles de confiance associés au Fscore pour le cas 4. Les valeurs indiquées représentent la borne maximale de l'intervalle de confiance, il faut donc comprendre $F_{\text{Score}} \pm I_{\text{c}_{\text{max}}}$.

Visualisation des performances iota^2

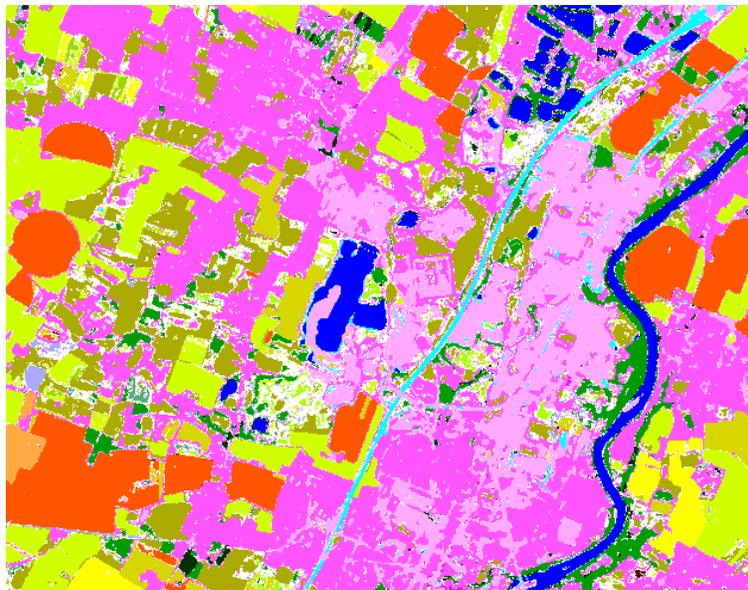


FIGURE F.1 : Carte de différence entre les classifications produites par le classifieur unique et bayésien

TABLE F.1 : Répartition du nombre d'images par tuile et par année

Tuile	Année		
	2016	2017	2018
T30TXN	50	79	101
T30TXP	46	65	74
T30TXQ	62	66	76
T30TXR	48	69	82
T30TYN	54	81	101
T30TYP	37	64	75
T30TYQ	39	61	74
T31TCH	46	76	102
T31TCJ	47	72	93
T31TCK	35	63	67
T31TDH	48	73	98
T31TDJ	43	69	79
T31TDK	39	63	81
T31TEH	53	75	114
T31TEJ	73	106	143
T31TEK	42	67	82
T31TFH	53	77	120
T31TFJ	54	76	115
T31TFK	50	74	104
T31TFL	40	56	95
T31TGH	58	82	115
T31TGJ	77	114	151
T31TGK	55	79	108
T31TGL	48	73	112
T32TLP	64	83	105
T32TLQ	56	85	107
T32TLR	57	86	116
Total	1374	2034	2690

Résumé

La production de cartes d'occupation des sols, à partir d'images de télédétection, requiert l'utilisation de méthodes de classification supervisée. Ces algorithmes dépendent de données de référence pour étiqueter les éléments exploités lors de l'apprentissage. Ces données sont très coûteuses à obtenir et sont associées à de long délais de mise à disposition, retardant la production de la carte correspondante. Dans un premier temps, nous proposons l'utilisation d'approches d'adaptation de domaine permettant d'exploiter un historique de données, afin de contourner la dépendance aux données de référence de la période à cartographier. Dans un second temps, nous proposons l'utilisation d'approches de fusion de classifieurs permettant d'exploiter la diversité de l'historique de données amassée aux cours du temps. Nous concluons ces travaux sur un constat d'égalité de performances des méthodes proposées mais à des coûts d'exploitation variables.

Abstract

The production of land cover maps, based on remote sensing images, requires the use of supervised classification methods to be effective. These algorithms rely on mandatory reference data to label the elements used in learning. These data are very expensive to obtain, and are associated with long lead times, delaying the production of the corresponding map. In this thesis work, we propose the use of domain adaptation approaches to exploit a data history, in order to bypass the dependence on reference data for the period to be mapped. In a second step, we propose the use of classifier fusion approaches to exploit the diversity of the data history collected over time. We conclude this work with a finding of equal performance of the proposed methods but at variable operating costs.