Vers une occupation du sol France entière par imagerie satellite à très haute résolution

Tristan Postadjian

▶ **To cite this version:**

Tristan Postadjian. Vers une occupation du sol France entière par imagerie satellite à très haute résolution. Réseau de neurones [cs.NE]. Université Paris-Est, 2020. Français. ⟨NNT : 2020PESC2018⟩. ⟨tel-03045637⟩

HAL Id: tel-03045637

https://theses.hal.science/tel-03045637v1

Submitted on 8 Dec 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.

HAL Authorization

THÈSE DE DOCTORAT

DE L'ECOLE DOCTORALE MSTIC
-
MATHÉMATIQUES, SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

PRÉSENTÉE POUR L'OBTENTION DU TITRE DE DOCTEUR DE L'

UNIVERSITÉ PARIS-EST

PAR

TRISTAN POSTADJIAN

VERS UNE OCCUPATION DES SOLS FRANCE ENTIÈRE À PARTIR D'IMAGES SATELLITES TRÈS HAUTE RÉSOLUTION

VERSION DÉFINITIVE,
SOUTENUE LE 12 FÉVRIER 2020

JURY

NICOLE VINCENT	RAPPORTEURE
THOMAS CORPETTI	RAPPORTEUR
ANNE PUISSANT	EXAMINATRICE
JORDI INGLADA	EXAMINATEUR
CLÉMENT MALLET	DIRECTEUR
ARNAUD LE BRIS	CO-ENCADRANT
HICHEM SAHBI	CO-ENCADRANT

Résumé

La connaissance de la couverture des territoires en terme d'occupation des sols est devenue un enjeu majeur du XXIème siècle. Que ce soit à l'échelle nationale ou à une échelle plus globale, les initiatives se multiplient pour proposer des cartographies d'occupation des sols qui répondent à des besoins propres à chacune. Consistant à classer des objets présents sur le sol selon des nomenclatures prédéfinies, la tâche est fastidieuse à l'heure actuelle avec des processus essentiellement manuels ou semi-manuels, nécessaires pour garantir le respect de certaines qualités et spécifications.
De son côté, la télédétection spatiale a connu un essor conséquent avec la multiplication des capteurs optiques d'observation de la Terre disponibles et de leur diversité en terme de résolutions spectrale, spatiale et temporelle. Ces capteurs optiques proposent chacun une description de la surface terrestre qui leur est propre, et donc caractérisant un ou plusieurs type(s) d'occupation(s) des sols. Ces types dépendent justement des caractéristiques de ces capteurs, caractéristiques adaptées davantage à l'observation des glaciers, des forêts ou des zones plus urbaines par exemple. Les satellites SPOT 6 et SPOT 7, lancés en 2012 et 2014 respectivement, sont dotés de capteurs optiques à très haute résolution spatiale, et acquièrent des images dans quatre bandes spectrales à haute résolution ainsi qu'une bande panchromatique à très haute résolution, permettant de porter la résolution des quatre canaux spectraux à 1,5 m. L'IGN, à partir de ces acquisitions SPOT disponibles sur le pôle de données surfaces continentales THEIA, produit chaque année une couverture d'orthophotos sur l'ensemble du territoire français. Il apparaît dès lors intéressant d'exploiter cette courverture pour générer une OCS millésimée.
La problématique de cartographie de l'occupation des sols automatique à partir d'images aériennes ou satellites occupe la communauté de télédétection depuis longtemps, par le biais de processus de classification supervisés, tels que les SVMs, ou les forêts aléatoires pour, entre autres, la vitesse d'exécution de ces derniers. Mais les résultats obtenus par ces méthodes n'ont pas encore permis une réelle automatisation, notamment en adéquation avec des spécifications existantes (erreurs encore trop importantes).
En parallèle de ces algorithmes depuis longtemps utilisés, des méthodes d'apprentissage automatique d'un genre nouveau, bien que reposant sur des concepts remontant aux années 1950, émergent depuis une décennie et sont étroitement liés aux recherches menées en *machine learning*. L'apprentissage profond, dont il est question ici, a fait ses preuves dans de nombreux domaines depuis le traitement naturel du langage, à la reconnaissance d'objets dans des images. Cet essor récent est la conséquence de la disponibilité de grandes bases de données d'apprentissage, ainsi que la démocratisation de l'utilisation de GPUs et de l'accroissement général des puissances de calcul. Représentants principaux de cette famille d'apprentissage, les réseaux de neurones profonds ont réellement bouleversé le monde actuel au quotidien. Que ce soit au niveau académique en terme de recherche, au niveau sociétal, au travers des smartphones par exemple (reconnaissances vocale, faciale, systèmes

de recommandation), ou même au niveau politique, avec les questions déontologiques que cela peut poser en terme de confidentialité des données (RGPD) et de protection des libertés individuelles, l'apprentissage profond est au cœur de technologies utilisées par la plupart des gens, de manière transparente et donc sans que ceux-ci s'en aperçoivent. En effet, pour afficher de telles performances dans tant de domaines, l'inconvénient pratique est le besoin très massif de données d'apprentissage lorsque l'on manipule ces algorithmes.
Les bases de données géographiques de l'IGN sont donc une opportunité dans notre cas, permettant d'exploiter au mieux les images très haute résolution monoscopiques acquises par les satellites SPOT 6 et 7 en les classifiant automatiquement par réseaux de neurones profonds appris sur ces mêmes bases de données. C'est cette approche que nous proposons dans ces travaux de thèse, avec une volonté d'étudier cette problématique tout en se plaçant dans un cadre plus large à visée opérationnelle, afin de proposer des cartographies sur de grandes étendues géographiques. Les expérimentations menées répondent aux questions soulevées lorsque l'on cherche à classifier de grandes zones : par exemple, la couverture annuelle SPOT produite par l'IGN étant unique, deux images adjacentes de cette couverture peuvent avoir été acquises à des époques différentes. Egalement, nous étudions les possibilités de transfert d'apprentissage par *fine-tuning* qui offre beaucoup d'avantages en matière de charges de calcul et de jeu d'apprentissage. Enfin, dans un contexte de mise à jour automatique de bases de données géographiques, l'exploitation jointe d'images aériennes et de réseaux de neurones profonds est étudiée, avec un accent mis sur la préparation des données d'apprentissage issues des bases de données géographiques de l'IGN qui présentent certains inconvénients.

Remerciements

En premier lieu, je remercie très sincèrement Clément et Arnaud à qui je dois beaucoup. Tout d'abord, ils ont su me convaincre d'entreprendre ces quatre années de doctorat, à un moment où j'étais encore fortement indécis sur l'après-école ! Je les remercie également de m'avoir accordé une grande confiance en me proposant ce sujet mettant en avant l'apprentissage profond, quasiment inexistant dans les laboratoires IGN et en télédétection au début de ces travaux, et pourtant aujourd'hui absolument omniprésent à plusieurs niveaux de la société. Le pari d'introduire cette notion à l'IGN au travers de cette thèse était grand mais a été réussi je crois ! Enfin, je les remercie pour un quotidien joyeux au sein de feu le MATIS, et de cette relation encadrants-doctorant naturellement décontractée qui est importante pour moi. C'est en qualité d'encadrant que je remercie Hichem qui a su nous éclairer sur les aspects théoriques en apprentissage, et pour ses conseils avisés sur la partie rédaction.
Merci aux rapporteurs, examinateurs qui ont pris le temps de parcourir ce manuscrit, d'y apporter correction et surtout d'avoir suscité une discussion intéressante sur divers points abordés dans ces travaux lors de la soutenance.
Merci à Sébastien sans qui je n'aurais sans doute jamais développé d'appétence particulière pour la télédétection. C'est en grande partie grâce à ses cours, pendant lesquels on pouvait ressentir une réelle passion pour ce domaine, et pendant ce stage à Forcalquier en école, que j'ai pu apprécier cette matière ! Il a en plus contribué avec Clément et Arnaud à cette ambiance si plaisante pendant ces travaux de thèse. Je n'oublie également pas mes co-doctorants d'alors de l'IGN, parmi lesquels il y avait Clément « Junior », Nathan, Oussama, Stéphane, et l'équipe de recherche Bruno, Mathieu, Sidonie, Loïc, Laurent, ce cher David V. et bien d'autres ! Merci à tous énormément.
En sortie de thèse, j'ai pu découvrir le travail en équipe, avec un sujet dans la continuité de celui de la thèse. Merci à mes collègues de l'équipe TERMOS, Camille, Anthony, Nicolas, Sylvain, Raphaëlle et Véronique !
Ces quatre années n'ont pas été simples tous les jours que ce soit pour moi ou mes proches. On dit que la thèse est une période solitaire, elle peut l'être oui sur le plan pratique, au quotidien, mais si ce ne fut déjà pas le cas pour moi grâce à un encadrement de qualité, j'avais également l'appui, sur le plan moral, de mes parents et mes sœurs que je remercie énormément. Chloé, tu as cru en moi jusqu'au bout, au même titre que beaucoup d'autres, sauf que pour toi c'était de très près et au quotidien, aussi je ne te remercierai jamais assez pour la motivation que tu m'as apportée et le soutien constant.
L'amitié est également une grande source de réconfort. Aussi, Laurane, Tim, Hippolyte, Régis, Tristan, Quentin, Pauline, Sylvain, Nicolas, Thomas, mais aussi Marion qui sait à quel point cette période peut être une succession de hauts et de bas, je vous dois beaucoup, même si ce n'est pas sur le plan technique. Votre présence, vos messages de soutien pouvaient suffire à me changer les idées, ou même apporter du recul sur certains aspects de ces travaux (si si !). Donc un grand merci à vous également !

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

OCS	Occupation des Sols
THRS	Très Haute Résolution Spatiale
AAE	Agence européenne pour l'environnement
GIEC	Groupe d'Experts Intergouvernemental sur l'Evolution du Climat
PLU	Plan Local d'Urbanisation
SCoT	Schéma de Cohérence Territorial
BRGM	Bureau de Recherches Géologiques et Minières
UMC	Unité Minimale de Collecte
OCS GE	OCS à Grande Echelle
IGN	Institut National de l'Information Géographique et Forestière
GMES	Global Monitoring for Environment and Security
CES	Centre d'Expertise Scientifique
OSO	Occupation des SOls
CESBIO	Centre d'Etudes Spatiales de la BIOsphère
RGE	Référentiel à Grande Echelle
BD	Base de Données
MOS	Mode d'Occupation des Sols
CRIGE-PACA	Centre Régional de l'Information GEographique en région Provence-Alpes-Côte d'Azur
RaDAR	Radio Detection And Ranging
SPOT	Satellite Pour l'Observation de la Terre
SWIR	Short-Wave Infra-Red
MODIS	Moderate-Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
OLI	Operational Land Imager
TIRS	Thermal Infrared Sensor
MERIS	MEdium Resolution Imaging Spectrometer
ESA	European Space Agency
CNES	Centre National d'Etudes Spatiales
EADS	European Aeronautic Defence and Space company
HRG	High Resolution Geometrical
ISODATA	Iterative Self-Organizing Data Analysis Technique Algorithm
EM	Espérance Maximisation
SVM(s)	Support Vector Machine(s) ou Séparateur(s) à Vaste Marge
RBF	Radial Basis Function
OOB	Out-Of-Bag
LiDAR	Light Detection And Ranging
NDVI	Normalized Difference Vegetation Index
PIR	Proche Infra-Rouge

IB	Indice de Brillance
MLP	Multi-Layer Perceptron
CNN(s)	Convolutional Neural Network(s)
TDNN	Time-Delay Neural Network
RNN	Recurrent Neural Network
LSTM	Long-Short Term Memory
RBM	Restricted Boltzmann Machine
DNN	Deep Neural Network
CPU	Central Processing Unit
GPU	Graphics Processing Unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ReLU	Rectified Linear Unit
SGD	Stochastic Gradient Descent
GAN	Generative Adversarial Network
VGG	Visual Geometry Group
VOC	Visual Object Classes
CIFAR	Canadian Institute For Advanced Research
MNIST	Modified National Institute of Standards and Technology
FCN	Fully Convolutional Network
ISPRS	International Society for Photogrammetry and Remote Sensing
RGE	Référentiel à Grande Echelle
RPG	Registre Parcellaire Graphique
VP	Vrai Positif
FP	Faux Positif
VN	Vrai Négatif
FN	Faux Négatif
IoU	Intersection over Union
OA	Overall Accuracy ou Précision Globale
IRSTEA	Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture
CEREMA	Centre d'Etudes et d'Expertise sur les Risques, l'Environnement, la Mobilité et l'Aménagement
DGALN	Direction Générale de l'Aménagement, du Logement et de la Nature
IFFSTAR	Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux
INRA	Institut National de la Recherche Agronomique

Chapitre I

Introduction

Si la connaissance de la couverture de la surface terrestre a historiquement été importante pour des raisons sociales et économiques, celle-ci est également devenue fortement liée aujourd'hui aux préoccupations en matière d'environnement et de climat. Les changements globaux en jeu sont visibles à la surface de la Terre, que ceux-ci soient anthropiques ou bien naturels.
L'observation à large échelle des dynamiques continentales s'avère donc cruciale. Les capteurs spatiaux d'observation de la Terre répondent à ce besoin au travers de données images à différentes résolutions spatiales et spectrales. Cette source de données quasi continue permet de donner matière à des processus de cartographie du sol.
Dans un monde où la surface terrestre évolue constamment, les changements sont très rapides, il est dorénavant essentiel de cartographier à une fréquence soutenue le paysage. Or, les capteurs fournissant énormément de données à analyser, il devient ardu de répondre à cela par des moyens purement manuels. C'est pourquoi les communautés scientifiques se tournent vers des

processus automatiques, soutenus par les prescripteurs et les utilisateurs finaux de la couverture des sols. En particulier, les algorithmes d'apprentissage profonds ont démontré leur efficacité dans de nombreux domaines et se révèlent prometteurs en matière de cartographie des sols.
Après avoir défini le terme d'occupation des sols et passé en revue les différents produits existants, un état de l'art des capteurs dédiés à l'observation de la Terre est réalisé, avec une focalisation sur les capteurs à Très Haute Résolution Spatiale (THRS). D'un point de vue méthodologique, on explicite pourquoi l'approche automatique est intéressante, les travaux existants en faisant usage. Enfin, la problématique de la thèse ainsi que ses tenants et aboutissants sont explicités en fin de chapitre, avec l'organisation du manuscrit.

1 Cartographie de l'occupation des sols

La définition que l'on peut donner de l'occupation des sols est la suivante : il s'agit de la description physique de l'ensemble de la surface terrestre selon une nomenclature donnée. On considère donc par ce terme aussi bien les surfaces anthropiques que naturelles.
Cette description est à différencier de l'usage du sol qui traduit la fonction sociale ou économique exercée par l'objet présent sur l'occupation des sols. Pour illustrer la différence entre les deux cartographies, en milieu urbain, là où l'occupation intègre généralement une classe unique *bâti*, l'usage, quant à lui, distingue au sein de cette classe les bâtiments à caractère résidentiel, commercial ou bien industriel. L'utilisation seule d'images satellite pour la classification du sol rend ardue la détection de classes d'usage du sol. En effet, l'analyse purement de la radiométrie ne permet pas de représenter de telles classes dont les définitions reposent purement sur des considérations sociales ou économiques. Occupation et usage sont souvent mêlées au sein de bases de données géographiques existantes. Par exemple, Corine Land Cover, base de données établie à l'initiative de l'Agence européenne pour l'environnement (AAE), regroupe aussi bien des classes d'occupation telles que *Forêts et milieux semi-naturels* et *Surface en eau*, que des des classes d'usages telles que *Zones industrielles ou commerciales et installations publiques*.
Outils aussi bien scientifiques que décisionnels, les cartes d'occupation des sols fournissent une description utile et nécessaire à la compréhension et au suivi de phénomènes naturels ou anthropiques, à de nombreuses échelles spatiales et temporelles. Elles constituent une base de travail indispensable à des organismes comme le GIEC (*Groupe d'experts intergouvernemental sur l'évolution du climat*), notamment sur la question du stockage du carbone dans le sol et dans les arbres, et le problème posé par la déforestation (libération du carbone dans l'atmosphère). La connaissance de la nature du sol à différentes dates guide la prise de mesures et la gouvernance à adopter dans un contexte où les changements climatiques sont globaux. Le rapport de la Cour des Comptes de 2012 (Cour des Comptes, 2012) en témoigne puisqu'il est révélé que la réaffectation des sols et le défrichement pour la production de biocarburants conduit finalement à une augmentation d'émissions de gaz

à effet de serre plus importantes que celles issues d'énergies hydrocarbures (en plus de soustraire une partie de la surface totale des terres arables à destination alimentaire). En terme de gestion des risques liés à l'environnement, la cartographie des sols joue un rôle crucial (prévention, simulation, acheminement de secours). Par exemple, les surfaces inondables, en croissance continue due au réchauffement climatique global et à la montée des eaux, les forêts susceptibles de subir un incendie, ou encore les zones d'avalanche sont référencées et disponibles en France grâce à l'outil mis en place en juillet 2014 par le Ministère de l'Ecologie, du Développement durable et de l'Energie et le BRGM (Bureau de recherches géologiques et minières) [1], suite à la tempête Xynthia de 2010.
En matière d'urbanisme et politiques territoriales en France, à des échelles spatiales plus fines, des documents tels que les Plans Locaux d'Urbanisme (PLU), Schémas de Cohérence Territoriale (SCoT) s'appuient sur la connaissance de l'occcupation des sols. En particulier, l'étalement urbain et l'imperméabilisation croissante des sols sont des sujets sur lesquels la législation a imposé des limites, afin de préserver les zones naturelles et les surfaces agricoles utiles. A cette fin la Loi d'avenir pour l'agriculture, l'alimentation et la forêt met notamment en place au niveau départemental une commission de la préservation des espaces naturels, agricoles et forestiers [2] qui émet des avis lors de l'élaboration de PLU et SCoT, au regard d'un objectif de préservation des terres agricoles et naturelles.
La connaissance de l'occupation des sols étant donc essentielle pour de nombreux usages et de nombreux utilisateurs. Il convient pour chacun d'entre eux de fournir l'outil adéquat. Pour cela, les cartes d'occupation sont caractérisées par plusieurs spécifications détaillés en première partie. En seconde partie, nous exposons un certain nombre de cartes d'occupation des sols existantes, à différentes échelles. Enfin, les principes de la télédétection spatiale ainsi qu'une présentations des divers capteurs disponibles sont abordés en troisième et quatrième parties respectivement.

1.1 Caractérisation d'une carte d'occupation des sols

Un certain nombre de caractéristiques permettent de décrire une carte d'occupation des sols d'un point de vue méthodologique :

1. La zone couverte sur la surface terrestre (échelle locale ? régionale ? nationale ? globale ?).
2. La sémantique incluant :
    - la nomenclature : les classes d'intérêt choisies pour décrire l'environnement que l'on cherche à étudier. Une nomenclature peut avoir une granularité plus ou moins fine, ce qui correspond au

1. http ://www.georisques.gouv.fr/cartes-interactives
2. Loi n° 2014-1170 du 13 octobre 2014 d'avenir pour l'agriculture, l'alimentation et la forêt

niveau de détail souhaité : les nomenclatures existantes adoptent bien souvent une structure hiérarchique emboîtée [3] ;

- l'adéquation de cette sémantique lorsque l'on compare les classes attribuées par rapport à la réalité observable ;

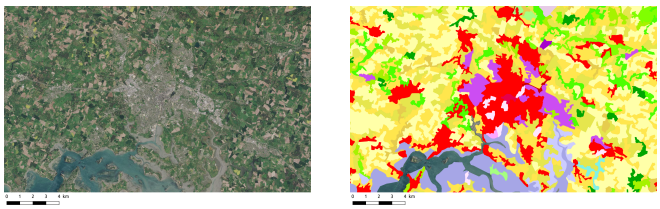
3. Les règles géographiques, appelées spécifications :
   - la résolution spatiale si la carte est sous format raster (pixels) ou l'unité minimale de collecte (UMC) dans le cas de carte vectorielle (base de données géographiques). Tout objet inférieur en surface à cette résolution ne sera pas inventorié par la carte ;
   - la précision géométrique permettant de qualifier la localisation des objets par rapport à la réalité : cette précision est souvent conditionnée par l'étendue géographique de la carte. Par exemple, celle-ci couvre le globe entier, et donc l'on privilégie une analyse statistique, la délimitation géométrique exacte des objets n'aura que peu d'importance ;
4. le millésime associé à la carte : pour le suivi de phénomènes temporels, il est indispensable de disposer de cartes à plusieurs dates afin de retracer leur évolution.

Tous ces éléments contribuent à l'élaboration du cahier des charges pour la carte d'occupation des sols souhaitée par l'utilisateur final. Si ces points peuvent apparaître triviaux, il est néanmoins nécessaire de les aborder avec rigueur car cela conditionne toute la suite du processus de cartographie d'occupation des sols.

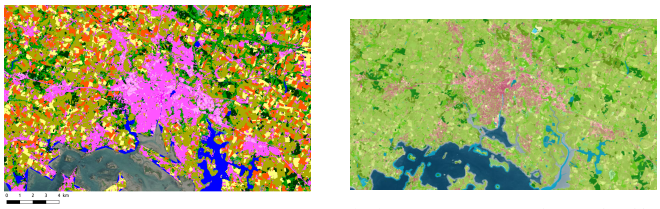
1.2 Offres actuelles en OCS

Au regard des paramètres passés en revue en 1.1, différentes cartes d'occupation des sols ont déjà été produites. Nous nous restreignons aux produits englobant le territoire français afin de respecter le cadre de notre étude.
La figure I.1 présente différentes cartes d'occupation, dont les caractéristiques sont ensuite détaillées carte par carte, pour la ville de Vannes, ainsi que l'image SPOT 6/7 de 2017 couvrant la zone. Il est visuellement possible d'apprécier un certain nombre des paramètres évoqués en 1.1. Deux notions d'échelle sont mis en avant dans la comparaison de ces différentes cartographies : (i) l'échelle au sens de la *résolution* de la cartographie, autrement dit par rapport à l'objet le plus petit observable, et (ii) l'échelle relative à la part du territoire couverte par la cartographie. Tout d'abord les différentes échelles d'information disponibles : Corine Land Cover (I.2) propose une classification à l'échelle de l'Europe, permettant une analyse pertinente du paysage à une échelle très globale et nationale. Cependant, un tel niveau d'étude rend difficile la description fine de la structure d'une ville ; en particulier les routes sont absentes, ayant une finesse morphologique trop importante pour rentrer dans les critères de représentation de la base de données (échelle spatiale trop

3. On entend par hiérarchie emboîtée la notion de classes « mères » et classes « filles » : la classe *bâti* peut être ainsi subdivisée en *habitat* et *bâti à usage commercial ou industriel*.



(A) Image SPOT. (B) Corine Land Cover.



(C) Carte produite par le CESBIO dans le cadre du CES OSO. (D) OCS-Grande Echelle de l'IGN.

FIGURE I.1. Différentes cartes d'occupation des sols sur la région de Vannes. Les nomenclatures sont différentes, les granularités de chaque OCS différant des autres.

grossière). Certaines cartes sont donc plus pertinentes pour des usages statistiques (calculs d'indicateurs, de détection grossière de changements) que pour une finalité cartographique pure (délimitation précise de tous les objets d'intérêt). Pour étudier un paysage en milieu urbain dense par exemple, on préfère une carte offrant une résolution spatiale plus fine telle que proposée par l'OCS GE (Occupation des sols à Grande Echelle) de l'IGN (Institut National de l'Information Géographique et Forestière), visible sur la figure I.1.

Initiatives globales

A l'échelle européenne, le projet européen GMES (Global Monitoring for Environment and Security), aujourd'hui connu sous le nom Copernicus, a la charge de fournir des données d'observation de la Terre dans un objectif de surveillance. Ce programme est à l'origine de la réalisation de la base de données d'occupation des sols Corine Land Cover. Celle-ci fait également partie du champ de la directive INSPIRE (Infrastructure d'information géographique dans la Communauté européenne), et a pour but de cibler les classes prioritaires d'un point de vue environnemental, de coordonner les efforts pour la collection des données relatives à ces classes et de s'assurer de la compatibilité entre toutes les données collectées par les différents organismes impliqués. Les données Corine Land Cover sont gratuites et libres d'accès.
Corine Land Cover couvre 39 pays d'Europe - Union Européenne et pays limitrophes (figure I.2) - et est produite manuellement par photo-interprétation

d'images satellites de 20 à 25 mètres de résolution, tout en s'aidant de données plus résolues tout de même pour les zones les plus denses notamment, afin de répertorier les objets ayant une surface (Unité Minimale de Collecte - UMC) de 25 hectares. Quatre millésimes ont été produits jusqu'à aujourd'hui : en 1990, 2000, 2006 et 2012. Les produits ultérieurs à 1990 s'accompagnent de cartes de changements entre deux itérations successives. La dernière itération en date de CLC a été mise à disposition fin 2018. Mêlant occupations et usages des sols, la nomenclature est hiérarchique comportant trois niveaux de détail, pour un total de 44 classes sur le niveau le plus riche, regroupées en cinq grands types d'occupation du territoire : *Territoires artificialisés*, *Territoires agricoles*, *Forêts et milieux semi-naturels*, *Zones humides* et *Surfaces en eau*. Avec une UMC de 25 hectares, l'échelle d'utilisation de la base de données Corine Land Cover avoisine typiquement 1 :100 000 et permet des analyses du territoire à un niveau européen, national ou régional. L'UMC contraint l'absence du réseau routier de la base de données, limitant fortement les possibilités d'utilisation en milieu urbain.
Si Corine Land Cover ne fournit pas de détails suffisants à l'échelle urbaine, le programme Copernicus met à disposition un second produit dédié à l'analyse urbaine appelé Urban Atlas. Il en existe deux versions, présentant toutes deux une UMC de 0,25 hectare, l'une produite en 2006, la seconde en 2012. Cette base de données est disponible sur toute l'Europe mais de façon localisée (discontinue). En effet, si l'on considère la version de 2012, seules les agglomérations de plus de 50 000 habitants y sont référencées, pour un total de 697 unités urbaines.

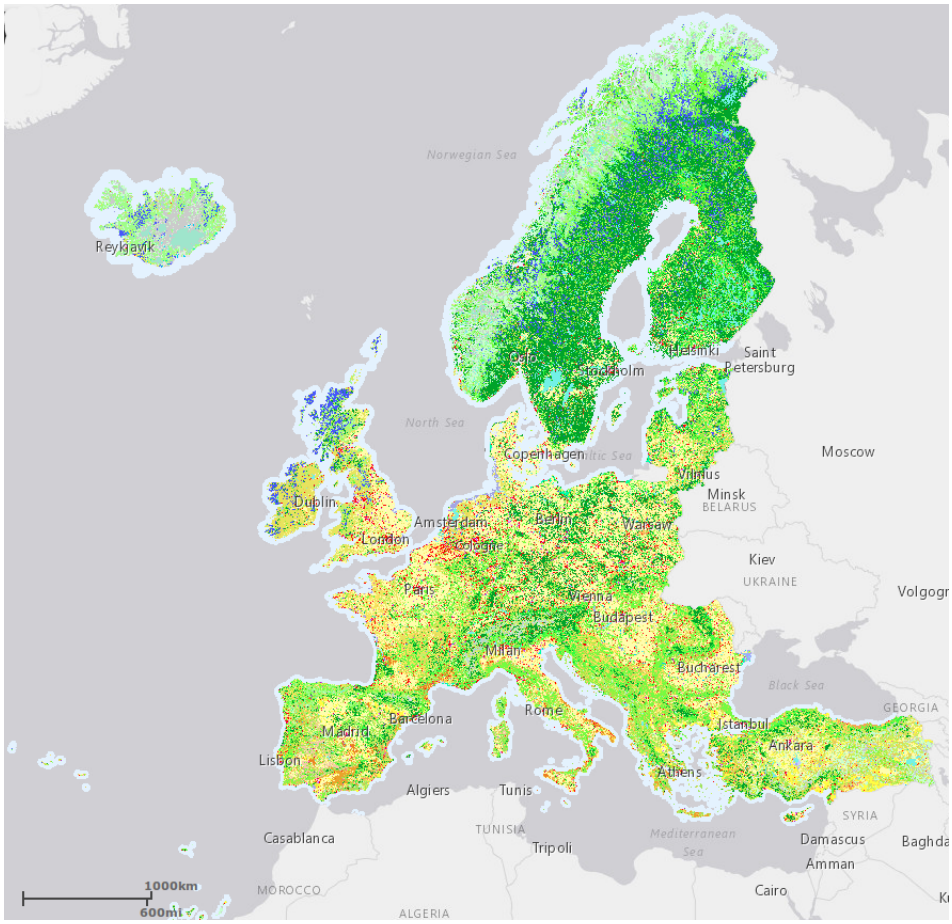


FIGURE I.2. Couverture de la base de données européennes d'occupation des sols Corine Land Cover en 2012.

Initiatives nationales

De multiples projets d'occupation des sols existent en France, et sont plus ou moins aboutis. Chacun d'entre eux répond généralement à un besoin local à court terme. Quelques exemples seront évoqués dans la partie suivante (Section 1.2). Même si ceux-ci restent pertinents dans un contexte régional, voire à une échelle (territoriale) plus petite encore, l'étude de certaines thématiques nationales telles que le parc forestier ou l'artificialisation des sols en France nécessite des cartes d'occupation des sols homogènes sur le territoire, tant au niveau de la nomenclature, que des processus de production mis en œuvre pour construire ces cartes. A cette fin, le pôle de données et de services surfaces continentales Theia a initié, au travers du Centre d'Expertise Scientifique (CES) Occupation des SOls (OSO) porté par le Centre d'Etudes Spatiales de la BIOsphère (CESBIO), la production automatique de cartes annuelles d'occupations des sols à l'échelle de la France entière (Inglada et al., 2017). L'utilisation de séries temporelles Sentinel-2 dans un processus de classification supervisée permet notamment de discriminer efficacement en milieu naturel les différentes entre résineux et feuillus pour les zones forestières, ou encore le type de culture en présence. Le but de ce CES est de proposer

chaque année une nouvelle carte d'occupation calculée sur les données satellites acquises l'année précédente. Ainsi, la première carte disponible pour la France entière a été mise à disposition début 2017, et produite à partir des données récoltées en 2016 par les deux satellites Sentinel-2 (Figure I.3). La carte résultante comporte entre 14 et 17 classes selon la nomenclature adoptée en milieu urbain, et est disponible avec une résolution à (i) 20 m nativement ou (ii) 10 m après sur-échantillonnage. Une seconde carte a été mise à disposition début 2018, après classification des séries temporelles acquises en 2017.
De son côté, l'Institut National de l'Information Géographique et Forestière (IGN) a pour mission l'élaboration d'une référence nationale unique sur la France entière. Depuis 2013, cette référence est définie par des acteurs nationaux mais aussi régionaux pour répondre également aux besoins locaux, notamment en terme de compatibilité avec les nomenclature présentes sur les cartes d'occupations locales. Cette référence répond aussi au besoin de conformité avec la directive INSPIRE. En plus, des spécifications supplémentaires contraignent davantage cette référence : l'adéquation avec le Référentiel Grande Echelle (RGE) de l'IGN qui rassemble diverses couches de données (BD Ortho, BD Topo, BD Adresse, BD Parcellaire, RGE Alti), une précision suffisante pour permettre son utilisation à une échelle locale (cohérence spatiale avec les SCoTs), une fréquence de mise à jour régulière et rapide. A l'issu de groupes de travail du CNIG (Conseil National de l'Information Géographique), la mission a été donnée à l'IGN d'établir une nomenclature prenant en compte les attentes mentionnées précédemment [4] : l'occupation des sols à grande échelle, ou OCS GE, visible sur figure I.1. La nomenclature ainsi constituée s'étire sur quatre dimensions distinctes et séparées les unes des autres : occupation des sols, usages des sols, attributs morphologiques et éléments de caractérisation. L'OCS GE propose une granularité sémantique proposant 24 classes au niveau le plus fin. Elle est générée en deux temps :

1. différentes bases de données sont superposées pour produire une première ébauche de la carte. A titre d'exemple, routes, bâtiments, réseau hydrologique sont extraits de la BD TOPO®, les forêts de la BD Forêt®, et les cultures du Registre Parcellaire Graphique (RPG). Ce processus d'extraction des bases de données permet une automatisation partielle de la production.
2. afin de combler les vides laissés par les bases de données utilisées à l'étape précédente, un travail de photo-interprétation manuelle est exécuté sur les images de la BD ORTHO®. Il s'agit également ici de corriger les erreurs liées aux bases de données qui n'ont pas nécessairement intégré des changements récents sur l'occupation. Cette étape, bien que nécessaire, est coûteuse à la fois en temps de traitement et en opérateurs humains mobilisés.

Il existe trois UMC pour les objets recensés dans l'OCS GE : (i) 0.02 hectare pour les zones bâties, (ii) 0.05 hectare pour les zones construites mais non

4. Lettre de mission du groupe de travail de la CNIG pour l'établissement d'une OCS nationale, 2010-2012.

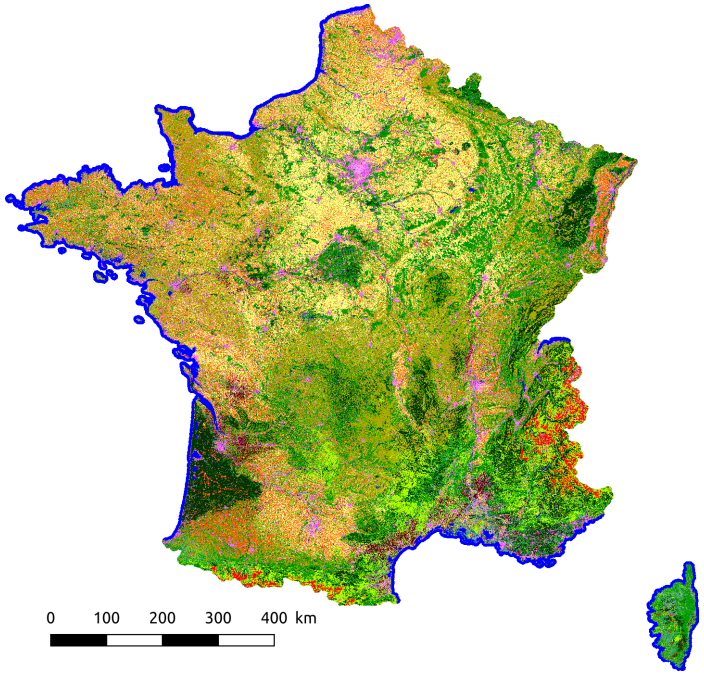


FIGURE I.3. Carte d'occupation des sols de 2017, calculées avec des images Sentinel-2 acquises en 2016, dans le cadre du CES OSO.

bâties (réseaux routiers, parking, réseaux ferrés...), (iii) 0.25 hectare pour les zones non construites (forêts, surfaces d'eau, parcelles agricoles...).

Initiatives locales

A un niveau plus local géographiquement parlant, deux modes de production existent. L'utilisation de bases nationales de données d'occupation des sols comme socle pour construire une carte avec une granularité plus fine spatialement est possible, justement parce que la portée restrainte géographique d'une cartographie locale permet de mettre plus de moyens pour améliorer ces aspects. Le Mode d'Occupation des Sols (MOS), actualisé neuf fois depuis 1982 (dernier millésime datant de 2017), a été réalisé par l'Institut d'Aménagement et d'Urbanisme d'Ile-de-France, et se concentre d'ailleurs sur cette région. Le MOS, visible en figure I.4, comprend 81 postes, dont certains ne sont présents dans aucun des produits plus larges (nationaux ou globaux), constituant la deuxième approche possible pour établir ces cartes locales d'occupation.
Avec ces nouvelles initiatives qui se multiplient (en Provence-Alpes-Côtes-d'Azur par exemple avec la cartographie du CRIGE-PACA), la télédétection spatiale sur images satellitaires à très haute résolution peut améliorer grandement les processus de production. En effet, à l'heure actuelle, ces cartographies locales s'appuient sur la photo-interprétation manuelle à partir d'images aériennes, que l'on sait coûteuse et dont les acquisitions sont peu fréquentes. La télédétection spatiale, à l'inverse, permet d'assurer une continuité avec des capacités de revisite régulière (quasi-hebdomadaires) des satellites à très haute résolution spatiale.

FIGURE I.4. La carte « Mode d'Occupation du Sol » d'Ile de France.

2 Capteurs THRS : une opportunité pour l'OCS

Etroitement liés aux activités militaires et aux applications environnementales et dans les géosciences, les capteurs satellites acquièrent des données aujourd'hui indispensables pour notre société. Du positionnement géodésique aux prévisions météorologiques, les capteurs embarqués en orbite autour de la Terre fournissent des mesures essentielles à l'élaboration de modèles d'évolution de divers phénomènes, et à la compréhension des écosystèmes dans lesquels nous vivons. En particulier, les capteurs de télédétection permettent d'imager de grandes régions en peu de temps. Dans le cas du problème d'occupation des sols, seuls les capteurs imageurs offrant des données sur les surfaces émergées nous intéressent.
Après avoir donné des clefs de compréhension concernant les caractéristiques des capteurs imageurs, un rapide recensement des capteurs satellites dans le contexte de la télédétection spatiale est effectué, avec un intérêt particulier pour les capteurs optiques à très haute résolution spatiale, en particulier pour le couple de satellites SPOT 6 et 7 (Satellite Pour l'Observation de la Terre).

2.1 Caractéristiques des capteurs satellites

La télédétection spatiale consiste à analyser, dans une ou plusieurs longueurs d'onde donnée(s), la fraction de rayonnement rétrodiffusé par la surface terrestre arrivant au capteur. Cette analyse conduit ensuite à l'interprétation du signal sous forme de classification par exemple dans notre cas : le comportement variable du signal par rapport à la surface sur laquelle il a été rétrodiffusé renseigne sur la composition et la nature de celle-ci.Si ce signal est à l'origine émis par le rayonnement solaire, le capteur sera dit *passif*, tandis qu'un capteur dit *actif* émet son propre signal et mesure la quantité rétrodiffusée ou réfléchie (selon le type de capteur) par la surface. Une conséquence de la nature des capteurs actifs est la possibilité de mesurer aisément et directement (i) la quantité de signal reçue par rapport à celle émise,

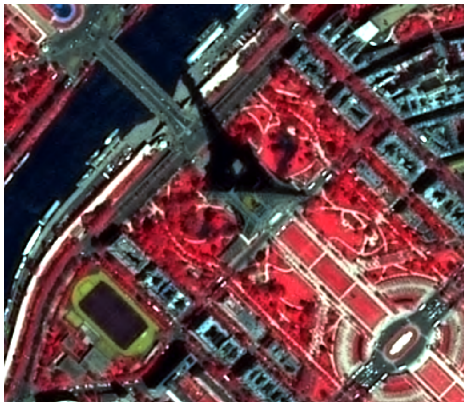
(ii) le changement de polarisation ou non de l'onde émise. Les capteurs RADAR (Radio Detection And Ranging) et LiDAR (Light Detection And Ranging) sont des exemples de capteurs actifs, émettant des signaux dans les longueurs d'onde radio et visible/proche-infrarouge respectivement. L'étude se limite uniquement à l'utilisation de capteurs mesurant le rayonnement rétrodiffusé à l'origine émis par le Soleil. L'inconvénient par rapport aux capteurs actifs est la nécessité d'acquérir les données pendant la journée sur la région d'intérêt.
Les images acquises par les capteurs actifs stockent en chaque photosite, ou pixel, la quantité de photons rétrodiffusée par la surface imagée. On appelle cette quantité la réflectance après corrections atmosphériques (modèles de transfert radiatif prenant en compte les aérosols, l'humidité...). Cette quantité mesurée par le capteur dépend (i) de la longueur d'onde (ii) du matériau sur lequel la rétrodiffusion a lieu. Il est donc intéressant d'enrichir une image en multipliant les bandes spectrales dans lesquelles est acquis ce signal rétrodiffusé, les comportements variant selon la nature du sol et permettant de construire une signature spectrale pour chaque type d'occupation. Pour illustrer ce concept de signature, la Figure I.6 permet de voir les spectres de réflectance de trois types d'occupation du sol, *sol nu*, *neige*, *végétation*. Les bandes grises visibles sur la figure correspondent aux bandes rouge, proche-infrarouge et SWIR (Shortwave Infrared) du satellite MODIS. C'est pourquoi les capteurs embarqués sur les satellites dédiés à la télédétection sont des capteurs qualifiés de *multispectraux* ou *superspectraux*.
Outre le nombre de bandes et la largeur de ces bandes (résolution spectrale), les images ont une taille de photosite intrinsèque, décidée à la conception du capteur. Cette taille est souvent assimilée, à tort, à la résolution spatiale du capteur qui se réfère en réalité à la distance minimale entre deux objets que le capteur peut séparer. La taille des photosites contraint la taille d'objet minimale que l'on peut distinguer sur l'image. Une résolution spatiale (on utilise la définition abusive de ce terme) fine permet de distinguer davantage de détails sur la surface terrestre, mais entraînent diverses contraintes :

— à surface à imager constante, le stockage nécessite plus d'espace que des images à résolution spatiale moins fine ;

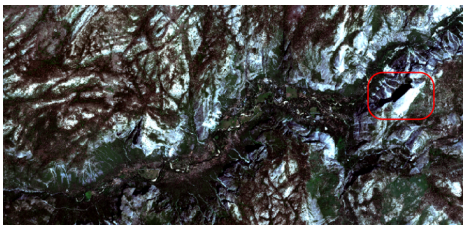
— les photosites de la matrice du capteur étant plus petits, peu de photons parviennent à chaque photosite, il faut donc élargir les bandes spectrales pour obtenir un rapport signal sur bruit suffisamment grand.

La décision de la résolution spatiale d'un capteur est en fait un compromis entre résolution spatiale, spectrale et rapport signal sur bruit.
Enfin, les capteurs peuvent être différenciés selon leur période de revisite, caractérisant la capacité du capteur à imager entre deux instants immédiatement consécutifs la même région. Cet aspect non négligeable conditionne la possibilité de mener des travaux mettant en jeu des séries temporelles d'images, qui représentent la même scène à différents instants, enrichissant cette fois, non plus spectralement, mais temporellement, les caractéristiques des objets vus au sol. Par ailleurs, ces capteurs sont très souvent embarqués

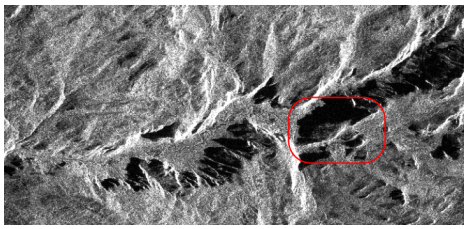
(A) La tour Eiffel vue par la caméra aérienne de l'IGN - résolution de 30cm.



(B) La Tour Eiffel vue par le capteur embarqué sur SPOT 6 et 7. Les canaux RVB contiennent respectivement les bandes IR, R, V fournies par le capteur.



(C) Image de la vallée de Yosemite, issue du capteur **multispectral** embarqué sur Sentinel-2. Le symbole de Yosemite, le Half Dome, est délimité en rouge.



(D) Image de la vallée de Yosemite, issue du capteur **radar** embarqué sur Sentinel-1. Le symbole de Yosemite, le Half Dome, est délimité en rouge.

FIGURE I.5. Des capteurs différents en réponse à des thématiques et besoins différents.

sur des satellites à orbite hélio-synchrone (un tel satellite repasse quotidiennement au dessus d'un même lieu à une heure solaire locale identique), couvrant ainsi tout le globe terrestre.

2.2 Satellites dédiés à la télédétection terrestre

Les capteurs d'observation de la Terre sont aussi variés que nombreux, tant au niveau des plateformes (aérien / satellite) que de leurs capacités spectrales et spatiales. La figure I.5 propose diverses interprétations du paysage terrestre, chacune issue d'un capteur dont les caractéristiques lui sont propres.
Le premier satellite dédié à des tâches de télédétection fait partie du programme civil de la NASA (National Aeronautics and Space Administration) Landsat. Il s'agit de Landsat-1, lancé en 1972. Il permettait l'acquisition de 7

bandes spectrales à une résolution de 80 m, via les capteurs embarqués *Return Beam Vidicon* et *Multispectral Scanner*. Six satellites lui ont succédé successivement jusqu'à Landsat-8 dernier satellite en date de ce programme, lancé en 2013. Les progrès techniques permettent à celui-ci d'accroître les performances en acquérant dans 11 bandes spectrales, à une résolution de 30 m pour les bandes acquises par l'*Operational Land Imager* (OLI) (utiles pour l'analyse des surfaces émergées). Ce capteur fournit des images à haute résolution, avec une période de revisite de 16 jours. Une nouvelle itération Landsat-9 est prévue pour 2020, dotée d'un capteur optique (OLI) similaire à celui de Landsat-8, mais avec une version améliorée du capteur thermique (*Thermal Infrared Sensor*, ou TIRS). Ce capteur thermique est notamment utilisé pour des thématiques liées au cycle de l'eau et à la gestion de cette ressource. Dans les années 2000, les capteurs multispectraux à faible période de revisite se sont multipliés avec tout d'abord les satellites américains Terra et Aqua, embarquant le radiomètre MODIS (*Moderate Resolution Imaging Spectroradiometer*), capturant des images dans 36 bandes de 0.4 µm à 14.4 µm, à une résolution de 250 m à 1 km. Les satellites couvrent la Terre entière en 2 jours environ, et sont particulièrement attrayants pour les analyses à l'échelle mondiale liées à l'océan (chlorophylle), mais aussi au contrôle de la végétation (santé, déforestation, incendies). Toutefois, dans le but de dresser la cartographie d'occupation par des objets de dimensions pouvant approcher celles de bâtiments, ces capteurs sont inadaptés pour nos travaux. Le capteur MERIS (*MEdium Resolution Imaging Spectrometer*) embarqué à bord d'Envisat, opéré par l'ESA (Agence Spatiale Européenne) et dont la mission s'est achevé en 2012, avait des spécifications techniques semblables à celles du capteur MODIS. Le lancement de RapidEye en 2009 par l'industriel allemand éponyme, a permis la multiplication des images grâce à une constellation de 5 satellites, pour une période de revisite conjointe de 24h. L'acquisition est faite sur 5 bandes spectrales, à une résolution spatiale de 5 m.
Plus récemment, dans le cadre du programme Copernicus, la constellation Sentinel de satellites d'observation de la Terre a été mise en place par l'ESA. Cette constellation, dont la vocation est la surveillance et le suivi de phénomènes temporels, constitue une nouveauté par (i) l'ampleur du projet, mettant en jeu divers moyens d'acquisition (actifs et passifs), (ii) la couverture du spectre électromagnétique et (iii) la transversalité des thématiques couvertes (océans, atmosphères et terres émergées). Les satellites Sentinel-1A et Sentinel-1B (Torres et al., 2012) ont initié ce programme lors de leur lancement en 2014 et 2016, embarquant chacun un capteur radar, acquérant ainsi en continu (de jour comme de nuit) mais également dans toutes les régions du globe sans difficulté (les longueurs d'onde utilisées n'étant pas ou rarement impactées par les effets atmosphériques). Ont suivi Sentinel-2A et Sentinel-2B (Drusch et al., 2012), lancés respectivement en 2015 et 2017, munis de capteurs optiques multispectraux : 10 bandes de 10 m à 20 m pour les applications d'occupation des sols, 3 bandes à 60 m permettant des corrections atmosphériques et des recherches dirigées sur l'atmosphère. Le binôme Sentinel-2 offre une période de revisite de 5 jours, alliant ainsi haute fréquence temporelle, haute résolution spatiale et une couverture spectrale

importante. Les données issues de ces deux satellites constituent donc une opportunité, sans précédent, très intéressante pour la cartographie de l'occupation des sols. En particulier, pour discriminer certaines classes non pérennes telles que les différents types de culture, la connaissance spectrale à différentes dates est cruciale. La troisième partie du programme, Sentinel-3 (Donlon et al., 2012), est dédiée à l'étude des océans et des températures océaniques et continentales ; le premier satellite a été lancé en 2016, le second en 2018.

La France, en collaboration avec la Belgique et la Suède, a mis en place en 1970 le programme SPOT (Satellite Pour l'Observation de la Terre), porté par le CNES (Centre National d'Etudes Spatiales). SPOT Image (société anonyme du CNES) était propriétaire des images jusqu'au rachat des dernières parts du CNES par Airbus Defence and Space en 2008. En 1986, SPOT 1 est lancé, doté d'un capteur acquérant dans 3 bandes spectrales (rouge, vert, proche-infrarouge) à 20 m et d'un canal panchromatique à 10 m. Ses successeurs, SPOT 2 et SPOT 3 partagent des spécifications techniques identiques que SPOT 1. SPOT 4, lancé en 1998, est également sensiblement similaire aux itérations précédentes, ajoutant toutefois une bande supplémentaire dans le moyen infrarouge. En 2002 est lancé SPOT 5, apportant une résolution spatiale accrue en proposant une image panchromatique à 5 m et un « supermode » permettant de produire une image à 2.5 m à partir de deux images panchromatiques acquises avec un nouvel instrument, le HRG (*High Resolution Geometrical*). L'acquisition superspectrale est à 10 m de résolution et 20 m pour la bande SWIR. Lancés en 2012 et 2014, SPOT 6 et SPOT 7 sont les deux plus récents satellites de ce programme. Les images acquises ont une résolution de 6 m pour les canaux spectraux (rouge, vert, bleu, proche infrarouge) et de 1.5 m pour le canal panchromatique. Les capteurs embarqués sur ces satellites sont qualifiés de capteurs à très haute résolution spatiale. En l'espace de 25 ans, la résolution a été affinée d'un facteur 6.5, offrant aujourd'hui la possibilité de cartographier des paysages difficilement discernables à des résolutions avoisinant 20 m. En effet, une résolution accrue permet d'enrichir les textures des images en milieu urbain dense où il est ardu de reconnaître des motifs caractéristiques de ce paysage à l'aide d'images Sentinel-2. De plus, cette résolution donne accès à la détection d'objets trop fins pour d'autres capteurs, tels que les haies ou les routes. Cette nouvelle richesse sémantique est accompagnée d'une amélioration de la géométrie d'objets auxquels des capteurs à haute résolution spatiale avaient déjà accès. Les parcelles agricoles sont ainsi mieux délimitées et la mise à jour de bases de données sur ce type de classe en est facilitée. Par rapport à des capteurs satellites mieux résolus tels que Pléiades, les satellites SPOT ont la possibilité de couvrir le territoire entier en un temps limité, avec des période de revisite et des fauchées plus importantes (60 km de large et jusqu'à 600 km de long). Il est par exemple difficile de séparer les différents types de cultures. Pour ces raisons, des capteurs tels que ceux présents sur SPOT 6 et 7, et Sentinel-2, ne peuvent être opposés, mais sont complémentaires.

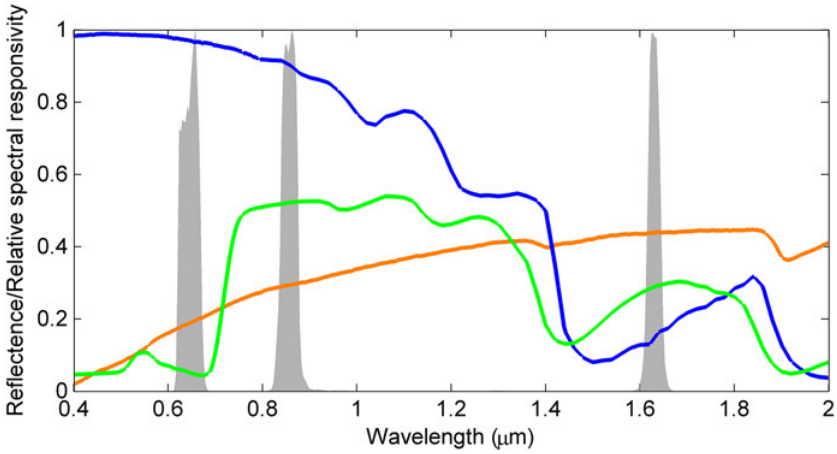


FIGURE I.6. Spectre de réflectance en fonction de la longueur d'onde des types de sol suivant : neige, sol nu, végétation. Source Wang et al., 2017

3 Automatisation de la cartographie d'occupation des sols

La disponibilité d'une variété très importante d'images satellites sur de très larges zones, et à des fréquences temporelles plus ou moins importantes, ouvre de nouvelles perspectives pour la description de ces zones en terme d'occupation des sols. Les délais d'acquisition étant très brefs, les capteurs toujours actifs et détaillés notamment en 2, rendraient possible la mise à jour de carte d'occupation des sols existantes, ou la génération de nouvelles cartes complètes.

3.1 Cartographie manuelle

Les processus de production de cartes d'occupation décrites dans la Section 1 sont principalement manuels. Si la photo-interprétation manuelle est naïvement la plus triviale à mettre en place d'un point de vue méthodologique, elle présente plusieurs inconvénients très limitants, toutefois mitigés grâce aux protocoles de qualification et de recette rigoureux auxquels sont soumis les produits finaux.
Tout d'abord, il apparaît évident que les approches manuelles sont très coûteuses et très longues, d'autant plus longues que les surfaces à couvrir sont importantes. A titre d'exemple, les efforts à mettre en œuvre pour générer la base de données d'occupation Corine Land Cover sont tels qu'un millésime n'est produit que tous les six ans depuis une vingtaine d'années, et la version calculée sur des données de 2012 n'a été diffusée qu'en 2015. Certains phénomènes lents peuvent être suivis avec une telle fréquence de mise à jour (accroissement des surfaces imperméabilisées), mais ceux-ci gagneraient de toute façon à bénéficier de versions plus rapprochées. En tout cas, beaucoup d'études ne se satisfont pas d'une telle fréquence (évolution des milieux naturels, suivi des cultures et de leurs rotations...) et nécessitent des mises à jour plus régulières.

Une limitation liée au temps de production est celle des données utilisées pour déterminer la classe d'occupation à laquelle appartient chaque objet. Un opérateur humain ne peut pas, dans les délais de production impartis, confronter différentes sources d'images (images très haute résolution spatiale, et images riches spectralement), ou bien suivre l'évolution dans le temps du comportement radiométrique de chaque objet à l'aide de séries temporelles. Ces séries temporelles peuvent en outre être sources de difficulté d'interprétation pour un opérateur humain, au même titre que l'imagerie radar. Cela peut générer des erreurs d'interprétation dues à un manque d'information discriminante.
Enfin, on peut citer une source de problème inhérente aux méthodes de production : diverses régions sont confiées à divers opérateurs humains, qui ont chacun leur perception des objets sur la carte, et leur expérience en matière de classification. Cela engendre une hétérogénéité spatiale qui peut également être gênante.

3.2 Classification supervisée

A l'inverse des méthodes manuelles, les procédures de production automatique permettent d'utiliser diverses sources de données (ou des séries temporelles) en des temps raisonnables, tout en garantissant une homogénéité du produit puisqu'un algorithme unique est utilisé pour traiter l'ensemble des données. Ces méthodes automatiques reposent sur le concept de *classification* qui consiste à attribuer à chaque pixel de l'image un label correspondant à l'une des classes de la nomenclature. Si la mise en place d'algorithmes de classification d'image ont été mis en place en télédétection depuis longtemps, la communauté de vision par ordinateur a été assez active également sur le sujet et a pu fournir des méthodes toujours plus performantes sous le terme de *machine learning*, englobant entre autre la classification automatique d'images. La notion d'apprentissage, ou de *learning*, est cruciale ici et se divise en deux familles d'algorithmes : l'apprentissage supervisé et l'apprentissage non supervisé. La différence principale entre les deux familles est la connaissance *a priori* (approche supervisée) ou non (approche non supervisée) des classes que l'on souhaite retrouver en fin de processus. En occupation des sols, la nomenclature a généralement été choisie au moment de l'établissement du cahier des charges de la carte souhaitée, donc en amont de la procédure de cartographie elle-même. La classification supervisée est détaillée ici.
Reposant sur la connaissance *a priori* des classes, la phase d'apprentissage nécessite un *jeu d'apprentissage* ou *jeu d'entraînement*. Celui-ci regroupe des échantillons dont la classe est connue. Ces échantillons peuvent être (i) des pixels ou (ii) des groupes de pixels. A partir de cette connaissance, la phase d'apprentissage permet au classifieur choisi de représenter chaque classe en fonction des échantillons lui appartenant (modèle génératif), ou en fonction des frontières entre classes (modèle discriminatif). Afin d'obtenir cet a priori sur les classes, on extrait souvent l'information d'occupation contenue dans

des bases de données géographiques (telles que celles mentionnées précédemment pour constituer l'OCS GE). La constitution du jeu d'apprentissage constitue l'étape la plus délicate d'un processus supervisé étant donné que l'on veut obtenir une représentation des classes qui soit la plus complète possible : si l'on considère la classe *bâti*, les instances de cette classe peuvent avoir des apparences très différentes les unes des autres, selon le type de bâtiment ou encore la région géographique. Par ailleurs, bien souvent, les bases de données employées pour construire ce jeu ne sont pas parfaitement cohérentes temporellement avec les images desquelles extraire ces échantillons. Mise à jour tardive, erreur de saisie (peu fréquente cela dit), imprécision géométrique sont des causes possibles d'erreur d'étiquetage dans le jeu d'apprentissage.
Parmi les produits d'occupation des sols existants, des cartes ont été calculées automatiquement sur la base de jeux d'apprentissage construits par photo-interprétation. L'effort est moindre par rapport à une solution complètement manuelle mais demande une minutie dans la conception des données de référence, et est d'autant plus ardu que la zone à classer est étendue spatialement. C'est le cas du produit FROM-GLC (Gong et al., 2013), offrant une occupation des sols sur le monde entier, à partir d'images Landsat acquises entre 1984 et 2011, pour une performance moyenne de 63,69% sur 9 classes d'occupation. Environ 90 000 échantillons ont été extraits sur l'ensemble de la Terre par photo-interprétation d'images Google Earth. Des experts de chaque pays ont contribué à la construction de ce jeu d'apprentissage, engendrant au total de lourds coûts en terme humain et financier. Toutefois, l'utilisation de séries temporelles aussi étendues dans le temps conduit à considérer des images représentant la même scène, mais avec des occupations qui ont souvent fortement varié sur la période considérée. L'entraînement des classifieurs est délicate à cause de la différence d'occupation d'une année à une autre, et il est difficile de millésimer le produit final car on ne sait pas déterminer à quelle époque correspondent les résultats (carte inhomogène temporellement). L'amélioration la plus récente en terme de performance de ce produit a été effectuée par Chen et al. (2015) qui introduit une analyse objet en plus de celle purement pixellique jusqu'alors utilisée. Des données MODIS ont été utilisées pour apporter l'information temporelle, réduisant la résolution réelle du produit final (plus grossière que la résolution initiale de 30 m de Landsat).
L'initiative portée par le pôle Theia mentionnée en Section 1.2 vise à produire par classification supervisée une carte chaque année en employant des séries temporelles Sentinel-2 denses (Inglada et al., 2017), mais avec des bornes plus rapprochées que pour le produit FROM-GLC, permettant une meilleure interprétation des classes présentes à une date donnée. Utilisant l'algorithme des forêts aléatoires, la méthode met en œuvre des règles de décision généralisable à d'autres régions que la France, tant que les échantillons d'apprentissage sont disponibles pour entraîner ces règles.

4 Travail de thèse

4.1 Problématique

La problématique générale de la thèse porte sur la cartographie à très haute résolution de l'occupation des sols sur de grandes zones géographiques, par classification automatique d'images satellites ou aérienne. Ces travaux s'ajoutent aux études répondant à un besoin exprimé qui a été détaillé en section 1 pour les utilisateurs de ce type de données. Ce besoin est lié à des contraintes de cohérence temporelle, fréquence de mise à jour, résolution des cartes, compatibilité, avec des OCS existantes.
La multiplication des capteurs imageurs spatiaux passés en revue en section 2 constitue une opportunité pour la cartographie et le suivi de phénomènes naturels ou anthropiques. Parmi ces capteurs, les capteurs optiques des satellites SPOT 6 et 7 acquièrent des images à très haute résolution spatiale sur quatre bandes spectrales, à partir desquelles l'IGN produit une couverture nationale annuelle. Peu exploitées dans un cadre de cartographie d'occupation des sols, ces images monoscopiques ont la bonne propriété de proposer une résolution spatiale à 1.5 m, décrivant ainsi le milieu urbain finement avec une séparation route / bâti par exemple. Quant aux milieux naturels à l'instar des forêts et des zones de culture, une telle résolution permet d'apporter une texture fine, et pour les parcs de tailles réduites en milieu urbain, la détection est également possible avec ces images.
Si SPOT 6 et 7 proposent des images qui peuvent être exploitées pour des tâches de suivi de phénomènes naturels ou anthropiques, l'utilisation d'orthophotos aériennes est plus adéquate pour mettre à jour des bases de données topographiques existantes qui ont bien souvent des spécifications précises en termes géométriques, sémantiques et surfaciques (unité minimale de collecte par exemple). De plus, le calcul d'orthophotographies s'accompagne depuis récemment du calcul du MNS du millésime correspondant à ces images. Or, nous le verrons, l'information de hauteur peut permettre de lever des ambiguïtés autrement non résolubles.
Du point de vue méthologique, l'effort est mis en entier sur l'utilisation de réseaux de neurones profonds. En effet, que l'on utilise des images satellites SPOT 6 et 7 ou des images aériennes, les résolutions étant très hautes, la variété d'une classe peut être très importante en raison du détail que l'on a sur les objets de cette classe. Or, si l'information spectrale permet de construire des attributs discriminants pour regrouper des objets appartenant à une même classe, le nombre de bandes spectrales des images issues d'acquisition SPOT ou aériennes est réduit. Les réseaux de neurones sont connus aujourd'hui pour leur grand pouvoir de généralisation, et l'IGN dispose de bases de données topographiques conséquentes. Le but est donc de mettre à l'épreuve ces réseaux dans un contexte de cartographie d'occupation des sols, sans une grande richesse spectrale, mais en reposant sur leur exploitation très efficace de la texture dans les images.

Toutefois, plusieurs verrous apparaissent relatif à l'automatisation de la

classification d'images très haute résolution spatiale sur des régions aussi variées que sont celles du territoire national français :

1. que l'on utilise les images SPOT ou aériennes, on ne dispose pas de séries temporelles, et les couvertures respectives issues de ces deux types d'images sont produites sur une année de données. On a donc, en plus de la diversité thématique inhérente à un territoire national, des diachronies pouvant mener à des erreurs de classification entre date d'acquisition, et ce, pour des objets appartenant à une même classe ;
2. pour cartographier efficacement le territoire national, on ne peut envisager un modèle à apprendre par région ou par date pour plusieurs raisons. Tout d'abord, il faudrait des jeux de données annotés cohérents avec chaque région ou date en question, ce qui rend cette approche impossible. Mais en plus, les réseaux de neurones peuvent être longs à entraîner, et multiplier les modèles induit automatiquement une multiplication conséquente des temps d'entraînement ;
3. on sort dans ces travaux d'aspects académiques visant à maximiser les performances d'un modèle sur un jeu de données. Les réseaux de neurones n'étant encore qu'une famille de méthodes récente, et leur utilisation au service de besoin à visée opérationnelle en matière d'occupation des sols étant inexistantes jusque là, il convient de placer ces modèles profonds dans divers scenarii pour avoir une idée de leur potentiel général ;
4. les classifications automatiques d'occupation des sols souffrent généralement d'un manque d'aspect cartographique justement, au sens des bases de données. En particulier, si l'on cherche à rendre une OCS automatisée compatible avec des bases de données topographiques, la méthode doit avoir été éprouvée sur certains postes de nomenclature existantes, avec un travail de fond sur les données d'apprentissage et les données images utilisées (nettoyage des bases de donénes, cohérence temporelle entre celles-ci, les images utilisées et le MNS le cas échéant).

Ces points sont parmi les plus importants et seront adressés dans l'ensemble de manuscrit.

4.2 Contributions

La suite du manuscrit détaille l'ensemble des travaux menés dans le cadre de cette thèse doctorale. Pour plus de transparence et de clarté, une liste des contributions principales est dressée ici :

— démocratisation de l'utilisation des réseaux de neurones convolutifs en télédétection spatiale ;

— preuve que les réseaux de neurones sont efficaces pour la cartographie de zones géographiques étendues mêlant des paysages variés (urbains, ruraux, forestiers, côtiers) ;

— réflexion sur les contraintes liées à l'axe opérationnel de la thèse quant à l'utilisation de ces réseaux de neurones ;

— établissement de scenarii réalistes pour la cartographie d'occupation des sols avec des approches d'apprentissage profond ;

— étude du caractère "adaptatif" des réseaux de neurones lorsqu'ils sont utilisés sur de nouvelles données, paysages ou nomenclatures ;

— utilisation de réseaux de neurones pour la classification d'images aériennes en réponse à un besoin réel.

4.3 Structure de la thèse

Le manuscrit est découpé en trois chapitres principaux qui font suite à cette section introductive.
Le chapitre II propose au lecteur une introduction générale à l'apprentissage automatique, avec un état des méthodes principales abondamment utilisées depuis plusieurs décennies. L'apprentissage profond étant une composante centrale dans ces travaux, on en propose une introduction tout d'abord par une approche historique, à laquelle succède une entrée en matière un peu plus théorique sur les parties principales constitutives d'un réseau de neurones profond.
Le chapitre suivant met en lumière la méthodologie suivie pour classifier des images satellites très haute résolution par le biais de ces réseaux de neurones. Une investigation particulière est menée dans un contexte de télédétection : on entend par là l'étude de l'impact de changement de temporalité, de géographie par exemple sur un classifieur. L'adaptation de domaine est un champ de recherche très intéressant et qui a le mérite de chercher à minimiser les efforts de collecte de données d'apprentissage qui sont coûteux à tout point de vue. Ce chapitre permet également d'étudier le comportement d'un modèle lorsque la zone géographique à classifier s'étend spatialement.
Si le chapitre précédent affiche des résultats prometteurs en matière de cartographie d'occupation des sols, et permet de faire un pas vers la cartographie opérationnelle, le chapitre IV s'attache davantage encore à placer ce manuscrit hors d'un cadre uniquement académique. Le choix d'images aériennes très résolues (sub-décimétriques) et l'ancrage des divers tests à la nomenclature OCS GE permettent de comparer en termes opérationnels les cartes obtenues avec les bases de données topographiques existantes, tout en remettant en question la quelité de ces dernières. On y confronte également plusieurs architectures de réseaux de neurones.
La conclusion de l'ensemble de ces travaux compile les résultats principaux et dresse les principales perspectives de recherche.

Chapitre II

État de l'art

1 Méthodes de classification en télédétection

Les cartes d'occupation des sols produites automatiquement sont généralement générées par des outils de classification (semi-) automatiques. Le principe de la classification automatique a déjà été décrit en Section 3.2 : il s'agit de fournir, pour une image, une information sémantique en chaque pixel. Les pixels de l'image sont alors regroupés en ensembles appelés classes. Ces classes contiennent en leur sein des pixels qui doivent partager des similarités visuelles présentes sur l'image, tout en séparant efficacement des pixels qui ne présenteraient pas de telles similarités. Les divers travaux conduits en *machine learning* permettent de proposer des algorithmes de classification, plus ou moins adaptés selon la tâche à exécuter.
Il est possible de classer ces algorithmes selon la connaissance que l'on a sur la donnée que l'on souhaite étiqueter en deux groupes, les approches non supervisées et les approches supervisées. Ces deux approches suivent toutes

deux un schéma d'apprentissage itératif, bien que différentes du point de vue de la conduite de l'apprentissage lui-même. Notons que des algorithmes de classification *semi-supervisés* existent aussi, tirant parti de données non étiquetées lorsque peu de données d'apprentissage étiquetées sont disponibles (Blum et Mitchell, 1998). Ces méthodes ont été utilisées pour la classification d'images optiques (Bruzzone et al., 2006a) et hyperspectrales (Bruzzone et al., 2006b).

1.1 Classification non supervisée

Les approches non supervisées ne disposent pas de connaissance a priori sur la nature des classes à détecter et sont donc dépourvues de tout jeu d'entraînement. L'attribution des classes est effectuée selon des critères de similarité (motifs récurrents entre objets d'une même classe par exemple), regroupant les éléments à classer (ici, les pixels d'une image) en *clusters*. Les *clusters* sont généralement constitués de façon à maximiser l'homogénéité *intra-cluster* et à maximiser la dissimilarité entre *clusters*. Les classes, au sens de la télédétection, sont ensuite attribuées a posteriori (Loveland et al., 2000). Le fait de ne pas nécessiter de jeu d'entraînement peut être approprié aux cas de classification d'occupation des sols pour lesquels peu de données de référence est disponible (ces données étiquetées sont souvent onéreuse, voire inexistante) (Eva et al., 2004), ou pour découper des classes d'un problème supervisé en sous-classes.
Deux catégories d'algorithmes de classification non supervisée existent : les méthodes dites de partitionnement et les méthodes probabilistes.

Les méthodes hiérarchiques sont un exemple des méthodes de partitionnement : elles fonctionnent soit (i) par agglomération de pixels (hiérarchique ascendante), soit (ii) par dissociation (hiérarchique descendante). Dans le premier cas, où l'on cherche à regrouper les pixels, comme dans le second, où l'on cherche à les dissocier, des critères ou métriques sont utilisés pour comparer les pixels. Dans le cas de regroupement de pixels, les méthodes associées fournissent un résultat pouvant être représenté sous forme de dendrogramme (arbre), à l'aide duquel il est possible de considérer diverses partitions des pixels selon la coupe choisie dans ce dendrogramme. Ces coupes définissent différents compromis entre similarité intra-classe et séparabilité entre classes, que l'on souhaite toutes deux maximiser.
Parmi les algorithmes de partitionnement, les k-moyennes sont l'un des plus connus et simples à mettre en place. Décrit en premier par Steinhaus (1956) puis introduit sous cette dénomination par MacQueen (1967), l'algorithme des k-moyennes permet de partitionner l'ensemble des pixels selon un processus itératif, en garantissant une meilleure partition à chaque nouvelle itération, et donc une convergence en un temps fini (le minimum global est loin d'être garanti cependant). L'objectif global de cet algorithme est de minimiser la somme des distances (norme euclidienne) entre les pixels et le centroïde du *cluster* auquel ils se rapportent, défini par la moyenne des points associés à ce *cluster*. Bien qu'offrant une méthode rapide de classification, et

une facile mise en œuvre, un inconvénient majeur de cet algorithme repose sur le besoin au préalable de connaître le nombre de classes à détecter, et la difficulté potentielle de donner une interprétation aux *clusters* ainsi formés. Par ailleurs, la partition finale dépend de l'initialisation préalable des centroïdes, conduisant souvent à tester différentes initialisations pour ne garder que celle qui produit le meilleur résultat. Il est possible de définir chaque *cluster* par des noyaux autres que la moyenne des points qui les constituent : le terme de *nuées dynamiques* en français fait référence à cette généralisation des k-moyennes. Ainsi, l'algorithme des k-médoïdes (Kaufman et Rousseeuw, 1987), par exemple, utilise un élément de chaque classe la représentant en lieu et place de la moyenne, permettant ainsi de produire de meilleurs résultats dans le cas de classes convexes.
Ball et Hall Dj (1965) ont introduit l'algorithme ISODATA (*Iterative Self-Organizing Data Analysis Technique*), qui se différencie des k-moyennes dont il est une variante car il ne nécessite pas un nombre prédéfini de classes en paramètre. L'algorithme repose sur deux seuils à paramétrer correspondant (i) au degré maximal d'hétérogénéité au sein d'une classe au-delà duquel on la sépare en deux classes, (ii) au degré minimal de dissociation entre deux classes en-deçà duquel on les fusionne. L'inconvénient réside dans la sensibilité du résultat vis-à-vis de ces deux seuils, pouvant facilement conduire à une seule classe au terme du processus.

La dernière famille de méthode d'apprentissage non supervisée, les méthodes probabilistes, a pour objectif l'estimation de densités de distribution des classes au sein du jeu d'apprentissage. A l'inverse des méthodes précédentes, les considérations géométriques ne sont plus prises en compte pour répartir les échantillons entre les classes données, mais on s'appuie sur l'analyse de la distribution de ces éléments. Les éléments appartenant à la même classe sont donc issues de la même distribution. Un intérêt de ces méthodes sur les k-moyennes, par exemple, est de fournir une probabilité d'appartenance à chacune des classes, offrant ainsi la possibilité de contrôler les échantillons dont la classe attribuée semble incorrecte. Chacune des classes étant représentée par une distribution, on parle de modèle de mélange de distributions (somme des distributions pondérées par le nombre d'échantillons affectés à chaque classe). Les distributions de ces échantillons à classer peuvent être représentées par des fonctions statistiques telles que des gaussiennes donc les paramètres sont estimés par EM (*Expectation–maximization*) pour trouver le maximum de vraisemblance.

Dans le contexte de la télédétection spatiale, l'approche non supervisée a été utilisée à des fins de cartographie, notamment par analyse multi-résolution, en milieu urbain (Kurtz et al., 2010 ; Zhang et Kerekes, 2011) ou sur des zones étendues (notamment si l'on n'a pas de nomenclature *a priori* ou pas assez d'échantillons d'apprentissage pour entraîner un classifieur supervisé) présentant des paysages plus variés (Sublime et al., 2017). Le dernier article s'intéresse à la classification d'images très résolues (Pléiades - 0.5cm). L'absence

de connaissance a priori des classes à trouver couplée à la faible information sémantique contenue dans un seul pixel d'une image superspectrale très résolue, rend la classification pixellaire très difficile. C'est pourquoi les auteurs calculent, en amont du processus de classification, une segmentation de l'image, pour une analyse au niveau objet, plutôt qu'à l'échelle du pixel. Dans tous les cas, les différents travaux sont confrontés à l'écueil de la correspondance entre les clusters découverts par les divers algorithmes et les nomenclatures d'occupation du sol ciblées. En particulier, associer un cluster à une classe d'occupation du sol nécessite une connaissance experte (i) de la nomenclature (ii) de la zone d'étude (type de culture, essences forestières en présence, etc). Par ailleurs, le décalage temporel entre dates d'acquisition peut entraîner la présence d'objets d'une même classe d'occupation dans différents clusters.
Ayant une connaissance préalable des classes voulues ainsi que des données de référence pour la classification d'occupation des sols, l'approche supervisée, décrite dans la section suivante, est favorisée.

1.2 Classification supervisée

Les algorithmes de classification supervisée permettent d'intégrer une connaissance préalable sur les classes à retrouver, tant au niveau du nombre de ces classes, que de leur représentation. Là où le processus itératif non supervisé a pour vocation de regrouper les objets similaires selon un critère défini, les approches supervisées mettent en jeu un *apprentissage* de l'algorithme choisi, qui définit des règles et modèles capables de répartir à partir des observations qui leur sont associées les éléments d'un jeu de données, labellisés selon la nomenclature choisie ; le classifieur doit alors minimiser l'erreur *empirique*, calculée sur l'ensemble du jeu d'apprentissage. Le modèle est ensuite utilisé pour inférer la classe d'occupation de nouveaux éléments à partir des règles apprises durant la phase d'apprentissage. La qualité du classifieur se quantifie via l'erreur *réelle*, ou capacité à bien généraliser, qui mesure l'erreur de classification pour de nouveaux échantillons qui ne font pas partie du jeu d'apprentissage. Cette bonne généralisation est complètement tributaire de la qualité du jeu d'apprentissage utilisé pour entraîner le modèle choisi. En l'absence de base de données représentatives, la constitution d'un tel jeu de données est coûteux en temps et doit être réalisé en gardant à l'esprit qu'il faut prendre en compte la diversité des objets au sein d'une classe.

Echantillons d'apprentissage

La constitution du jeu d'apprentissage, comme dit précédemment, est le principal facteur d'influence sur la qualité du modèle de classification, indépendamment du modèle lui-même. Pour une classe donnée, le contenu des objets choisis pour représenter cette classe au sein du jeu d'apprentissage doit être varié afin que le classifieur prenne justement en compte cette variété. Omettre des échantillons de *bâti* représentatif de cette classe dans la moitié Nord de la France dans le jeu d'apprentissage empêcherait sans doute

le classifieur de retrouver, lors de la phase d'inférence, ce type de bâti sur l'image. On peut appliquer ce raisonnement au paysage lui-même : le milieu urbain en agglomération dense offre un paysage tout à fait différent d'habitations dans des communes plus rurales et beaucoup moins dense. L'exhaustivité des objets à retrouver est donc un facteur clé de succès pour obtenir un modèle qui généralise bien.

Par ailleurs, bien dimensionner le jeu d'apprentissage est crucial : bien qu'une taille de jeu de référence accrue améliore systématiquement la qualité du classifieur (Huang et al., 2002 ; Foody et Mathur, 2004) un modèle entraîné à partir d'un jeu au contraire trop réduit peut conduire à un sur-apprentissage du modèle. Ce phénomène est dû à une taille de jeu d'apprentissage trop faible au regard du nombre de paramètres à estimer au sein du modèle. Le sur-apprentissage survient lorsque les paramètres du modèle, en trop grand nombre, "mémorisent" parfaitement les échantillons d'apprentissage (compromis biais / variance). Au lieu de modéliser les caractéristiques communes entre chaque échantillon d'une même classe, le jeu de paramètres modélise les caractéristiques propres à *chaque* échantillon, condamnant le modèle à ne pas généraliser sur de nouveaux échantillons extérieurs au jeu d'apprentissage.
En théorie, multiplier le nombre d'échantillons pour éviter ce phénomène est donc la démarche à adopter. Toutefois, le temps et les moyens humains nécessaires pour annoter des données de référence est un facteur très limitant. Il s'agit donc de trouver le bon équilibre entre complexité du modèle (afin de donner suffisamment de liberté à celui-ci pour modéliser des phénomènes eux-aussi complexes) et taille du jeu d'apprentissage. Enfin, il reste à considérer les attributs de chaque échantillon, utilisés pour la classification : si une augmentation du nombre d'attributs permet instinctivement d'améliorer les résultats, la pratique est différente. Le phénomène de *malédiction de la dimension* réduit les performances de classification lorsqu'un trop grand nombre d'attributs est utilisé. Il a été mis en évidence par Hughes (2006) (on l'appelle également "phénomène de Hughes") : la plupart des algorithmes de classification nécessitent le calcul de distances, qui est d'autant plus difficile à mener que l'espace des attributs est grand. Si le nombre d'échantillons n est trop faible par rapport au nombre d'attributs p qui les décrivent, le classifieur peinera à regrouper les échantillons d'une même classe qui seront isolés dans l'espace des attributs. L'état de l'art suggère un nombre $n = 30p$ d'échantillons par classe (Foody et Mathur, 2004 ; Mather et Koch, 2011). Les classifieurs modernes résistent toutefois mieux à ce phénomène de Hughes, nous le verrons avec les SVMs dans la suite.

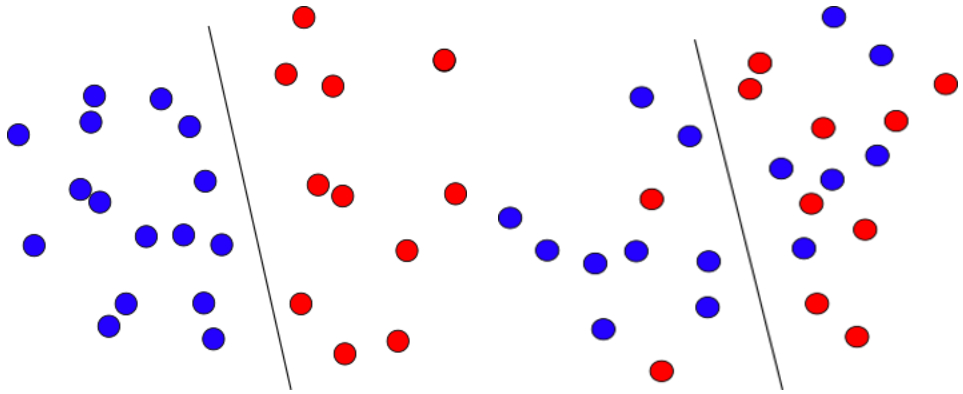
Algorithmes de classification

La télédétection automatique de classes d'occupation des sols a été expérimentée à l'aide de beaucoup de méthodes de classification supervisée. Ces méthodes se regroupent sous deux principales familles. Les premières approches établissent des hypothèses quand à la nature des distributions des

classes au sein du jeu d'apprentissage : ce sont les méthodes paramétriques (ou probabilistes). Chaque classe est modélisée par une distribution dont les paramètres sont estimés à partir des échantillons d'apprentissage. L'estimation est souvent effectuée par maximum de vraisemblance. Maas et al. (2018), dans le contexte de la détection du changement, prennent en compte l'obsolescence potentielle de certaines des données d'entraînement qui ont pu changer entre deux dates, en affectant à chaque échantillon un poids et une transition issus d'un modèle probabiliste gaussien. Dans le cas de la cartographie d'occupation des sols, les modèles paramétriques ne sont pas aussi efficaces que les modèles non paramétriques sur des données images (Foody, 2002). Les modèles paramétriques peinent en effet à capturer la diversité d'apparence dans les classes d'occupation ; cet effet est accentué avec des données très haute résolution.
Les modèles non paramétriques se définissent par une absence d'hypothèse sur la distribution des classes et sont particulièrement efficaces sur les données spectrales. Plusieurs algorithmes existent, plus ou moins élaborés : les *k plus proches voisins* (Altman, 1992), les arbres de décisions sont des algorithmes simplistes mais dont la mise en œuvre a le mérite d'être aisée. En télédétection, deux types de méthodes en particulier sont quasi systématiquement utilisées : les *Support-Vector Machines* (SVMs), introduits par Cortes et Vapnik (1995), et les Forêts Aléatoires (*Random Forests* (Breiman, 2001)). Les deux algorithmes sont détaillés dans les paragraphes suivants.

SVMs

Abondamment utilisé en télédétection (Mountrakis et al., 2011), les SVMs ont été employés aussi bien dans le cadre de classification de séries temporelles (Waldner et al., 2015), MODIS (Jia et al., 2014) ou Sentinel-2 (Hawryło et al., 2018) par exemple, que dans le cadre de cartographie d'occupation des sols à partir d'une seule image satellite (Saini et Ghosh, 2018). Des travaux portant sur l'estimation de données biophysiques s'appuient également sur une classification effectuée par SVMs, c'est le cas des travaux de Knudby et al. (2010) portant sur l'analyse de récifs coraux à Zanzibar et la corrélation de données structurelles sur ces coraux à la diversité de la faune sous-marine environnante, à partir d'images IKONOS.
Les SVMs sont appréciés du fait de l'impact très limité du phénomène de Hughes et du faible besoin de données d'apprentissage pour obtenir des résultats satisfaisants Camps-Valls et al. (2004) tout en conservant un grand potentiel de généralisation. Toutefois, les échantillons d'apprentissage doivent être bien choisis et permettre de définir clairement les frontières entre classes, les SVMs optimisant la frontière entres classes.
Initialement, les SVMs ont été pensés dans le cas où (i) seules deux classes sont présentes et (ii) lorsque ces deux classes sont linéairement séparables. La figure II.1 montre le cas où deux ensembles d'échantillons peuvent être séparés linéairement. Cependant, la plupart des cas présentent, à l'instar du schéma de droite, des classes *a priori* plus complexes à retrouver car non linéairement séparables.



(A) Cas linéaire (B) Cas non linéaire.

FIGURE II.1. Illustration de la tâche à résoudre : la droite noire au milieu doit au mieux discriminer les deux ensembles d'échantillons rouges et bleus.

Les SVMs se généralisent évidemment au-delà de la dimension 2 illustrée sur la figure précédente. Dans le cas général, on cherche donc un hyperplan séparant **au mieux** les deux ensembles. L'expression "au mieux" résume l'objectif que les SVMs : pour comprendre cela, on introduit la notion de *marge* qui se rapporte à la distance minimale d'un échantillon à l'hyperplan. L'objectif est ainsi exprimé : trouver l'hyperplan qui maximise la marge. Si les SVMs résistent bien au phénomène de Hughes, c'est grâce à l'introduction d'une « tolérance » d'erreur, généralement notée C. La Figure II.2 représente la situation rencontrée par les SVMs. Les échantillons sur les droites en pointillés sont les vecteurs supports, c'est-à-dire les échantillons les plus proches de l'hyperplan et donc des deux classes : la droite en noire constitue la frontière délimitant le mieux les deux hyperplans précédemment calculés. Formellement, si x est notre vecteur de n observations, dont les échantillons sont répartis dans les classes $\{-1, 1\}$, on veut trouver les paramètres de l'hyperplan $(\boldsymbol{w}, b)$ tels que :

$$\forall x_i,\ i \in \{1...n\},\ y_i(w \cdot x_i + b) > 1 \tag{II.1}$$

La distance géométrique entre les deux hyperplans "support" étant de $\frac{2}{\|w\|}$, maximiser cette distance revient à minimiser $\frac{1}{2}w^t w = \frac{1}{2}\|w\|^2$ sous la contrainte précédente. La forme étant quadratique, le problème est convexe, garantissant donc l'existence d'un minimum global.

Evidemment, le cas dans lequel nous nous sommes placés est uniquement binaire et considère que les classes sont linéairement séparables, ce qui n'est pas un scénario réaliste en télédétection compte tenu de la variété spectrale intra-classe et de la richesse des nomenclatures étudiées. Revenons dans un premier temps à la Figure II.1b, illustrant un problème dont les données ne sont pas linéairement séparables. Tel que posée dans l'Equation II.1, la question ne peut être résolue directement de façon à proposer un classifieur

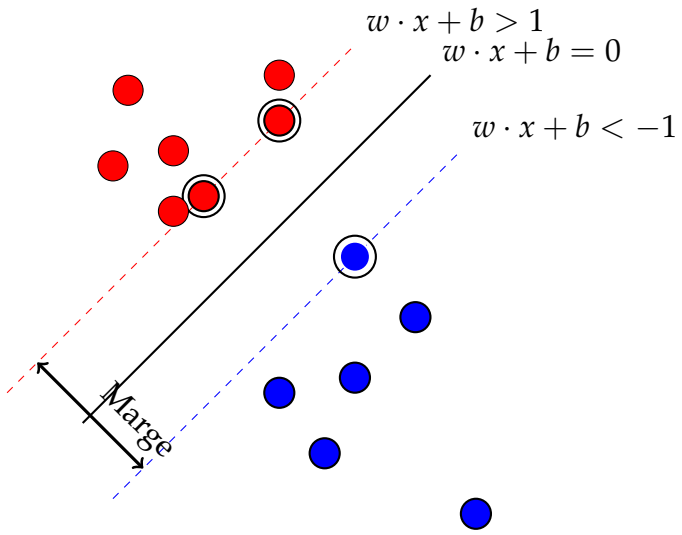


FIGURE II.2. Représentation de l'objectif recherché par les SVMs : les *vecteurs supports* sont les échantillons vérifiant $w \cdot x + b = 1$ et $w \cdot x + b = -1$ (droites représentatives en pointillés). C'est pour cela qu'il est important d'avoir des échantillons bien définis à la frontière des classes.

efficace. Afin de pallier cet obstacle, l'utilisation de fonctions noyaux (Cristianini, Shawe-Taylor et al., 2000 ; Müller et al., 2001) permet, sans en calculer explicitement la transformation, de projeter les données dans un espace d'attributs de dimension plus élevée, facilitant la séparation entre classes. La figure II.3 représente le passage d'un espace de dimension 2 à un espace de dimension 3 dans lequel la séparation par un hyperplan entre les deux classes est rendue possible. Sans détailler davantage la démarche non linéaire, notons que la fonction noyau plus utilisée est le noyau gaussien *Radial Basis Function* - ou RBF.

Pour conclure sur cette partie consacrée aux SVMs, discutons du passage au cas multi-classes. Deux approches sont possibles :

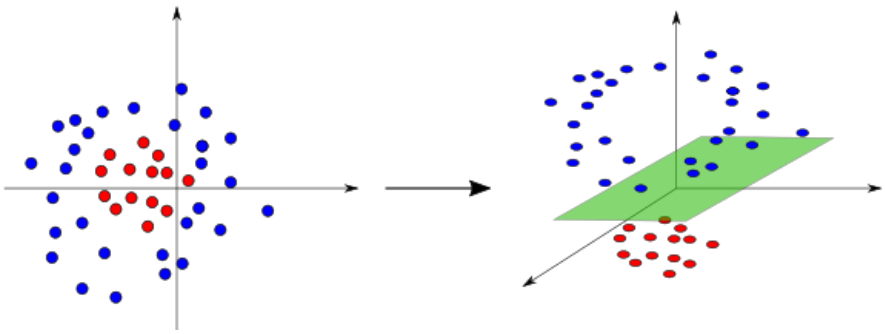


FIGURE II.3. Le *kernel trick* est une méthode permettant à un classifieur linéaire d'identifier, au moyen d'une projection dans un espace d'attributs de plus grande dimension, des classes initialement non séparables linéairement.

1. les classes sont toutes confrontées par paire : il y a donc autant de classifieurs SVMs que de paires de classes possibles. Cette approche est nommée *"one-versus-one"*.
2. chaque classe est confrontée à l'ensemble des autres, alors rassemblées en une seule classe *autre*. On parle d'algorithme *"one-versus-all"*.

Ces deux stratégies ont été comparées dans le cadre de plusieurs travaux (Hsu et Lin, 2002 ; Melgani et Bruzzone, 2004). Si les deux aboutissent à des résultats sensiblement similaires, la première est préconisée, même si elle suggère le calcul de davantage de classifieurs binaires, car ceux-ci convergent plus rapidement vers une solution optimale (il est plus facile de séparer deux classes bien définies, qu'une classe et un regroupement des autres classes, pouvant être différentes entre elles).

Forêts Aléatoires

Méthode ensembliste, l'algorithme des Forêts Aléatoires applique le principe de *bagging* (Breiman, 1996) aux arbres de décisions. Ces derniers sont, pris séparément, des classifieurs dits *faibles*. En effet, ils sont très sensibles au sur-apprentissage et donc sujets à une faible capacité de généralisation (sensibilité accrue au bruit de mesure). Toutefois, les arbres de décision ont un certain nombre d'avantages qui les rend attrayants. Leur simplicité les rend très faciles à interpréter, les résultats de règles de décisions pouvant très souvent être réduits à *Vrai* ou *Faux*. Aussi, par construction, une sélection des attributs les plus pertinents parmi tous ceux disponibles est également effectuée. Un avantage très précieux est leur simplicité algorithmique leur permettant d'analyser de très gros volumes de données avec des temps de traitement réduits ainsi qu'un besoin matériel raisonnable par rapport à d'autres algorithmes. Le *bagging* agrège les résultats de classification issus d'un certain nombre d'arbres (d'où le terme de *forêts aléatoires*) pour obtenir un classifieur *fort* ou *robuste* à partir de classifieurs *faibles*.
Le mécanisme de fonctionnement des Forêts Aléatoires consiste à :

1. constituer N sous-ensembles d'échantillons *bootstrap* (= tirage avec remise) issus du jeu d'apprentissage initial ;
2. calculer un arbre décisionnel pour chacun de ces sous-ensembles : chaque arbre est binaire et est constitué de nœuds, qui sont les tests auxquels sont soumis les variables d'apprentissage. Une spécificité des Forêts Aléatoires est de ne sélectionner aléatoirement qu'une partie des attributs disponibles en chacun de ces nœuds. Ce tirage est cette fois-ci sans remise. En cela, les Forêts Aléatoires vont plus loin que des méthodes de *bagging* classiques.
3. décider d'une classe finale par une règle, à définir, de vote majoritaire sur l'ensemble des arbres pour chaque échantillon.

Construire ainsi l'algorithme permet d'établir deux métriques intéressantes, la première évocatrice directe de la capacité de généralisation de chaque arbre est l'écart ou erreur *Out-Of-Bag (OOB)* (terme provenant du *bagging*

effectué à l'étape 1). Ce dernier est calculé simplement en mesurant l'erreur de chaque arbre sur les échantillons non utilisés pour leur apprentissage, erreur ensuite moyennée sur la forêt. La seconde métrique est l'importance de chaque attribut : en effet, en échangeant les valeurs d'un attribut donné entre plusieurs échantillons, et en re-calculant l'erreur OOB après perturbation du jeu d'entraînement, il est possible d'apprécier l'influence de cet attribut sur la décision finale de la classe. Cette dernière est une métrique bienvenue dans un problème où le nombre d'attributs est très important et nécessite une sélection des plus intéressants et des plus déterminants (même si les attributs très redondants peuvent poser problème).

En télédétection, les Forêts Aléatoires sont très abondamment utilisées du fait de leur propension à choisir les meilleurs attributs parmi tous ceux disponibles pour discriminer les classes entre elles, de donner un score de confiance dans la classe choisie, et pour leur implémentation logicielle hautement parallélisable, avec des performances presque aussi bonnes que les SVMs. Récemment, ils ont été utilisés pour la cartographie de cultures à l'aide d'images Sentinel-2 monodates (Immitzer et al., 2016) comme de séries temporelles (Inglada et al., 2017) pour la production de cartes d'occupation des sols à l'échelle nationale. L'utilisation de cet algorithme, intrinsèquement multi-modal, ne se limite aux données spectrales mais aussi au LiDAR (Chehata et al., 2009) dans le cadre de l'étude du milieu urbain, et plus spécifiquement du choix des attributs dérivés du LiDAR les plus pertinents pour une telle étude, avec l'utilisation de l'erreur *OOB* et de la mesure d'importance de variable.

2 Apprentissage profond et occupation des sols

2.1 Processus classique de classification de l'OCS

L'importance cruciale de la connaissance de la couverture biophysique (occupation) des territoires a conduit à l'utilisation intensive des méthodes décrites précédemment. La thématique très spécifique de l'occupation des sols nécessite une connaissance accrue des données utilisées et des classes manipulées de la part d'un opérateur cherchant à mettre en place une chaîne de traitement de classification automatique pour cette tâche. Cette expertise requiert des spécialistes (opérateurs, restituteurs...) formés spécialement pour cette tâche afin, qu'en fonction d'une nomenclature donnée, ils puissent dégager un certain nombre d'attributs discriminant les différentes classes à partir des données observées. Ainsi, dans le cas où ces données sont issues des capteurs embarqués sur les satellites SPOT 6 et 7, les observations sont les images quatre bandes - rouge, vert, bleu, proche infra-rouge. En l'état, celles-ci auraient bien du mal à différencier l'ensemble des classes d'occupation présentes dans les bases de données géographiques. Dans le cadre d'un processus classique de classification, s'ensuivent donc des travaux méticuleux et nécessaires d'extraction d'information pertinentes pour séparer les classes, à partir de ces images. Les travaux mettant en jeu ce type d'approche

sont multiples et aussi variés que les cas d'études traités.
Les attributs dérivés des bandes spectrales peuvent être catégorisés en deux familles : spectraux, et de texture. Les premiers correspondent à des combinaisons de canaux spectraux, le plus connu étant le *Normalized Difference Vegetation Index* (*NDVI*) (Rouse Jr et al., 1974) :

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

avec PIR, le canal proche infra-rouge, et R le canal rouge. Cet indice est, comme son nom l'indique, un indicateur quant à la présence de végétation ou non. Les ondes du domaine infra-rouges et rouge étant sensibles à la chrolorophylle (réflectance élevée des ondes infra-rouge, absorption importante des ondes rouges) , un indice élevé révèle vraisemblablement la présence de végétation, tandis qu'une valeur faible est caractéristique de végétation malade, éparse ou même absente. L'exemple de la végétation peut être étendu à d'autres objets présents sur les images et sensibles à un ou plusieurs domaines du spectre électromagnétique : il convient donc d'avoir une connaissance préalable des différents types d'objets à détecter ainsi que la meilleur combinaison des longueurs d'onde disponibles dans l'image pour les mettre en évidence. Les zones artificielles ou anthropisées ont été étudiées, entre autres, en utilisant l'indice de brillance (Bannari et al., 1996) qui dissocie surfaces minérales et couverture de végétation :

$$IB = \sqrt{R \times R + PIR \times PIR}.$$

Il est important de noter que les canaux infra-rouges sont très sensibles à l'humidité et que les caractéristiques du sol au moment de l'acquisition sont à considérer (de même que la présence d'ombre, fréquente en milieu urbain, peut mettre en défaut ce type d'indicateur). Les attributs de textures sont calculés en considérant le voisinage autour de chaque pixel de l'image : même s'il n'existe pas de définition universelle concernant la texture (Tuceryan et Jain, 1993), l'effet attendu est la caractérisation de la présence de motifs récurrents dans une image et / ou de sa régularité (aspect homogène, ou rugosité). La texture d'une image peut être considérée (i) à l'échelle globale, auquel cas la texture est la répétition d'une primitive (région de l'image dont les pixels partagent des propriétés semblables : on parle aussi de texton (Julesz, 1981)) à travers une partie de la scène, ou (ii) à l'échelle locale, ou la texture correspond à la distribution locale au voisinage d'un pixel des niveaux de gris (approche probabiliste). L'analyse de la texture est intéressante dans la plupart des domaines car les images présentent souvent une structure facilement interprétable pour l'œil humain. Toutefois, si la texture d'une image est un facteur discriminant évident pour l'œil humain, cette notion est difficile à traduire efficacement sous forma de mesure. Ainsi, même sans présenter de répartition strictement régulière d'une primitive, on peut retrouver sur une images des orientations dominantes de texture (Haralick, Shanmugam et al., 1973).
L'inconvénient de ce besoin d'attributs pour entraîner des algorithmes tels

que SVMs ou Forêts Aléatoires, en plus de nécessiter une connaissance experte des objets à détecter et des données manipulées, est dans la définition "figée" de ces attributs. En outre, régis par des formules indépendantes des données manipulées, ces représentations des données initiales peuvent nécessiter un travail conséquent d'ingénierie afin d'être adapté au problème considéré (identifier les bons types d'indices et le bon paramétrage).

Un fort renouvellement de l'état de l'art a eu lieu en apprentissage automatique avec l'émergence de l'apprentissage profond, qui a permis de lever le verrou précédemment évoqué quant à la détermination d'indices spectraux pertinents, et de textures efficaces. Les réseaux de neurones profonds, en particulier, se sont rapidement répandus parmi les différentes communautés friandes d'outils d'apprentissage, grâce à une multiplication considérable des données numériques qui a eu lieu, et d'évolutions techniques en matières de parallélisation des calculs. Cette ère du "Big Data" a permis de lever le verrou principal quand on en vient à parler d'apprentissage profond ou *deep learning*, qui est le besoin très massif en données d'apprentissage.

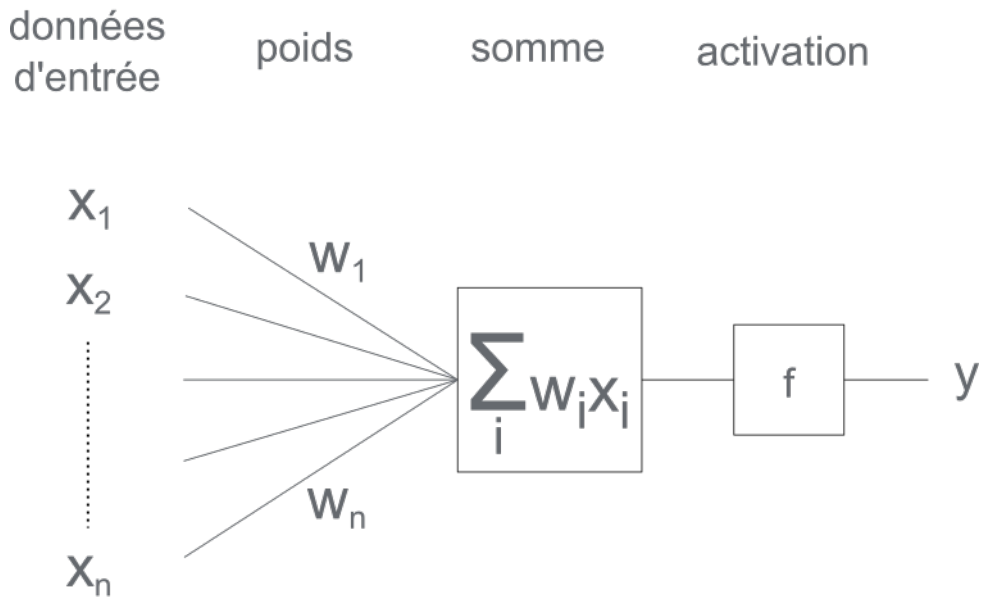


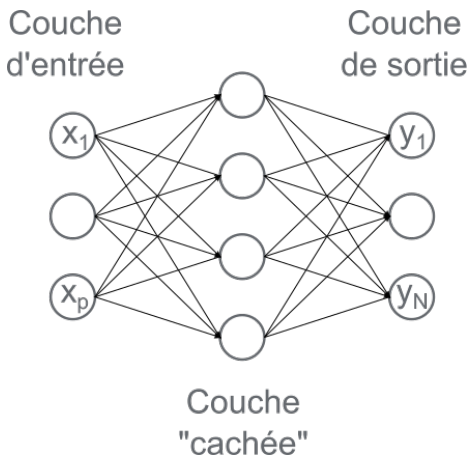
FIGURE II.4. Schéma du neurone formel construit par McCulloch et Pitts, aussi appelé *perceptron*.

2.2 Généralités et essor de l'apprentissage profond

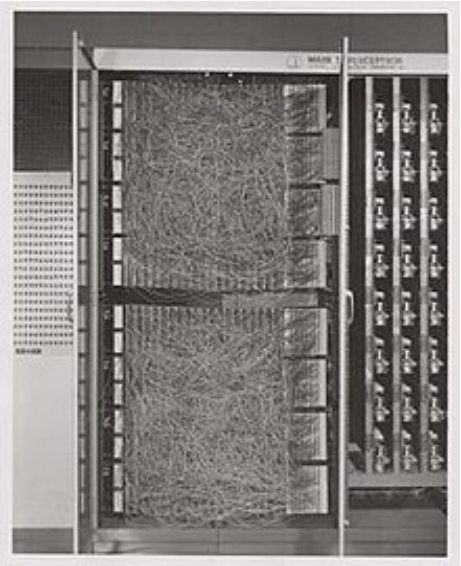
Une description de l'évolution historique de l'apprentissage profond, de ses prémices en 1943 à ce qu'il est devenu aujourd'hui, est proposée dans ce paragraphe. Notons ici que l'apprentissage profond transparaît principalement au travers d'un vecteur, d'une grande famille d'algorithmes : les réseaux de neurones profonds (réseaux convolutifs ou non, réseaux de croyances profondes, etc). Cette frise chronologique est suivie d'une section consacrée aux mécanismes plus théoriques d'apprentissage sur lesquels s'est progressivement bâti cet ensemble de méthodes d'apprentissage automatique.

Evolution historique en apprentissage profond

Deux chercheurs américains en neurologie et psychologie cognitive, Warren McCulloch et Walter Pitts, rangent l'activité du système neuronal dans la catégorie "all-or-none", le réduisant à une succession d'enchaînements logiques (au sens booléen du terme). Ils construisent ainsi en 1943 (McCulloch et Pitts, 1943) une formulation mathématique du neurone, représentée sur la Figure II.4, qui agit comme un opérateur transformant un ensemble d'états d'entrée en un état de sortie actif ou non ; état de sortie qui constitue lui-même par la suite l'un des états d'entrée d'un neurone qui succède au premier.
Entre temps, en 1950, Alan Turing interroge dans son ouvrage (Turing, 1950) sur la capacité des machines à penser, et donc sur d'éventuelles capacités neurologiques ou cognitives que celles-ci pourraient avoir. Cette interrogation sur une possible « intelligence artificielle » est la première en la matière. Toutefois, au lieu de répondre à la question « *Can machines think ?* », il la remplace très vite par ce qu'on traduit aujourd'hui comme le Test de Turing ou *Imitation Game*, titre de son introduction. Ainsi, plutôt que de s'intéresser directement à la pensée supposée de la machine, celle-ci peut-elle se faire passer pour un être pensant et tromper un être humain qui n'a pas conscience d'être en interaction avec une machine ? Turing n'apportera toutefois aucune solution.



(A) Perceptron multicouches imaginé par Rosenblatt en 1958.



(B) Le perceptron multicouches prend vie : le « Mark I Perceptron ».

FIGURE II.5. Le perceptron multicouches : premier réseau de neurones artificiels.

En 1958, Frank Rosenblatt, psychologue lui aussi, construit sur la base du

neurone formalisé par McCulloch et Pitts, le perceptron multicouches (Rosenblatt, 1958). Premier réseau de neurones artificiels (Figure II.5a), il empile un certain nombre de neurones élémentaires regroupés dans une seule couche, tous reliés à l'ensemble des entrées, et à la sortie. Cette architecture est le fruit de l'unification des domaines de la recherche en intelligence artificielle d'une part, initiée par Alan Turing, et de la neurologie cognitive, initiée par McCulloch et Pitts d'autre part, puisqu'une implémentation logicielle va en être faite pour l'IBM 704 dans un premier temps, à laquelle succède une machine conçue spécialement pour la reconnaissance d'image, baptisée « Mark I Perceptron », visible sur la Figure II.5b. La figure Figure II.5a indique le caractère *feed-forward* (ou propagation direct / avant) du perceptron : le signal d'entrée progresse uniquement dans le sens qui conduit à la sortie. Cette architecture est toutefois sujette à un défaut majeur : sa structure ne lui permet intrinsèquement de ne résoudre que des problèmes dont la solution est linéaire. C'est ce que démontrent Minsky et Papert dans Minsky et Papert (1969), notamment en se penchant sur l'impossibilité pour le perceptron d'implémenter la fonction logique XOR. Les travaux mis en avant dans cet ouvrage ont contribué à la mise en veille de la recherche sur les réseaux de neurones artificiels jusqu'au début des années 1980, la communauté impliquée s'étant reposée sur les espoirs que des réseaux simples à une couche intermédiaires comme les perceptrons suffiraient pour approcher n'importe quelle fonction, linéaire ou non. En effet, les auteurs indiquent qu'il est nécessaire d'utiliser plusieurs couches de neurones (*multilayer perceptron*) pour apprendre des fonctiones non linéaires, mais qu'aucun moyen d'entraîner ces réseaux n'existent, condamnant ceux-ci pour deux décennies au regard de la communauté. Paul Werbos s'y intéresse tout de même en 1976 dans sa thèse (Werbos, 1974), en construisant pour la première fois l'algorithme de **rétropropagation** (*backpropagation*), fondé sur la *chain rule*, et en le proposant comme procédé d'entraînement des réseaux de neurones artificiels. Toutefois, selon ses mots dans (Werbos, 2006) : « *Minsky was merely summarizing the experience of hundreds of sincere researchers who had tried to find good ways to train MLPs (multilayer perceptron), to no avail* », autrement dit, malgré les efforts de beaucoup de chercheurs, aucun n'a pu résoudre le problème d'entraînement des réseaux multicouches, engendrant un réel désintérêt de ce problème. Les travaux de recherche de Paul Werbos restent alors méconnus.
Douze ans plus tard, David Rumelhart, Geoffrey Hinton et Ronald Williams parviennent finalement à convaincre de la puissance des réseaux multicouches pour des problèmes d'apprentissage lors de la publication de Rumelhart et al. (1986), en redécouvrant et en formulant explicitement l'algorithme de rétropropagation du gradient dans le cadre des réseaux de neurones artificiels (sans toutefois citer les travaux de Paul Werbos). Le théorème d'approximation universel stipulant sommairement que toute fonction peut être approchée par un perceptron multicouches (et donc par n'importe quelle architecture à propagation avant, dotée d'au moins une couche intermédiaire) est généralisé en 1991 (Hornik, 1991). Pour satisfaire ce théorème, le perceptron multicouches doit posséder des fonctions d'activation non linéaires, dont certains exemples seront passés en revue un peu plus loin dans ce chapitre.

Quelques années auparavant, le *Neocognitron* est publié par Kunihiko Fukushima (Fukushima, 1980) dans le cadre de la reconnaissance de chiffres digitaux plus ou moins déformés sur des images. Fukushima construit son modèle en prenant en compte les travaux de Hubel et Wiesel (1959), qui établissent une **hiérarchie** dans l'organisation des réseaux de neurones : ainsi les neurones appartenant aux premières couches (dénommés *S-cells*) reflètent des caractéristiques simples des objets décrits dans les images, tandis que les neurones situés dans des couches plus profondes (les *C-Cells*) correspondent à des caractéristiques complexes, complexité qui s'accroît avec la profondeur. Fukushima évoque également la notion de *receptive field*, se rapportant à la quantité d'information (le voisinage de pixels dans le cas d'une image) à laquelle a accès un neurone dans une couche donnée (tous les neurones d'une même couche perçoivent le même *receptive field*). Ce champ d'information augmente également avec la profondeur du réseau, les neurones des premières couches propageant l'information de différentes portions d'une image aux neurones de couches ultérieures. Clairement adapté pour le traitement d'images, le *Neocognitron* a fortement inspiré l'architecture imaginée par LeCun et al. (1989) : en effet, bien que l'architecture soit construite sur la base d'un perceptron multicouche, une couche d'un nouveau genre apparaît en entrée du réseau, une couche dite **convolutive**. Cette nouvelle structure matérialise véritablement la notion de *receptive field* décrit par Fukushima auparavant, puisque les neurones composant cette couche convolutive sont en réalité des filtres de convolution, dont les paramètres sont déterminés par la méthode de rétropropagation. Les neurones de cette couche ont un nombre réduit de paramètres (les coefficients du filtre), et sont les mêmes à travers l'ensemble des voisinages (de la taille du filtre) de pixels de l'image (on parle de *weight sharing*), à l'inverse d'une couche type perceptron qui connecte le neurone à l'ensemble des pixels de l'image, multipliant démesurément le nombre de paramètres. Cette propriété de *weight sharing* impose une contrainte forte sur le problème tout en faisant intervenir un nombre réduit de paramètres.

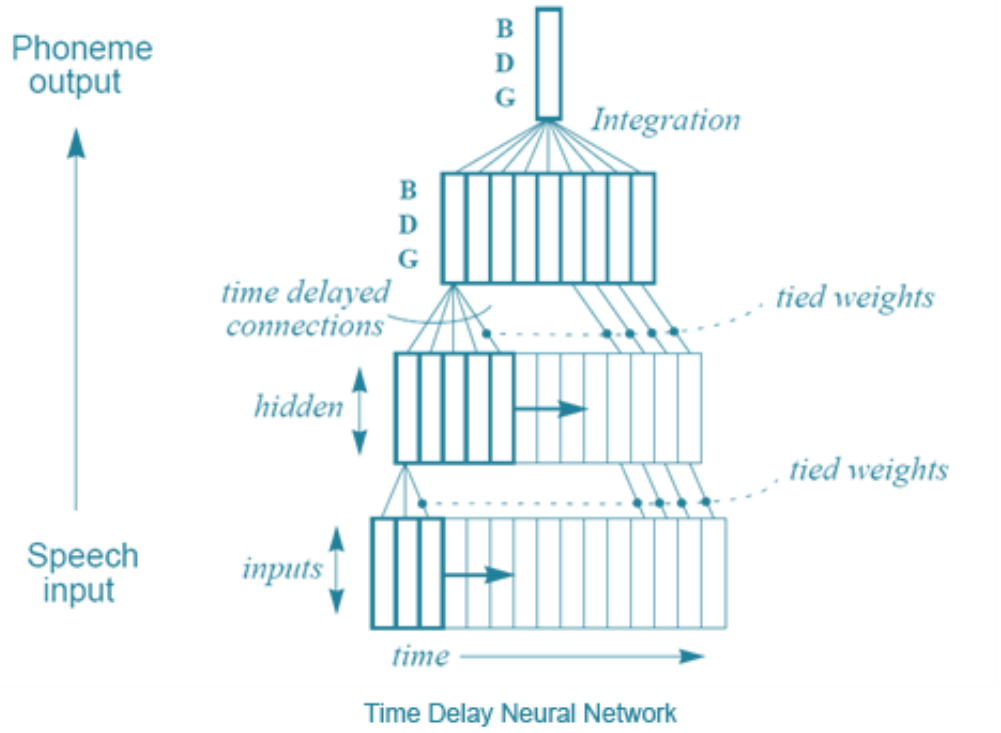


FIGURE II.6. « Time-Delay Neural Network »
Source

Avant de poursuivre du côté de la recherche en apprentissage profond dans des cadres non supervisés, abordons la question des réseaux dont la propagation du signal n'est pas uniquement *avant*. En effet, les réseaux considérés jusqu'alors ne sont constitués que de couches dont les sorties des neurones agissent comme l'entrée de neurones appartenant à des couches uniquement ultérieures. Par conséquent, si dans le cadre de données images, les réseaux convolutifs sont tout à fait adéquats par leur capacité à capturer l'information à des voisinages différents selon la profondeur dans les couches du neurone considérés, ils sont mal adaptés à d'autres domaines. Ainsi, des tâches telles que la retranscription écrite automatique d'une conversation enregistrée oralement (reconnaissance vocale), ou la traduction d'un texte d'une langue vers une autre, mettent en échec de tels modèles. La raison est intuitive et réside dans leur impossibilité de construire en leur sein, des liens de dépendance entre des données éloignées les unes des autres : une phrase linguistique est systématiquement construite selon une grammaire, grammaire pouvant varier d'un idiome à un autre, mais qui induit une *logique* dans l'enchaînement des mots (le sujet est toujours placé avant le verbe dans une phrase déclarative), et que doit prendre en compte un modèle à visée de traduction ou de retranscription automatique. En pratique, les réseaux prenant en compte cet aspect de dépendance à plus ou moins long terme, intègrent la notion de **mémoire**, autorisant une influence de certaines

parties du signal d'entrée sur d'autres. Les premiers travaux sur la reconnaissance vocale remontent en 1989. Alexander Waibel publie, notamment avec Geoffrey Hinton, dans (Waibel et al., 1995) le concept novateur de *Time Delay Neural Network* (*TDNN)*. Ces réseaux sont en quelques sortes précurseurs des CNNs, mais appliqués à un signal sonore, dans la mesure où chaque neurone n'a accès qu'à un morceau du signal d'entrée à la fois, là où un perceptron multicouche verrait l'ensemble du signal. Le signal est traité de manière séquentielle par des neurones parcourant ce signal à la manière d'une fenêtre glissante. Chaque neurone a plusieurs types de poids en fonction des différentes parties du signal analysé et donc des différents décalages pris en compte : un neurone est lié non seulement aux sorties des neurones de la couche précédente, mais également aux sorties *différées dans le temps*, de ces mêmes neurones. La figure II.6 illustre le propos précédent. Même si l'architecture *TDNN* intègre dans une certaine mesure une notion de temps, elle demeure une architecture à propagation avant, les neurones d'une couche donnée ne tenant compte que des sorties des neurones de couches précédentes. De plus, les connexions « différées » doivent être construites au préalable, induisant une restriction implicite de la part du signal passé considéré pour étudier le morceau de signal analysé.
Les réseaux de neurones récurrents (RNN) matérialisent véritablement le concept de *mémoire* en faisant apparaître des boucles en leur sein. Ainsi, ces boucles peuvent relier la sortie d'un neurone à l'entrée de ce **même neurone**, ou la sortie d'une couche de neurone à l'entrée d'une **couche antérieure** à cette dernière. L'avantage de cette approche est principalement une grande *souplesse mémorielle* apportée par les boucles : à l'inverse des TDNNs dont la mémoire est fixée à leur construction, chaque boucle s'intéresse à une partie du signal passé qu'elle détermine elle-même (mécanisme d'intérêt) au cours de l'entraînement au travers des couches cachées représentant cette boucle. Cette partie du signal est ainsi déterminée en fonction du but recherché. En revanche, le mécanisme de rétropropagation utilisé jusqu'alors avait été construit pour l'entraînement de réseaux dont les neurones était uniquement à propagation avant, l'idée étant, rappelons-le, de propager l'erreur contenue en sortie du réseau sur l'ensemble des poids en remontant successivement les couches. Dans le cas des RNNs, la notion d'ordre nécessaire pour appliquer cet algorithme semble compromise. Cette difficulté est en réalité surmontée par l'algorithme de *Backpropagation Through Time* (Werbos, 1988 ; Rumelhart et al., 1986 ; Williams et Zipser, 1995), qui consiste simplement à « déplier » le réseau récurrent en considérant chaque boucle sur un neurone comme une succession finie de ce neurone (dupliqué autant de fois que l'on boucle). Les poids de ce neurone sont ensuite moyennés à travers l'ensemble de ses répliques.
Au début des années 1990, Yoshua Bengio travaille sur le problème de *credit assignment* au sein des RNNs dans le cadre de la reconnaissance de paroles : modélisent-ils efficacement les relations de dépendance entre différentes parties du signal plus ou moins éloignées dans le temps (début et fin de phrase par exemple) ? Le mécanisme de *credit assignment* dont il fait mention renvoie à l'algorithme utilisé pour diffuser l'erreur globale sur chaque

paramètre, donc de l'algorithme de rétropropagation en ce qui concerne les RNNs. Ainsi, il montre que cet algorithme favorise fortement les dépendances à court terme lors de l'entraînement et ne prend pas en considération les dépendances à long terme au sein du signal (Bengio, 1993 ; Bengio et al., 1994), et affiche de meilleures performances pour des algorithmes types *Expectation - Maximization* (Bengio et Frasconi, 1994). La raison pour laquelle les RNNs peinent à modéliser ces longues dépendances réside dans le phénomène de gradients évanescents (*vanishing gradients*) qui apparaissent au cours de l'entraînement. Ce problème ne se limite pas aux RNNs mais à tous les réseaux, et survient lorsque le réseau est construit avec beaucoup de couches de neurones : au moment de l'application de l'algorithme de rétropropagation, la *chain rule* calcule le gradient de l'erreur par rapport à chaque activation et le multiplie avec les gradients des couches précédentes au fur et à mesure que l'on remonte dans le réseau. Les activations sont généralement des fonctions asymptoptiques, telle que la fonction tangente hyperbolique, bornées entre $[0, 1]$: le calcul de gradient fait intervenir la dérivée de la fonction tanh, qui est également comprise entre $[0, 1]$, donc si certains gradients sont faibles, le résultat de leur multiplication avec d'autres gradients les réduit davantage encore (réduction exponentielle en fonction du nombre de couche). Or, un RNN est entraîné par rétropropagation en le dépliant comme décrit précédemment, ajoutant donc un nombre considérable de couches. La dépendance à long terme nécessite beaucoup de boucles et donc d'ajouter un nombre de couches conséquent au réseau, c'est pourquoi c'est ce type de dépendance qui est difficilement modélisé.
En 1997, Sepp Hochreiter et Jürgen Schmidhuber modifient l'architecture RNN vers ce qu'ils dénomment *Long Short Term Memory Network* (Hochreiter et Schmidhuber, 1997), ou LSTMs. Ce modèle adresse spécifiquement les limitations posées par les RNNs pour capturer les relations de longues dépendances. Sepp Hochreiter avait d'ailleurs mis en évidence au même titre que Yoshua Bengio ces limitations concernant les RNNs en 1991 (Hochreiter, 1991). Les neurones d'un RNN peuvent être dépliés en plusieurs cellules identiques au neurone concerné, les sorties des uns reliés aux entrées des suivants. Les LSTMs explorent la même idée, mais en ajoutant des fonction de *gating* au sein de chaque cellule, fonctions venant modifier ou non une variable appelée *cell state* C_t qui circule entre toutes les cellules. C_t est une représentation de l'information issue du réseau à l'époque t. Au sein de chaque cellule, quatre fonctions de *gating* agissent comme filtres d'informations, ou fournisseurs de nouvelles informations. Le calcul de gradient met en jeu cette fois-ci l'une des *gates*, la *forget gate* qui modélise la capacité de la cellule à oublier ou non une information provenant de la cellule précédente. Ce gradient peut donc être non nul, résolvant le problème de gradients évanescents. L'information contenue dans la *cell state* peut ainsi rester inchangée au cours de ses transitions au sein des différentes cellules, préservant les relations de dépendances longues au sein du signal. En 2014, Cho et al. (2014) propose une architecture fondée sur le principe de *gates functions* à l'instar des LSTMs, mais en en réduisant le nombre. La structure du réseau est donc moins complexe, conduisant à un nombre moindre de paramètres nécessaires comparés

aux LSTMs, pour des performances similaires.

Si la majeure partie de l'effort fourni en matière de recherche en apprentissage profond s'est concentré jusque là essentiellement sur les processus supervisés, les réseaux artificiels ont également pu tirer leur épingle du jeu dans des applications non supervisées. En vérité, G. Hinton, co-auteur de l'article qui avait fait ressurgir les réseaux de neurones artificiels (Rumelhart et al., 1986), introduisait en 1985, et avec David Ackley et Terry Sejnowski, le concept de *Restricted Boltzmann Machine*s (RBM) (Ackley et al., 1985), réseaux artificiels similaires à ceux connus jusqu'alors, mais dont les neurones ont un état stochastique. Une RBM est une variante d'une *Boltzmann Machine* ; cette dernière est un champ de markov non-orienté, dont tous les nœuds peuvent être interconnectés, rendant le processus d'entraînement difficile. Les RBMs imposent une restriction supplémentaires, partitionner le graphe en deux couches dont les nœuds (neurones) d'une même couche sont indépendants entre eux (pas de connexion intra-couche).
Modélisant la distribution des données recueillies, les RBMs sont des modèles génératifs non supervisés. Ainsi, en plus de pouvoir résoudre des problèmes de classification, à l'instar de réseaux évoqués jusqu'alors, les RBMs peuvent être employées pour générer des données suivant la distribution de probabilité que le modèle a appris, ou encore pour reconstruire des données initialement incomplètes. Toutefois, l'utilisation de réseaux de neurones artificiels dans des configurations non supervisées remonte avant cela, en 1982, avec notamment les cartes de Konohen (Kohonen, 1982) (*Self-Organizing Maps*) pouvant servir dans des tâches de discrétisation d'un espace de données de grandes dimensions : cette réduction de dimension est désirable dans les tâches de classification ou de visualisation, deux autres domaines où ces cartes trouvent donc leur utilité.
Hinton et Zemel (1994) ont introduit une structure dite « auto-encodeur » et ont démontré leur capacité à modéliser des distributions de probabilités à partir d'observations. Dans sa forme la plus élémentaire, un auto-encodeur est constitué d'une couche intermédiaire dont le nombre de sorties est inférieur au nombre d'entrées, compressant ainsi l'information contenue dans les données d'entrée. Les données sont ensuite reconstruites à l'issue de la couche de sortie. En 2006, dans la poursuite de ses investigations sur les auto-encodeurs, Hinton propose une architecture, les *Deep Beliefs Networks* ou réseaux de croyance profonde (Hinton et Salakhutdinov, 2006), qui n'est ni plus ni moins qu'une succession de RBMs, et dont l'entraînement est conduit couche par couche, par ajout ensuite d'une nouvelle RBM. L'ensemble du réseau ainsi pré-entraîné est par la suite affiné dans son ensemble. Cette approche d'apprentissage couche par couche, publiée dans Hinton et al. (2006) est l'un des facteurs de la résurgence des réseaux de neurones profonds et du regain d'intérêt suscité au sein de la communauté de *computer vision*, puisqu'il est dorénavant possible d'entraîner des réseaux avec un nombre de couches important.

Avant d'en venir aux avancées majeures du XXIème siècle en matière

d'apprentissage profond et de réseaux de neurones artificiels, nous abordons dans ce paragraphe la notion d'**apprentissage par renforcement** (*reinforcement learning*). Ce troisième processus d'apprentissage diffère des deux précédentes méthodes (supervisées et non supervisées) par le concept mis en œuvre. Plutôt que de fournir à un classifieur supervisé des données avec leurs étiquettes (processus supervisé), ou de chercher à regrouper les objets selon des critères de similarité (processus non supervisé), l'objectif de l'apprentissage par renforcement est d'apprendre à un agent autonome à *opter pour les bons choix / actions* pour une tâche donnée. Ainsi, cet agent qui évolue dans un environnement donné cherche à maximiser ce que l'on appelle une fonction de récompense, permettant audit agent d'évaluer si les actions qu'il a pu choisir pendant un épisode d'apprentissage mènent à une issue favorable (récompense gratifiante), ou un échec (pénalité). Au cours d'un épisode, l'agent change d'état par le biais des actions qu'il choisit d'accomplir (le périmètre des actions possibles est prédéfini). C'est ce type d'algorithme qui a été le vecteur initial de publicité de l'apprentissage profond auprès du grand public, lors du succès sans appel de l'algorithme AlphaGo (Silver et al., 2016), mis au point par l'entreprise britannique DeepMind : en 2016, cette *Intelligence Artificielle* parvient pour la première fois à battre les meilleurs joueurs de Go de la planète. Cette tâche était le défi principal pour une intelligence artificielle, la complexité du problème étant très élevée : le nombre de partie estimée est de 10^{600}, et le nombre de positions en accord avec les règles de 10^{140}, rendant la détermination du meilleur coup possible par exploration des différentes possibilités impossible à conduire. L'algorithme a été entraîné par observation des coups successifs de parties conduites par des joueurs humains. Depuis, l'équipe de DeepMind a grandement accru les capacités de leur modèle, permettant à ses versions successives, la dernière étant AlphaZero (Silver et al., 2017) de ne s'améliorer qu'en jouant contre lui-même, surpassant l'ensemble de ses versions précédentes, y compris AlphaGo. Aujourd'hui, l'apprentissage par renforcement est essentiellement utilisé dans le domaine du jeu vidéo, de la robotique et dans certains systèmes de recommandation : la bibliothèque Gym (Brockman et al., 2016) met a disposition un environnement de *reinforcement learning* compatible avec les frameworks d'apprentissage profond existants aujourd'hui (Keras notamment).

Nous avons évoqué l'avancée majeure effectuée par Hinton en 2006 (Hinton et al., 2006) en matière algorithmique afin de rendre l'entraînement des réseaux de neurones profonds possibles. Toutefois, si les DNNs (*Deep Neural Networks*) ont su convaincre les différentes communautés intégrant une part de *computer vision* ou de *machine learning* méthodologiquement parlant, un écueil demeurait en terme de moyens matériels. En effet, malgré l'augmentation des fréquences de calcul des CPU, ceux-ci restaient insuffisants pour gérer les, parfois, quelques dizaines de millions de paramètres, des modèles envisagés alors. En 2009, Raina et al. (2009) comparent les performances entre une approche CPU, et une approche GPU, démontrant que l'utilisation de GPU versus CPU réduisent les temps d'apprentissage d'un facteur allant

jusqu'à 72. Cela a permis d'adopter des stratégies *brute force*, à savoir accroître les jeux de données d'entraînement et la taille des réseaux de neurones utilisés (Claudiu Ciresan et al., 2010), qui ont pris le pas sur des avancées réelles en matière d'algorithmie. Nous venons de parler de l'accroissement des jeux de données lors de l'utilisation de GPUs ; cela a été rendu possible en conjonction avec la multiplication des données numériques disponibles (jeux de données labellisés), que l'on appelle « Big Data ». Le challenge ILSVRC (*ImageNet Large Scale Visual Recognition Challenge*) de 2012 a été le point culminant de l'ascension progressive des réseaux de neurones depuis 2006. L'unique noyau proposé mettant en pratique un réseau de neurone convolutif (Krizhevsky et al., 2012) a obtenu un taux d'erreur *top-5* de 15.3% tandis que le second meilleur résultat n'atteignait que 26.2%. Le taux d'erreur *top-5* correspond au taux d'échantillons dont la classe réelle est parmi les cinq classes les plus probables. Le challenge ILSVRC intègre environ 1000 classes. La performance affichée par le réseau convolutif proposé a dès lors suscité un très vif intérêt pour ce type d'architecture et pour l'apprentissage profond plus largement qui demeure aujourd'hui.
Le défi principal auquel font continuellement face les réseaux de neurones profonds, en plus des problèmes de gradients évanescents, est le risque de **sur-apprentissage**. L'ère du « Big Data » a permis de pallier ce problème, mais en partie seulement, les réseaux étant toujours de plus en plus profonds. A partir de 2010, nombre de travaux ont porté sur différents mécanismes à mettre en œuvre au moment de l'apprentissage pour faciliter l'apprentissage et accroître le pouvoir de généralisation des réseaux. Les paragraphes suivant détaillent les parties composantes d'un réseau de neurones, le processus d'apprentissage classique d'un point de vue plus théorique, et les mécanismes évoqués précédemment et facilitant cet apprentissage .

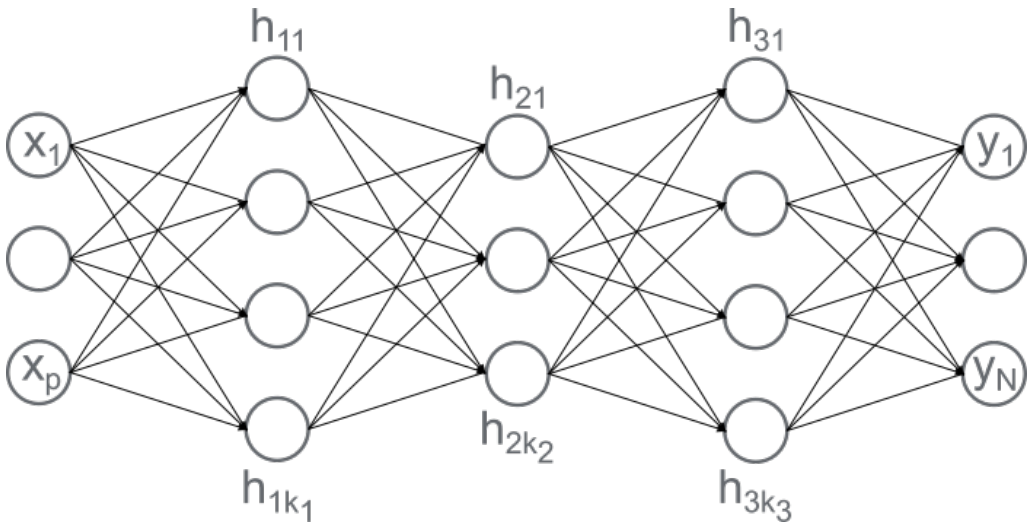


FIGURE II.7. Perceptron multicouche.

Composants d'un réseau de neurones

Ce paragraphe présente les différentes parties d'un réseau de neurones. L'unité minimale constitutive d'un réseau est donc le neurone, illustré Figure II.4, et formalisé mathématiquement comme suit :

$$y = f(\sum_{i=1}^{p} w_i x_i + b) \tag{II.2}$$

Plusieurs variables à détailler dans l'équation précédente :

— les $(w_i)_{i\in\{1,..,p\}}$ désignent les paramètres ;

— les $(x_i)_{i\in\{1,..,p\}}$ renvoient aux valeurs d'entrée ;

— un biais b permettant de donner plus de flexibilité au neurone dans le cas où les valeurs d'entrée sont toutes nulles par exemple ;

— f est la fonction d'activation ; inspirée du fonctionnement des neurones biologiques, elle « active » le neurone si celui-ci est suffisamment stimulé au regard du produit scalaire entre les paramètres et les valeurs d'entrée.

Le perceptron (Figure II.5) présente un empilement de neurones dans une seule couche. Il est également possible de multiplier ces couches, tout en reliant les sorties $(y_i)_{i>0}$ d'une couche m aux entrées de la couche $m+1$, succédant immédiatement à m. C'est ce que met en pratique un perceptron multicouches, dont un exemple est donné Figure II.7. Les couches comprises entre les données d'entrée et les résultats en sortie sont appelés « couches cachées » et intègrent l'ensemble des paramètres optimisés pendant le processus d'entraînement. Un réseau ne présentant pas de boucle à l'instar des RNNs est dit à propagation avant, l'information circulant depuis les données d'entrée vers la sortie, avec un passage dans chaque neurone de chaque couche une fois exactement, et dans l'ordre d'apparition des couches.

Les fonctions d'activation de la littérature sont multiples mais chronologiquement et intuitivement, la première est la fonction de Heaviside, ou fonction « marche » qui renvoie 1 si le résultat du produit scalaire est supérieur à 0, 0 sinon : $f(x) = \frac{1}{2}(sgn(x)+1)$ avec sgn la fonction signe. Toutefois, étant constante par morceaux, les gradients de cette fonction sont tous nuls sauf en 0. Le paragraphe suivant dédié à l'algorithme de rétropropagation, nous verrons que celui-ci est fondé sur le calcul de gradients, rendant par conséquent la fonction de Heaviside peu adaptée. L'unique bonne propriété de la fonction de Heaviside est sa non-linéarité qui apporte à l'ensemble des couches leur raison d'être : en effet, dans le cas de fonction linéaire, la sortie du réseau ne serait somme toute rien de plus qu'une combinaison linéaire des valeurs d'entrée, et donc l'ensemble des couches pourraient finalement être concaténées en une seule couche de neurones.
Cette propriété se retrouve dans l'ensemble des fonctions que l'on passe en revue dans la suite. En revanche celles-ci présentent également l'intérêt de fournir un ensemble de définition dont les valeurs sont à gradients non nuls

(exception faite de quelques points éventuellement). La fonction logistique, aussi appelé sigmoïde en raison de l'allure de sa courbe représentative, a été abondamment utilisée historiquement pour des raisons de commodités mathématiques : $\sigma(x) = \frac{1}{1+e^{-x}}$. Non linéaire, et bornée entre 0 et 1, la sigmoïde était appréciée du fait de cette dernière propriété ; elle est une interprétation théorique claire du phénomène d'activation des neurones biologiques, avec une activation nulle ou faible en cas de valeur d'entrée faible, et une activation forte si cette valeur est élevée. Malheureusement, le bornage de cette fonction est également ce qui l'a desservie au fur et à mesure de la recherche en apprentissage profond. Le comportement de la fonction est asymptotique et saturant à ses bornes, avec donc des valeurs de gradients très faibles voire nulles pour des valeurs faibles ou élevées du terme issu de la somme dans l'équation II.2. Cela renvoie à ce qui a été discuté dans la section précédente 2.2 à propos des RNNs : un réseau dont les neurones sont dotés de la fonction sigmoïde est sujet au problème de gradients évanescents lors de l'apprentissage, ceux-ci étant de plus en plus faibles en remontant progressivement au travers des différentes sigmoïdes (par multiplication). De même, la fonction tangente hyperbolique tanh présente le même inconvénient de saturation à ses bornes, et a été moins utilisée ces dernières années, même si nous verrons que le processus de *batch normalization* permet un emploi de fonctions saturantes sans que cette même propriété de saturation n'ait d'impact réellement significatif.
En 2011 est proposée la fonction rampe appelée *Rectified Linear Unit* (Glorot et al., 2011) par les auteurs, ou ReLU : $ReLU(x) = max(0, x)$. Deux différences majeures avec la fonction sigmoïde et *tanh* sont visibles graphiquement. La fonction ReLU n'est pas bornée, et n'est saturante que dans une direction, celle des valeurs négatives. De plus, la fonction n'est pas différentiable en 0, mais cette fausse difficulté est levée en affectant la valeur 0 à la dérivée de la discontinuité en 0. Krizhevsky et al. (2012) la compare avec la fonction *tanh* et démontre par l'expérience que la fonction rampe améliore d'un facteur 6 la vitesse vers laquelle le réseau converge. La tendance peu saturante de la fonction ReLU réduit la fréquence d'apparition de gradients évanescents, et contribue à l'accélération du processus. Mais c'est aussi sa formulation simple qui permet de réduire considérablement le nombre d'opérations nécessaires pour appliquer la fonction : comparer aux fonctions sigmoïdes et *tanh* qui utilisent des fonctions exponentielles, ReLU ne requiert qu'un seuillage à 0 de l'activation, le résultat étant l'identité dans le cas où l'activation est supérieure à 0. Une version modifiée du ReLU, le *Leaky ReLU* (Maas et al., 2013) vise à résoudre le problème de « ReLU mourant » qui survient lorsqu'un neurone n'est pas actif, et ne voit donc pas ses poids modifiés par la méthode de rétropropagation (gradient nul). De tels neurones seraient inactifs pour le reste de l'entraînement dans le cas de ReLUs standards comme fonctions d'activation, mais ajouter un faible gradient non nul à la place permet de potentiellement lever cette inactivité au cours des itérations.

Maintenant que nous avons passé en revue les parties constitutives que doit avoir *a minima* un réseau de neurone, le paragraphe suivant décrit les

mécanismes liés à l'entraînement de tels modèles, et aux différents processus qui y sont liés et qui facilitent cet entraînement.

Apprentissage des réseaux de neurones : algorithme de rétropropagation et bonnes pratiques

Un réseau de neurone peut en théorie approcher n'importe quelle fonction continue par morceau en vertu du théorème d'approximation universelle (Hornik, 1991). L'algorithme de rétropropagation du gradient (Werbos, 1974 ; Rumelhart et al., 1986) est utilisé pour, itérativement, mettre à jour l'ensemble des poids des paramètres du réseau jusqu'à convergence par rapport à une métrique donnée.
Cette méthode met en œuvre la descente de gradient pour estimer les paramètres du réseau. Remontant au moins au XIXème siècle, la descente de gradient a pour but de minimiser une fonction objectif (différentiable) au cours des époques d'apprentissage : une fonction différentiable décroît le plus rapidement dans le sens inverse de son gradient (plus grande pente). En effet, le gradient d'une fonction pointe dans la direction où cette fonction croît le plus rapidement. Lorsque le gradient est nul, la fonction étudiée a atteint un minimum global dans le cas convexe, un minimum local sinon. Ce dernier cas est le plus courant lorsqu'on utilise un réseau de neurone ; l'algorithme ne nous garantit donc pas de trouver *le* meilleur jeu de paramètres, mais *un* jeu de paramètres permettant d'obtenir une solution plus ou moins satisfaisante parmi d'autres. Cela a des conséquences non négligeables sur l'entraînement d'un réseau puisque plusieurs points de convergence existent mais ne mènent pas aux mêmes performances. Plusieurs « méta-paramètres » appelés hyperparamètres (terme défini dans la suite) permettent de contrôler différents aspects de l'apprentissage et peuvent modifier le minimum local atteint. Le choix du meilleur réseau est donc couramment effectué par une approche *trial-error*, ce qui est important à noter, une telle approche étant particulièrement coûteuse en ressources matérielles et en temps de calcul dans le cadre de réseaux de neurones.
L'objectif étant de minimiser une fonction objectif f, ou *énergie*, par analogie à l'énergie en physique d'un système stable si celle-ci est minime. Si l'on formalise les choses, l'algorithme est ainsi :

Algorithme de la descente de gradient. *Soit* $f : \mathbb{R}^N \to \mathbb{R}$ *une fonction différentiable sur* $\mathbb{R}^N$. *Soit* ∇f *le gradient de* f *et soit* $\eta,\ \epsilon \in \mathbb{R} \mid \eta > 0,\ \epsilon > 0$.

1/ On initialise l'algorithme à $a_0 \in \mathbb{R}^N$.

2/ l'algorithme définit une suite $(a_n)_{n>0} \in \mathbb{R}^N$ *telle que :*

$$a_{n+1} = a_n - \eta \nabla f(a_n) \tag{II.3}$$

3/ Tant que $\nabla f(a_n) > \epsilon$, *on itère en suivant 2/.*

L'algorithme met en jeu deux constantes ϵ et η. ϵ est un seuil de tolérance en lien avec la précision / finesse de la solution apportée. Si la méthode du

gradient permet d'obtenir la direction de plus grande pente pour une fonction donnée, elle ne donne cependant pas la *quantité* à parcourir dans cette direction : c'est pourquoi la seconde constante, η, appelée *learning rate* ou vitesse d'apprentissage, a lieu d'être. Le *learning rate* est un des *hyperparamètres* du modèle : il est défini avant l'entraînement et n'est pas contrôlé par le gradient de la fonction analysée. η est l'un des hyperparamètres les plus importants et doit être choisit avec précaution. η permet de contrôler la quantité d'information issue du gradient à prendre en compte lors de la mise à jour des paramètres. Une valeur η trop élevée provoque un effet de « rebondissement » du gradient autour du minimum local vers lequel il se dirige, tandis qu'une valeur trop faible réduit considérablement les temps de calcul (les valeurs de paramètres étant quasi similaires entre deux itérations). Une bonne pratique est de considérer une fonction de décroissance du *learning rate* pendant l'entraînement (Zeiler, 2012 ; Kingma et Ba, 2014). Cela permet en début d'entraînement de converger vers un minimum local « grossier » puis d'affiner ensuite le réseau en modifiant petit à petit les paramètres.

Les réseaux de neurones ont également pour objectif de minimiser une fonction d'énergie E dont les paramètres W sont ceux du réseau. La fonction E est située en sortie de réseau et doit être minimisée pour l'ensemble du jeu d'apprentissage $\mathcal{D}$ à disposition :

$$\hat{W} = Argmin_W E(W, \mathcal{D}).$$

En adaptant la notation de l'algorithme avec cette formulation, l'équation II.3 devient

$$W \leftarrow W - \eta \nabla_W E(W, \mathcal{D}). \tag{II.4}$$

W dénote le vecteur de coefficients qui paramétrisent notre réseau de neurones profonds. Constitué de plusieurs couches, celui-ci ne donne pas accès directement au gradient de la fonction objectif par rapport à chaque coefficient $\frac{\partial E}{\partial w_i}$ puisque E est calculé à la sortie du réseau, chaque coefficient en étant donc séparé par plusieurs couches (sauf pour ceux de la dernière couche).
Le problème est résoluble en considérant le réseau comme une composition de fonctions d'activation dont chaque fonction (neurone) est l'argument du (des) neurone(s) suivant(s). Les fonctions d'activation étant différentiables, on peut appliquer la *chain rule*, terme anglophone désignant le théorème de fonctions composées. En pratique, cela se traduit par des multiplications de matrices jacobiennes intégrant l'ensemble des gradients par rapport à toutes les activations de chaque neurone en chaque couche du réseau pour atteindre chacun des paramètres $w_i \in W$. Prenons l'exemple suivant en guise d'illustration :
Soit $f : \mathbb{R} \to \mathbb{R}^2$ et $g : \mathbb{R}^2 \to \mathbb{R}^2$, différentiables sur $\mathbb{R}$ et $\mathbb{R}^2$ respectivement, telles que :

$$\forall x \in \mathbb{R},\ f(x) = [f_1(x), f_2(x)] \ \ et \ \ \forall y \in \mathbb{R}^2,\ g(y) = [g_1(y_1, y_2), g_2(y_1, y_2)],$$

$$avec f_1,\ f_2,\ g_1,\ g_2 : \mathbb{R} \to \mathbb{R}.$$

Par composition, $g \circ f(x) = g(f(x)) = [g_1(f_1(x), f_2(x)), g_2(f_1(x), f_2(x))]$. La composition de fonctions peut être interprétée comme une succession de couches d'opérations, $f_1,\ f_2,\ g_1,\ g_2$ étant les activations. Le gradient qui nous intéresse ici est celui de la sortie g par rapport à la valeur d'entrée x, soit $\frac{\partial g}{\partial x}$. Pour y parvenir, on calcule progressivement les gradients en remontant le long des couches :

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \frac{\partial}{\partial x} g_1(f_1(x), f_2(x)) \\ \frac{\partial}{\partial x} g_2(f_1(x), f_2(x)) \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial x}(f_1(x)) + \frac{\partial g_1}{\partial x}(f_2(x)) \\ \frac{\partial g_2}{\partial x}(f_1(x)) + \frac{\partial g_2}{\partial x}(f_2(x)) \end{pmatrix}$$

Finalement, en utilisant la *chain rule*, il vient :

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \frac{\partial g_1}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_1}{\partial f_2}\frac{\partial f_2}{\partial x} \\ \frac{\partial g_2}{\partial f_1}\frac{\partial f_1}{\partial x} + \frac{\partial g_2}{\partial f_2}\frac{\partial f_2}{\partial x} \end{pmatrix}$$

Le terme de droite de l'égalité précédente n'est rien de plus que la multiplication de deux matrices jacobiennes :

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \frac{\partial g_1}{\partial f_1} & \frac{\partial g_1}{\partial f_2} \\ \frac{\partial g_2}{\partial f_1} & \frac{\partial g_2}{\partial f_2} \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x} \\ \frac{\partial f_2}{\partial x} \end{pmatrix} = (J_g \circ f).J_f$$

Le mécanisme de propagation du gradient à travers les couches grâce à la *chain rule* est illustrée Figure II.8.

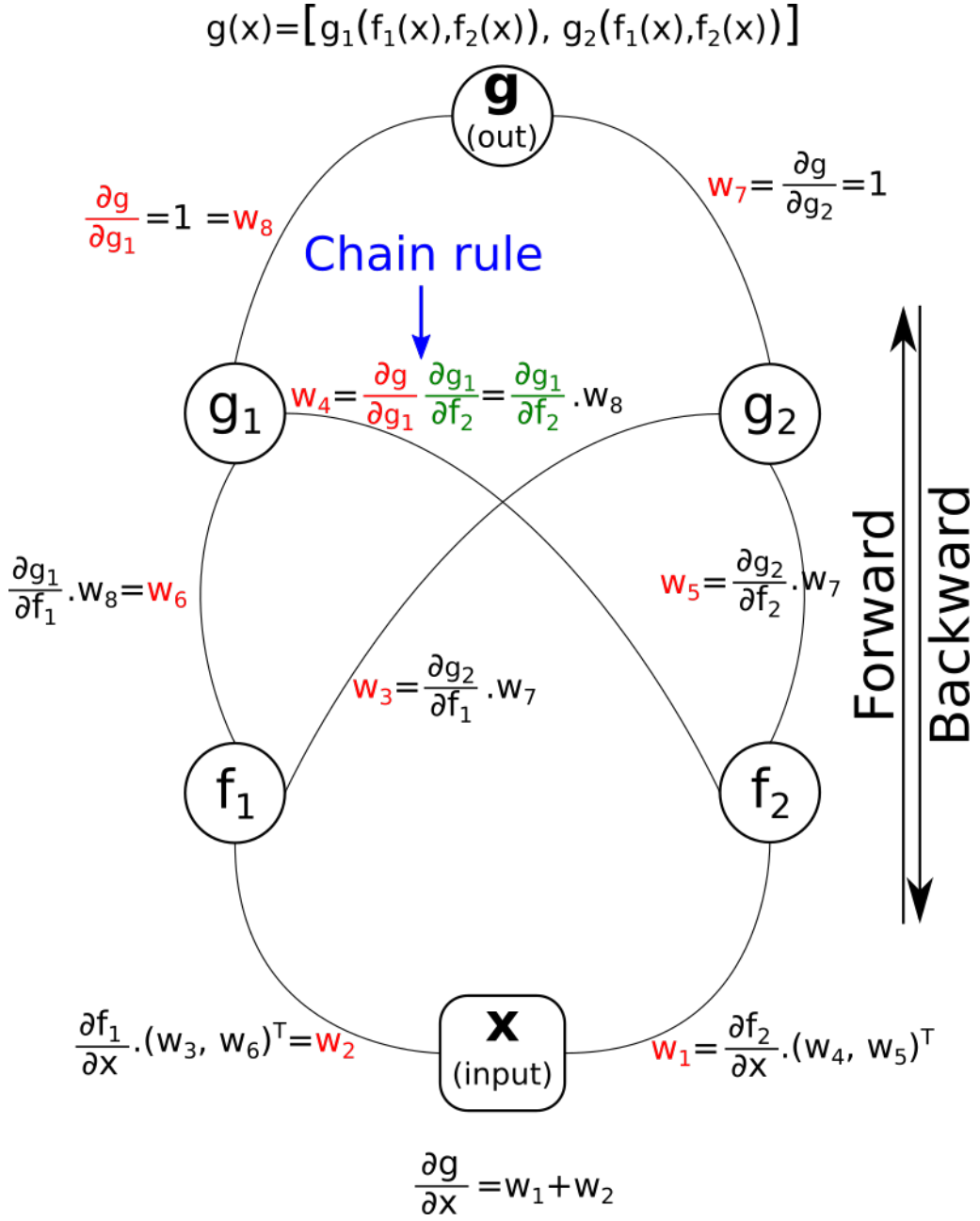


FIGURE II.8. On matérialise dans cette figure la rétro-propagation du gradient au moyen de la *chain rule*.

Avec l'algorithme tel qu'il est décrit à l'équation II.4, le jeu d'apprentissage total doit être soumis au réseau avant de procéder au calcul des nouveaux paramètres lors d'une itération. Nous l'avons précisé en fin de la section 2.2, lorsque le terme de « réseau de neurones » est employé, les données massives d'entraînement lui font souvent écho. Or, derrière le terme $\nabla_W E(W, \mathcal{D})$ se cache une sommation sur l'ensemble des échantillons avant de procéder à une mise à jour des paramètres W, pouvant conduire à de très longs temps de calcul. Par ailleurs, les données au sein du jeu d'entraînement

peuvent être corrélées, impliquant le calcul de gradients qui seraient similaires et n'apporteraient ainsi aucune nouvelle information. Une simple modification de l'algorithme permet de réduire considérablement les temps de calcul : l'évaluation du gradient et la mise à jour des paramètres est conduite après soumission d'un seul échantillon tiré aléatoirement dans le jeu de données (*online learning* dans le jargon de machine learning pour désigner les méthodes ne pouvant traiter l'ensemble du jeu d'apprentissage en une fois). C'est ce que l'on appelle la descente de gradient stochastique, ou SGD (*Stochastic Gradient Descent*) (Robbins et Monro, 1951), qui a été abondamment étudiée et comparée à une descente de gradient classique (Bottou et Cun, 2004 ; Bottou et Bousquet, 2008), avec des conclusions en faveur de l'approche stochastique, à la fois en terme de temps de convergence et de performance. Le terme différencié dans l'équation II.4 est modifié ainsi : $\nabla_W E(W, s)$ avec $s \in \mathcal{D}$ un échantillon du jeu de données initial.
Ainsi posé, l'algorithme SGD présente un inconvénient majeur : en procédant à une mise à jour des paramètres après soumission de *chaque échantillon*, le chemin parcouru vers un minimum local sera vraisemblablement bruité, avec pour conséquence un nombre nécessaire d'itérations important pour atteindre ce minimum. Finalement, le juste équilibre se situe entre SGD et descente de gradient classique : plutôt que de fournir au réseau (i) l'ensemble du jeu d'apprentissage comme dans le premier cas ou (ii) un seul échantillon dans le second cas, avant de recalculer les paramètres, l'utilisation de « paquets » d'échantillons, ou *batches* est un compromis qui considère un sous-ensemble constitué de N échantillons du jeu de données à chaque époque d'apprentissage. L'équation II.4 doit être modifiée pour prendre en compte, cette fois-ci, la somme des gradients sur les N échantillons appartenant à un batch $\mathcal{B} \subset \mathcal{D}$:

$$W \leftarrow W - \eta \frac{1}{N} \sum_{k=1}^{N} \nabla_W E(W, \mathcal{B}_k) \tag{II.5}$$

Ce simple changement offre plusieurs avantages :

1. le modèle converge vers un minimum en suivant un chemin moins bruité que dans le cas stochastique, puisque le réseau n'est mis à jour qu'en considérant la moyenne des gradients calculés sur les N échantillon d'un batch ;
2. de plus, comparé au cas purement stochastique, les temps de calcul sont grandement réduits, les paramètres n'étant recalculés qu'en fin de batch plutôt qu'à chaque échantillon ;
3. la convergence est plus rapide que dans le cas non stochastique, le gradient calculé par batch étant représentatif du gradient du jeu de données total, et la fréquence de mise à jour des paramètres étant plus élevée ;
4. le système est hautement parallélisable, chaque batch étant indépendant des autres, on peut affecter à plusieurs *jobs* un batch diffférent (*multithreading* sur GPU).

Le jeu de données est généralement divisé en batches contenant un nombre d'échantillons suivant les puissances de deux pour des questions de gestion

mémoire et d'optimisation des opérations vectorielles.

La fonction F représentée par le réseau est la traduction mathématique de la tâche à accomplir par celui-ci. Pour approcher cette fonction au mieux (on n'en connaît pas de formule explicite), on cherche à optimiser une fonction d'énergie (mentionnée précédemment) utilisant la connaissance que l'on a a priori sur les données (jeu d'entraînement), et ce que le réseau produit à chaque itération. Les fonctions d'énergie varient selon la tâche à effectuer, mais, dans le cas supervisé, elles mettent très souvent en jeu la différence entre la prédiction du réseau et la donnée réelle du jeu d'entraînement, ce qui correspond à l'erreur d'approximation du réseau pour un échantillon du jeu :

$$E(x) = E(\hat{F}_{net}(x) - F(x)) = E(\hat{y}, y),$$

avec $\hat{F}_{net}$ la fonction recherchée, modélisée par le réseau, $\hat{y}$, y respectivement la sortie du réseau et la sortie attendue, pour l'échantillon x. Si cette différence tend vers 0, on a un bon estimateur de cette fonction, d'où la nécessité de **minimiser** l'énergie E.
Une énergie couramment utilisée dans le cas où F est continue est la norme L_2, on suit alors l'algorithme des moindres carrés. Toutefois, dans notre cas, on cherche à résoudre un problème de classification, rendant l'espace d'arrivée de F discret et fini. Dans ce cas, $\hat{y}$ et y sont des vecteurs de mêmes dimensions (nombre de classes à séparer), y étant égal à 0 partout sauf à l'indice de classe correspondant (on parle de *one-hot encoded label*). Pour comparer et évaluer l'erreur d'approximation, l'entropie croisée est très abondamment utilisée dans la littérature. Cette fonction permet de comparer le vecteur $\hat{y}$ de probabilités d'appartenance à chaque classe, issu du réseau, au vecteur y, représentant les classes réelles de chaque item i, selon la relation suivante :

$$E(\hat{y}, y) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i),$$

avec N le nombre de classes. En d'autres termes, on compare les distributions entre y_i et $\hat{y}_i$. Les y_i étant nuls pour toutes les classes sauf une, on ne leur applique pas l'opération log. Par ailleurs, afin de comparer véritablement des distributions statistiques, les $\hat{y}_i$ doivent être homogènes à un comportement probabiliste tels que $\sum \hat{y}_i = 1$ et, $\forall i\ \hat{y}_i \in [0, 1]$. On utilise pour cela le *softmax* sur les activations a_i de la dernière couche du réseau, obtenant ainsi l'intervalle $[0, 1]$ comme image de n'importe quel intervalle réel continu :

$$\hat{y}_i = softmax(a_i) = \frac{e^{a_i}}{\sum_{k=1}^{N} e^{a_k}}.$$

La multiplication par -1 permet de rendre l'énergie positive (on somme des termes négatifs par application de log sur des nombres plus petits que 1).

Nous avons précisé en tout début de paragraphe que l'unique cas où la descente de gradient garantit que la solution trouvée correspond au minimum global est le cas d'une énergie E convexe. Or, cette bonne propriété n'est

jamais vérifiée avec les énergies utilisées dans les réseaux de neurones, qui traduisent, au contraire, des problèmes fortement non convexes. Le risque est donc grand de fournir un jeu de paramètres correspondant à un minimum local éloigné du minimum global. Sans chercher à atteindre ce dernier, plusieurs techniques ont été mises au point pour donner plus de flexibilité et de pouvoir exploratoire aux réseaux. Nous avons déjà indiqué le principe et rôle du *learning rate* ; son implémentation rend impossible toute sortie d'un minimum local puisqu'il est doté d'une décroissance exponentielle, donc chaque itération ne peut que réduire la portée de la direction vers laquelle les gradients sont modifiés. Le mouvement d'une bille le long d'une pente est rapidement perturbé voire stoppé au passage de celle-ci au-dessus d'un creux si sa vitesse initiale n'est pas suffisamment élevée ; cela a inspiré l'introduction du **momentum** (Sutskever et al., 2013 ; Qian, 1999) offrant la possibilité pour le réseau de sortir d'un minimum local en considérant non seulement le gradient de l'énergie par rapport à chaque paramètre au moment de la mise à jour du réseau à une époque donnée, mais également le gradient de l'époque précédente pondéré par ce *momentum*.
Les techniques énoncées jusqu'ici n'adressent aucunement le problème majeur en apprentissage supervisé, à savoir le sur-apprentissage. En tout cas, pas d'un point de vue méthodologique, étant donné que l'utilisation massive de données est la seule réponse que l'on a apportée pour tenter de contrer cet effet. Aujourd'hui, l'utilisation de l'**augmentation de données** (Simard et al., 2003 ; Wan et al., 2013) est très répandue et permet de générer artificiellement de nouvelles données. Par exemple, sur des images aériennes ou spatiales, la détection d'objets au sol est indépendante de l'azimut d'acquisition (angle entre la direction de l'objet et la direction de vol). On peut donc appliquer à chaque image une rotation dont l'angle est aléatoire pour accroître la représentation d'une classe et ainsi le pouvoir de généralisation du réseau pour détecter cette classe. Cette augmentation de données est souvent *online*, à savoir que les images « augmentées » sont générées à la volée, une fois le batch constitué et juste avant son passage dans le réseau. Cela évite d'avoir à stocker des quantités gigantesques de données, et réduit les temps de calcul en réduisant considérablement les temps d'accès mémoire.
Récemment, Cubuk et al. (2019) a entraîné un modèle ne modifiant pas les images directement, mais estimant les fonctions de transformation les plus adaptées pour un problème donné. Par ailleurs, les *Generative Adversarial Networks* (GANs) (Goodfellow et al., 2014) ont été abondamment utilisés afin de générer cette fois-ci directement les images transformées (Antoniou et al., 2017 ; Perez et Wang, 2017). Les GANs utilisent deux réseaux de neurones en compétition afin de générer des données possédant la même distribution que les données du jeu d'apprentissage. Le premier réseau, dit « génératif », modélise de nouvelles données et a pour objectif de maximiser l'erreur du second réseau (cela revient à le tromper), dit « discriminant » qui doit comparer l'image générée au jeu d'apprentissage afin de décider si elle a été synthétisée par le générateur, ou non.
La technique du *dropout* (Srivastava et al., 2014) consiste à amputer le réseau d'une partie de ses neurones lors de la phase d'apprentissage. Lors de chaque

itération, une fraction des neurones sont fixés à 0. Cela provoque deux effets : tout d'abord, cela empêche l'activation permanente de chaque neurone au détriment d'autres, ce qui réduirait le pouvoir de modélisation du réseau. Par ailleurs, l'information ne pouvant pas circuler dans l'ensemble du graphe, mais seulement sur un sous-graphe de celui-ci, le mécanisme de *dropout* permet de générer un réseau différent à chaque itération, plus petit que le réseau initial duquel on a désactivé un certain nombre de neurones. Grâce au *dropout*, le signal d'entrée est analysé par plusieurs graphes tirés aléatoirement, enrichissant cette analyse. Ce comportement n'est pas sans rappeler les approches ensemblistes qui mettent en jeu plusieurs classifieurs faibles conjuguant leurs efforts pour produire un classifieur robuste, à l'instar des *Random Forests*. En pratique, seuls les neurones activés à une itération donnée sont mis à jour, et les couches *fully connected* sont les seules à subir ce traitement pour deux raisons : celles-ci regroupent la majeure partie des paramètres du réseau global, et amputer les filtres convolutifs de certains neurones pourraient avoir un effet néfaste sur la structure spatiale des sorties des filtres (les cartes d'activation qui sont détaillées dans la suite).
Les mécaniques d'*augmentation de données* et de *dropout* ne sont effectives que pendant l'apprentissage. Cela est évident pour l'augmentation de données ; le dropout, quant à lui, doit être désactivé au moment de la prédiction, pour que celle-ci soit conduite sur le réseau à son plein potentiel, l'information circulant dans un maximum de neurones qui ont tous eu l'opportunité d'être sollicité au moment de l'entraînement. Enfin, la *batch normalization* (Ioffe et Szegedy, 2015) (normalisation par lot) est une technique visant à réduire l'effet de l'*internal covariate shift* en normalisant chaque activation du réseau. Le terme *covariate* dénote les données d'entrée au réseau ; avec l'adjectif *internal*, on désigne le signal transformé au sein du réseau. Entre deux couches, le signal s'éloigne pendant l'entraînement d'une distribution centrée et à moyenne nulle, même si les données d'entrée ont quant à elles été normalisées. Cela est dû à la présence des activations non linéaires. Cet *internal covariate shift*, même faible entre deux couches, peut conduire à des écarts importants en sortie du réseau. La normalisation par lot entre chaque couche permet de garder un flot d'information toujours à moyenne nulle et normalisé : le réseau n'est plus forcé à apprendre cet écart important entre distributions, réduisant les temps de calcul et améliorant les résultats. Notons également que les fonctions d'activation saturantes posent moins de problèmes, leurs arguments étant toujours normalisés les sorties sont moins sujets à des états de saturation. Finalement, ce mécanisme régularise le processus, étant moins sensible à la distribution des données d'entrée ; si des images représentant des objets, semblables sémantiquement, mais légèrement différent en radiométrie à ceux sur les images utilisés pendant l'entraînement, sont soumis au réseau, le réseau a de bonnes chances de détecter les objets en question.

En matière de télédétection, toutes ces pratiques sont mises en œuvre pour l'analyse d'images aériennes et satellites depuis quelques années. Les réseaux de neurones sont de nouveau utilisés depuis moins d'une dizaine d'années, les méthodes supervisées précédentes reposant principalement sur les SVMs ou Random Forests. Pourtant, la richesse spectrale d'une classe de

télédétection sur des images satellites est si diversifiée que la traduire, la compresser vers des attributs significatifs pour un classifieur est ardu. C'est pourquoi les réseaux de neurones sont tout indiqués pour **apprendre** ces attributs dans le même temps que l'on apprend le classifieur à répartir les pixels dans chacune des classes de la nomenclature considérée.

2.3 Télédétection et apprentissage profond

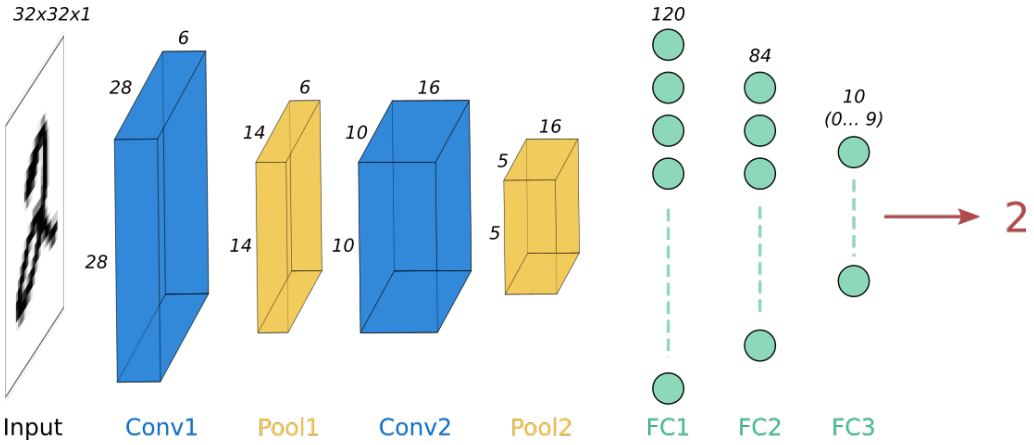


FIGURE II.9. Architecture employée par LeCun et al. (1989) pour la classification d'images de chiffres manuscrits (de 0 à 9).

Avant-propos sur les réseaux convolutifs

Avant d'aborder leur utilisation en télédétection, nous présentons ce que sont les réseaux convolutifs introduits par LeCun et al. (1989) pour la première fois (figure II.9). Prenons le cas où un MLP (*multi-layer perceptron*) est utilisé pour classifier une image quelconque. La première couche, voit ses neurones connectés à *chacun* des pixels de l'image : en découle un nombre très élevé de paramètres, $N \times P_1$ pour être précis, si N est le nombre de pixels dans l'image (généralement plusieurs millions) et P_1 le nombre de neurones dans la première couche. La seconde couche multiplie P_1 par P_2, nombre de paramètres dans cette dernière. En bref, le nombre de paramètres croît très rapidement et peut conduire à des problèmes de sur-apprentissage. Ce n'est toutefois pas l'unique raison qui rend les MLPs inadaptés aux images : ce type de réseau ne tient aucunement compte de l'information spatiale, pourtant très importante et structurante lorsque l'on analyse une image. Une image représente une scène du monde faisant face à l'objectif, or ce monde présente une continuité et une organisation particulière des objets. Cette information doit être utilisée pour optimiser l'interprétation qu'en fait le classifieur.

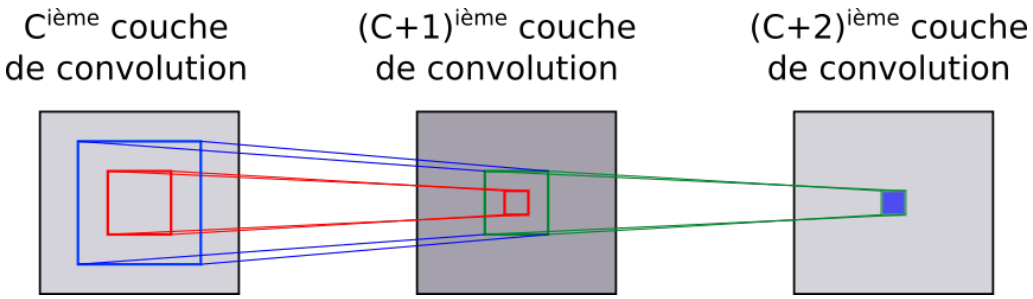


FIGURE II.10. A chaque nouvelle couche de convolution, les filtres « voient » de plus en plus de contexte issu de l'image originale.

Les filtres convolutifs sont parfaitement en adéquation avec ce besoin de décomposer le monde et son organisation, tout en préservant l'information spatiale dans les couches de convolution : une convolution 3×3 détecte des zones saillantes, et locales dans l'image, des contours d'objets par exemple. En enchaînant les convolutions au travers des couches successives, les filtres de convolution ont accès à une zone de l'image de plus en plus importante, fait illustré sur la figure II.10, renvoyant des informations de plus en plus globales sur l'image (notion de *receptive field* mentionnée précédemment dans la section 2.2).
En outre, chaque filtre de convolution utilise les mêmes paramètres sur l'ensemble de l'image, réduisant considérablement le nombre de poids à optimiser en comparaison d'un perceptron multicouches. De plus, cette contrainte de poids identiques à travers l'image pour un filtre donné assure une cohérence dans l'analyse de l'image par le réseau : le filtre effectue la même opération sur l'ensemble des pixels, avec les mêmes coefficients, mettant ainsi en évidence les mêmes motifs que ce filtre caractérise. Les objets à retrouver dans les images ne sont ainsi plus dépendants de leur localisation dans celle-ci.
Chaque couche du réseau C est composée de N_C filtres de convolution, dont les coefficients, on le rappelle, sont précisément les paramètres à optimiser. En sortie de chaque couche, on obtient ce qu'on appelle des cartes d'activations (on rencontre très souvent le terme de *feature maps* dans la littérature), chacune de ces cartes étant le produit de convolution de l'un des filtres de la couche avec l'image en entrée, de taille $nLin \times nCol$. On a donc N_C cartes d'activations, de taille $nLin \times nCol$ (en supposant que l'on utilise le mécanisme de *padding* pour conserver la taille de l'image initiale). Chaque convolution effectuée dans un réseau convolutif standard est en trois dimensions : on peut observer sur la figure II.11 que les opérations de convolution sont bien entendu spatiales d'une part, mais combinent également les cartes d'activations de la couche qui les précède. En reprenant les notations précédentes, si les filtres de convolution sont de taille 3×3 en 2D, alors les filtres de la couche $C + 1$ sont de taille $3 \times 3 \times N_C$.

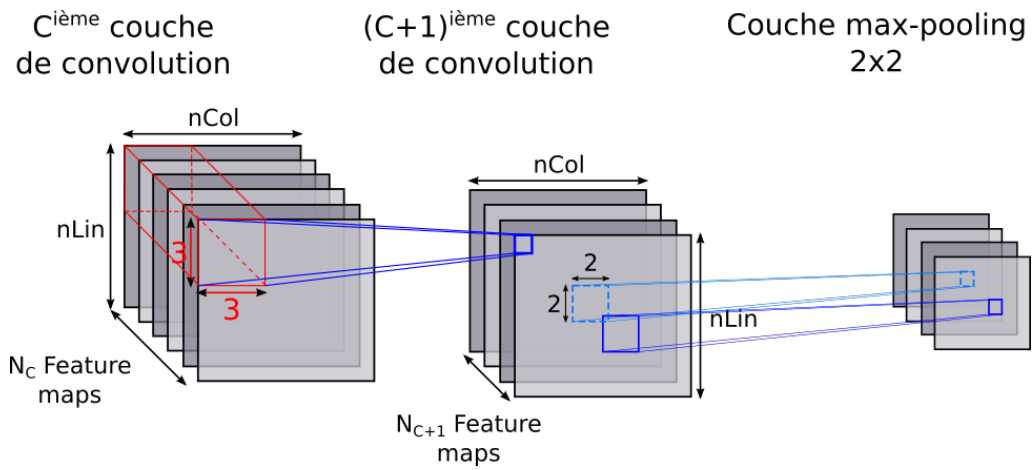


FIGURE II.11. Enchaînement de deux couches de convolution suivi d'une couche de pooling : les deux premières couches contiennent des filtres donc des paramètres à estimer, tandis que l'opération de pooling est libre de tout paramètre, mais offre des gains en temps de calcul et en performances.

Un autre type de couche apparaît sur la figure II.11. Le *max-pooling* (Boureau et al., 2010) est une opération sans paramètre à optimiser, mais qui a permet de réduire les dimensions spatiales des cartes d'activations (c'est pourquoi cette opération-ci est en 2D, sur chaque carte indépendamment les unes des autres). L'opération parcourt l'image avec un pas de deux pixels généralement (sans recouvrement) dans les deux directions, et ne conserve en sortie que le maximum parmi le voisinage 2×2 considéré. Le *max-pooling* a de bonnes propriétés telles que l'introduction d'une invariance locale en translation aux attributs appris par les filtres, la réduction du bruit, l'image étant sous-échantillonnée. Enfin, en réduisant les tailles d'image, on accélère sensiblement les processus d'entraînement. Une variante du *max-pooling* existe et ne conserve pas le maximum du voisinage 2×2, mais sa moyenne, d'où son nom d'*average pooling* (Gong et al., 2014).
Enfin, sont visibles sur l'architecture LeNet-5 de la figure II.9 les couches *fully connected* (entièrement connectées) reliant chacunes des sorties de la dernière couche de convolution à l'ensemble des neurones, cette fois-ci non convolutifs. Le terme *fully connected* renvoie justement à cette connexion complète entre chacun des neurones de deux couches successives, ce qui n'est en fait rien d'autre qu'un perceptron multicouches. L'architecture VGG de Simonyan et Zisserman (2014) comprend ainsi dans sa forme la plus profonde 135 millions de paramètres, dont environ 100 millions sont contenus dans ces trois couches seulement, prouvant numériquement que les perceptrons multicouches multiplient très rapidement le nombre de paramètres. L'utilisation récurrente de couches de *pooling* allège sensiblement le nombre de paramètres de ces couches entièrement connectées par rapport à une architecture conservant les dimensions de l'image initiale au travers des différentes couches de convolution ; les ultimes cartes d'activation contiennent moins de pixels que les images initiales, réduisant par la même le nombre de

connexions à la première couche du perceptron multicouches final.

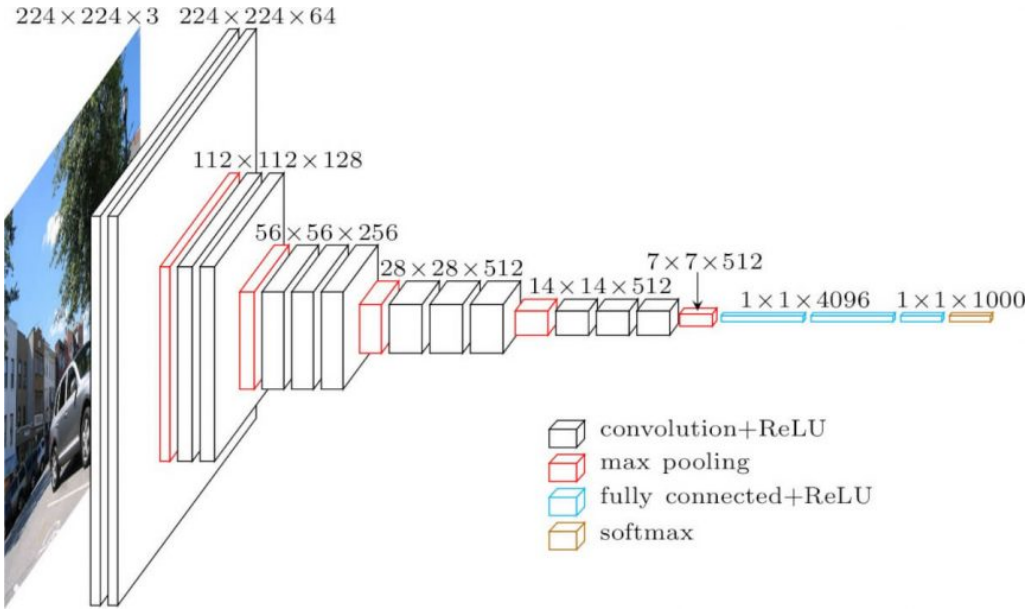


FIGURE II.12. Architecture convolutive construite par Simonyan et Zisserman (2014). Le papier s'intéressait à démontrer la croissance des performances de classification avec le nombre de couches. A titre d'information, l'architecture la plus profonde contenait 135 millions de paramètres environ.

Les réseaux de neurones convolutifs vus jusqu'alors emploient en fin d'architecture, et suite aux diverses couches de convolution et de *max-pooling*, ces couches dites *fully connected* décrites rapidement précédemment, qui concatènent l'ensemble de l'information contenue dans la dernière couche de convolution. Ce processus s'effectue avec **perte de la spatialisation de l'information**. Ces couches *fully connected* sont nécessaires (i) pour produire des attributs de haut niveau d'abstraction, et discriminants entre les classes, (ii) pour faire la passerelle vers l'opérateur *softmax* (la dernière couche du réseau LeNet-5 (figure II.9) correspond au score établi pour chaque classe par l'image en entrée, il y a donc 10 unités pour chaque chiffre de 0 à 9). Ces réseaux sont très utilisés en classification d'images.
Faisons ici un point de vocabulaire :

— la **classification** d'image renvoie à l'attribution d'un **label unique pour l'ensemble de l'image**. C'est le cas des jeux de données ImageNet (Deng et al., 2009 ; Su et al., 2012), Pascal VOC (Everingham et al., 2010), CIFAR-10 et CIFAR-100 (Krizhevsky, Hinton et al., 2009) ou MNIST (Deng, 2012) qui fournissent des jeux d'apprentissage sous la forme de couples (image, label), afin d'entraîner des algorithmes de reconnaissance d'images.

— En anglais, la **segmentation** d'images est dénotée *semantic segmentation* et consiste à attribuer cette fois-ci **un label par pixel de l'image étudiée**. Les jeux d'apprentissage nécessaires à l'apprentissage d'un modèle exécutant cette tâche sont sensiblement différents de ceux cités

précédemment, puisque les annotations doivent être faites au pixel et non à l'image.

La figure II.13 renvoie visuellement à la différence fondamentale des deux approches. Si les réseaux convolutifs sont adaptés dans les deux cas, une modification conceptuelle au niveau des couches *fully connected* vers des couches convolutives pour en faire des réseaux *fully convolutional*, offre de bonnes propriétés pour la segmentation d'image en particulier, comparé à leur architecture native. En cause, la perte de l'information spatiale citée est gênante pour la différenciation et la labellisation des objets *au sein d'une image*.

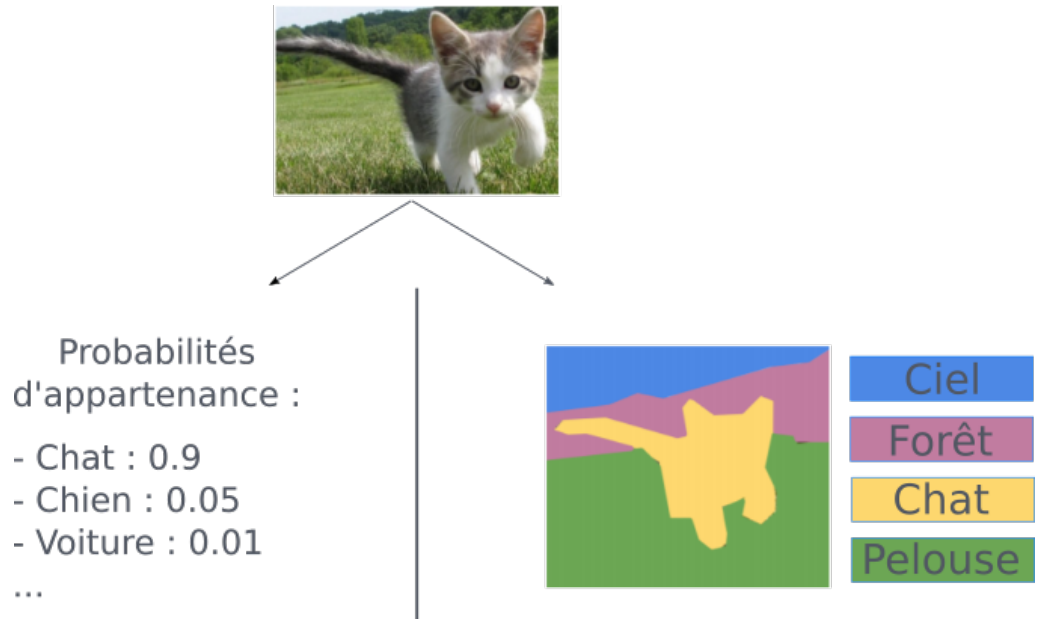


FIGURE II.13. Deux approches différentes : la classification (à gauche) et la segmentation, *semantic segmentation* (à droite). Source

En 2015, Long et al. (2015)) introduisent les *fully convolutional network*, ou réseaux entièrement convolutifs, qui se substituent peu à peu aux réseaux de neurones convolutifs utilisés jusqu'alors. A l'inverse des réseaux convolutifs dotés de couches *fully connected*, les réseaux entièrement convolutifs conservent l'information de localisation des premières couches jusqu'à la couche finale de classification. Les données d'entraînement étant labellisées *au pixel*, elles permettent d'apprendre au sein de chaque patch des arrangement spatiaux entre classes au niveau du pixel. Ces dernières délimitent notamment une géométrie plus précise des objets à classer. Ces approches denses reposent sur des architectures « encodeur-décodeur ». La première partie (*encoder*) permet de traduire l'information contenue dans l'image sous forme de vecteurs d'attributs de haut niveau, et la seconde (*decoder*) produit une carte de chaleur relative aux probabilités d'appartenance aux classes considérées, de résolution semblable à l'image d'origine, en utilisant l'information issue de l'encoder.

Dans notre cadre de travail, les réseaux de neurones convolutifs, qu'ils soient convolutifs ou entièrement convolutifs (nous parlerons de *FCN* dans la suite, en référence à *Fully Convolutional Network*), sont donc en parfaites adéquation avec les images satellites qui sont très fortement structurées *spatialement* au niveau des objets qui y sont représentés. Néanmoins si l'utilisation de ces réseaux doit être réfléchie, elle convient dans notre cas puisqu'on ne s'attache pas à analyser des séries temporelles, mais une zone géographique étendue pour laquelle nous n'avons qu'une mosaïque d'images. Les données temporelles ont été éprouvées à l'aide de réseaux de neurones, mais rarement convolutifs ou alors en les intégrant, comme une étape intermédiaire d'extraction et de sélection d'attributs pertinents, dans des processus bien plus adaptés pour le traitement de données présentant une *cohérence temporelle*.

Classification d'images aériennes et satellites

La communauté de télédétection s'intéresse depuis récemment aux CNNs pour les tâches de classification mono-temporelles d'images aériennes et satellites. La plupart des travaux de télédétection s'appuient sur le modèle *FCN* décrit précédemment. Les FCNs permettent une classification au pixel très précise, levant des ambiguïtés que l'approche par patch aurait, par exemple sur la distinction entre une voiture et le parking sur lequel elle se situe. L'utilisation jointe de l'information issue de l'image et d'un Modèle Numérique de Surface a été exploitée par Marmanis et al. (2018) en créant deux réseaux en parallèle, un pour chaque modalité, puis en les fusionnant à haut niveau d'abstraction. L'ajout de « skip-connections » permet de réinjecter de l'information haute fréquence sur la partie decoder pour retrouver la résolution initiale (réseaux Sharpmask ou RefineNet par exemple). Les mêmes auteurs se sont également efforcés d'améliorer la géométrie des objets détectés dans Marmanis et al. (2016).
En utilisant les mêmes données, Sherrah (2016) crée un FCN sans l'inconvénient du sous-échantillonnage dû aux couches de max-pooling, grâce à l'algorithme « a trous » (Chen et al., 2014a ; Yu et Koltun, 2015) qui remplace ces couches de max-pooling, préservant ainsi la dimension initiale de l'image. Les convolutions « a trous », aussi appelées *dilated convolutions*, peuvent être considérées comme des convolutions classiques sur une image deux fois moins résolues, en sautant un pixel sur deux au moment d'effectuer la convolution sur l'image à pleine résolution. Leur résultat est l'un des meilleurs sur les jeux de données de Potsdam et Vaihingen de l'ISPRS avec Volpi et Tuia (2017). Ces derniers utilisent un FCN avec une couche de déconvolution 3×3 en remplacement des couches fully-connected. Une comparaison entre approches « au patch » et *fully convolutional* montre que la seconde permet des temps de calculs au moment de la prédiction bien plus rapide qu'en utilisant une approche « au patch ».
Audebert et al. (2016) utilisent des données similaires dans le réseau SegNet (Badrinarayanan et al., 2015) pour effectuer une analyse multi-échelles dans

la partie decoder. L'utilisation de réseaux de neurones convolutifs nécessitant un volume d'apprentissage très conséquent, l'extraction automatique de jeux d'apprentissage issus de bases de données géographiques est devenue indispensable : dans Kaiser et al. (2017), les auteurs utilisent OpenStreetMap pour générer leurs données d'entraînement. L'utilisation massive de telles données permet de s'affranchir de paramétrages spécifiques quant à la génération automatique d'attributs (notamment de textures) à partir de ces données comme il a pu être nécessaire dans d'autres travaux (Gressin et al., 2013 ; Inglada et al., 2015 ; Maas et al., 2016 ; Pelletier et al., 2016 ; Dechesne et al., 2017).
Plus récemment, Chen et al. (2018b) utilisent le « shuffling operator », introduit par Shi et al. (2016) afin d'améliorer la détection d'objets de taille réduite en accroissant la résolution des couches en sortie du réseau. Le réseau utilisé s'appuie sur la mécanique « a trous » (*dilated convolutions*) (Chen et al., 2018c) permettant d'accroître le champ réceptif des filtres sans augmenter le nombre de paramètres pour autant. Chen et al. (2018a) améliorent le processus en utilisant conjointement des attributs géométriques issus d'un MNS et radiométriques, pré-calculés en amont du réseau « atrous ». Enfin, les processus de fusion multi-capteurs sont abordés dans (Audebert et al., 2018), où les différentes stratégies (fusion précoce ou tardive) apportent des avantages différents.

L'ensemble de ces travaux mettant en jeu l'apprentissage profond pour le calcul de carte d'occupation des sols contraste avec notre propre approche pour plusieurs raisons : l'utilisation de strictement plus de trois bandes spectrales n'a pas été explorée à notre connaissance en raison des réseaux qu'utilisent les travaux existants. Ceux-ci utilisent en particulier des architectures pré-entraînées (Simonyan et Zisserman, 2014 ; Szegedy et al., 2015) à l'aide d'images 3 bandes. Par ailleurs, le contexte dans lequel nous nous plaçons est éloigné des études citées précédemment d'un point de vue de l'objectif voulu : s'ils visent à améliorer plus ou moins sensiblement les classifications existantes sur des données issues de benchmark tels que les jeux de données de Potsdam et Vaihingen proposés par l'ISPRS, le but de ces travaux de thèse est de fournir des méthodes pour que le calcul de carte d'occupation des sols soit opérationnel et ainsi déployable à large échelle.

En revanche, suite aux expérimentations mise en œuvre pendant les travaux de thèse, l'IGN a jugé que des tests plus approfondis pour des perspectives de production de cartes d'occupation des sols étaient nécessaires. Nous avons donc, au sein d'une petite équipe, menés des tests en utilisant cette fois-ci des architectures type U-Net (Ronneberger et al., 2015) pour des classes d'objets particulières et cruciales dans un contexte de préservation de l'environnement. Nos résultats concordaient avec ceux mis en avant plus tard par Huang et al. (2018). S'efforçant de se prononcer sur les réseaux à utiliser pour mener à bien une classification d'images aériennes ou bien satellites, les auteurs s'accordent sur le fait que U-Net sert de baseline à de nombreux papiers, lui apportant quelques modifications parfois, mais affichant

toujours de très bons résultats. Nous reviendrons sur l'architecture U-Net à ce moment-ci.

FIGURE II.14. Base de données géographiques utilisée pour l'évaluation des modèles par validation croisée :
● Bâti, ● Route, ● Culture, ● Végétation, ● Eau.

3 Evaluation de classification - métriques

Les résultats obtenus ont été qualifiés par validation croisée avec les bases de données géographiques existantes (Figure II.14), qu'elles soient constituées à l'IGN (RGE) ou non (RPG). Il s'agit de la vérité terrain la plus exhaustive à disposition même s'il faut bien indiquer qu'elles comportent un certain taux de fausses étiquettes (erreurs ou changements depuis la génération de ces bases) (Foody, 2010). Les polygones de la base de données ont subi une érosion d'un pixel pour s'affranchir des incertitudes liées aux frontières des objets. L'évaluation porte donc davantage sur les pixels situés à l'intérieur des objets : plutôt que de se concentrer sur les frontières des objets, on cherche à quantifier le nombre d'objets bien retrouvés dans l'image.
Il faut bien comprendre qu'un algorithme de classification automatique nécessite effectivement une phase d'entraînement, mais la validation est cruciale : à ce titre, lorsque l'on construit un tel algorithme, on doit s'assurer de séparer les données labellisées en deux sous-ensembles, n'en utilisant qu'un pour la phase d'apprentissage afin de garantir une validation non biaisée, et donc une bonne estimation de la capacité du classifieur à généraliser à de nouvelles données.
Pour évaluer les différents résultats, plusieurs métriques ont été utilisées. Nous plaçant dans un cadre supervisé, il est possible de comparer directement les objets détectés ou non par rapport à ceux présents dans les bases de données géographiques. Concernant les métriques, elles sont généralement dérivées de la matrice de confusion confrontant directement les résultats prédits par le réseau à ce qu'il en est réellement. Cette matrice de confusion est construite en comparant, pixel à pixel, le résultat du modèle à la connaissance

que l'on a *a priori* de la classe de chacun de ces pixels. En reprenant la notation précédente, les $(y_i)_{i\in\{1..n\}}$ représentent les **classes réelles** de n pixels, parmi C classes tandis que les $\hat{y}_{i\,i\in\{1..n\}}$ correspondent aux **classes prédites** par le modèle pour ces mêmes n pixels. La matrice de confusion M est de taille $C \times C$, et est mise à jour de la manière suivante :

— si $\hat{y}_i \neq y_i$, la classe prédite pour le pixel i n'est alors pas la classe réelle, on ajoute $+1$ au coefficient à la coordonnée de la classe prédite en colonne, et de la classe réelle en ligne.

— si $\hat{y}_i = y_i$, la prédiction est correcte, $+1$ au terme diagonal correspondant à la classe en question.

La figure II.15 schématise la logique de lecture de la matrice de confusion dans un cas binaire, où chaque pixel peut appartenir à une classe exactement parmi $\{n, p\}$ (positif, négatif). Dans ce cas, quatre cas sont possibles pour chaque pixel i :

1. $\hat{y}_i = p$ et $y_i = p'$: dans ce cas, la classe prédite est égale à la classe réelle, on parle de pixel « Vrai Positif (VP) » ;
2. $\hat{y}_i = p$ et $y_i = n'$: on classe par excès p au lieu de n', le pixel i est un « Faux positif (FP) » ;
3. $\hat{y}_i = n$ et $y_i = p'$: on omet p' en classant le pixel i en n, ce pixel est un « Faux Négatif (FN) » ;
4. $\hat{y}_i = n$ et $y_i = n'$: classe prédite et classe réelle sont en accord, le pixel i est un « Vrai Négatif (VN) ».

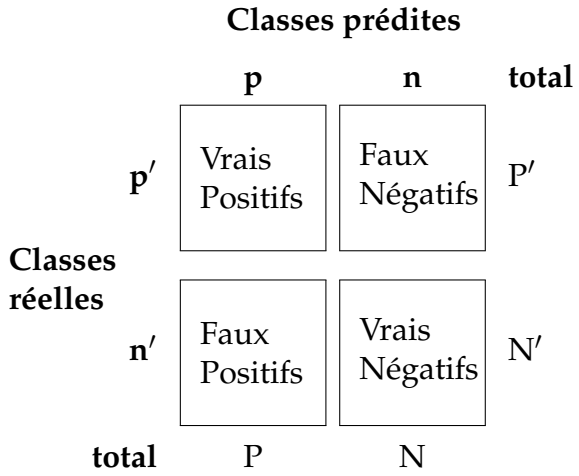


FIGURE II.15. Matrice de confusion dans un cas binaire

La matrice de confusion peut évidemment être étendue au cas multi-classes. De la matrice de confusion calculée par comparaison pixel à pixel peuvent être dérivés un certain nombre d'indicateurs. Ces indicateurs peuvent être révélateurs de la bonne classification de l'ensemble des classes ou de chaque classe indépendamment des autres. Les indicateurs utilisés pour quantifier la **bonne classification de chaque classe** sont les suivants :

— la **Précision** : $\frac{VP}{P}$. Cette mesure indique pour chaque classe si la détection effectuée par le classifieur est « pure » (ou exact), c'est-à-dire si les faux positifs sont nombreux (des pixels classés *à tord* dans la classe considérée) ou non. Une précision élevée (proche de 1) signifie que l'ensemble des pixels classés dans une classe donnée appartiennent effectivement à cette classe ;

— le **Rappel** : $\frac{VP}{P'}$. Le rappel renvoie à la capacité du classifieur d'être exhaustif dans sa détection des pixels d'une classe donnée. Un rappel élevé (proche de 1) signifie que l'ensemble des pixels appartenant à la classe considérée ont bien été classés comme tel (peu de faux négatifs) ;

— le **F-score**, ou F1-Score ou coefficient de Sørensen–Dice, combine rappel et précision selon la formule suivante (moyenne harmonique) :

$$Fscore = \frac{2}{\frac{1}{precision} + \frac{1}{rappel}}$$

Le F-score est proche de 1 si, pour une classe donnée, l'ensemble des pixels classés comme tels (i) le sont à juste titre et (ii) correspondent à tous les pixels de cette classe.
Obtenir une précision et un rappel proches de 1 étant l'objectif, il n'en demeure pas moins ardu. C'est pourquoi il faut généralement, à la définition du problème, déterminer s'il est plus important de trouver l'ensemble des objets d'une classe, quitte à commettre des sur-détection (faux positifs), ou au contraire de ne récupérer des objets dont on est sûr de la classe prédite, quitte à en omettre (faux négatifs). Dans le premier cas, on favorise alors un score de rappel élevé, dans le second la précision. Si on sort de l'analyse d'images, un cas pratique est la détection de fraude bancaire : il est « courant » qu'un compte bancaire soit bloqué parce qu'une transaction jugée frauduleuse a été détecté par un algorithme de détection d'anomalies. On favorise dans ce cas le rappel sur la précision pour s'assurer que l'ensemble des transactions réellement frauduleuses soient retenues, avec un effet collatéral sur des transactions elles, non frauduleuses.

— une mesure proche du F-Score existe, il s'agit de l'**Intersection over Union**, abrégé en IoU, ou encore appelé Indice de Jaccard, qui donne davantage d'importance aux vrais positifs par rapport au Fscore :

$$IoU = \frac{VP}{VP + FP + FN}$$

En segmentation sémantique (classification au pixel), cela revient à calculer le rapport de la surface prédite correctement par le classifieur sur la somme de la surface totale prédite et de la surface omise (extraites grâce à la vérité terrain).

Après avoir passé en revue les métriques « par classe », on dresse la liste des indices permettant de juger globalement du résultat de classification sur le jeu de donnée de validation :

— **Overall Accuracy** (OA) : $\frac{VP+VN}{P'+N'}$. L'OA donne une indication très globale de la qualité de la classification en comparant le nombre d'objets bien classés, toutes classes confondues, au nombre total d'objets. Si elle permet une première appréciation des performances du classifieur, elle n'en reste pas moins superficielle. Cela est d'autant plus vrai dans des tâches où les classes en question sont fortement déséquilibrées en terme de représentation. Par exemple en imagerie satellite, la classe relative à l'ensemble des parcelles agricoles est écrasante en nombre de pixels par rapport à une classe plus rare telle que les routes ou cours d'eau. Ainsi, derrière une OA de 95% sur 100 pixels répartis entre les classes *champs* et *routes* peuvent en réalité se cacher 95 pixels de *champs* bien classées, et 0 de *route*.

— l'**indice Kappa** κ (Pontius Jr et Millones, 2011) :

$$\kappa = \frac{P_0 - P_e}{1 - P_e},$$

avec :

— P_0 la proportion de pixels bien classés par rapport au total des pixels, autrement dit, l'overall accuracy ;

— $P_e = \frac{P \times P' + N \times N'}{(P+N)^2}$, quantifiant l'accord aléatoire entre classes prédites et classes réelles des pixels.

κ mesure donc l'accord entre données prédites et données réelles, par rapport à une assignation au hasard des classes. Un *kappa* proche suggère un accord total entre vérité terrain et prédiction tandis qu'un *kappa* négatif oppose totalement les deux, $kappa = 0$ révélant que les pixels bien classés ne le sont uniquement que par hasard.

— les mesures de rappel, précision, F-Score et IoU peuvent être rendues globales en calculant la moyenne de chacune de ces grandeurs sur l'ensemble des classes.

Chapitre III

Investigations sur la cartographie d'Occupation des Sols par apprentissage profond sur images satellites très haute résolution

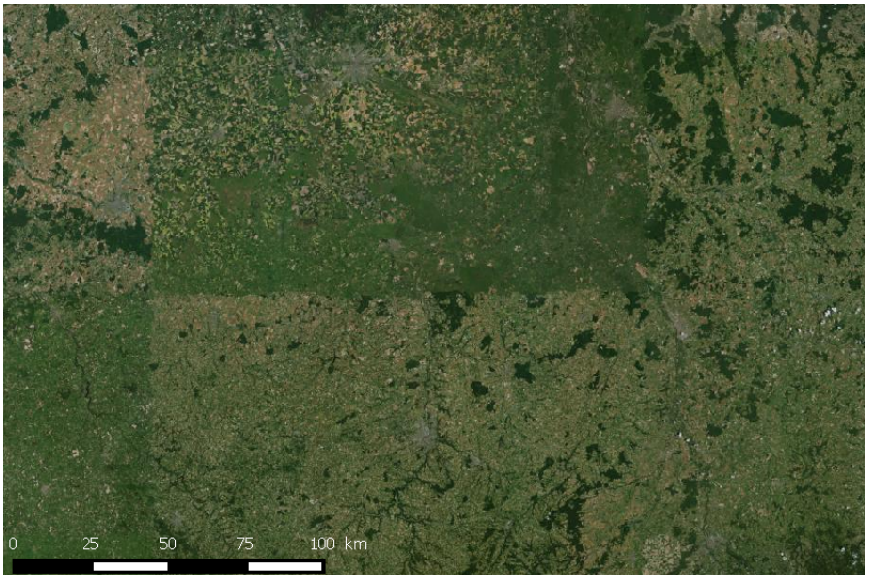
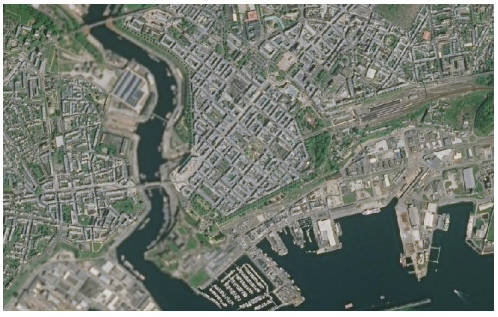


FIGURE III.1. Effet du phénomène de diachronie : les images d'une converture étant acquises sur l'ensemble de l'année, deux régions voisines présentent des réflectance différentes pour une même classe.



(A) Paysage urbain sur l'agglomération de Brest : paysage urbain moyennement dense, toits de zinc ou d'ardoise, eau très sombre (eau profondes).

(B) Paysage urbain sur l'agglomération de Bordeaux : paysage urbain très dense, toits de tuile, eau très turbide et "claire" (eaux peu profondes, courant important de la Garonne).

FIGURE III.2. A résolution égale (**1 cm <-> 230 m**), deux villes peuvent être très différentes telles que représentées par les images satellites.

1 Objectif d'investigation

Les travaux exploratoires mentionnés dans ce chapitre ont un but exploratoire avant tout. Disposant d'une couverture annuelle acquise par les capteurs embarqués sur les satellites SPOT 6/7, les possibilités offertes par ces images spatiales très haute résolution en matière d'occupation des sols sont multiples, mais nécessitent des investigations d'ordre méthodologique, notamment en ce qui concerne l'automatisation des processus de labellisation.

SPOT 6/7 délivrent des images à 1,5 m de résolution, permettant une analyse fine du territoire géométriquement, par rapport à des acquisitions Sentinel-2. Les travaux portant sur l'apprentissage automatique mentionnés dans le Chapitre II utilisent pour la plupart des données d'une modalité supplémentaire, en particulier une information d'altitude en chaque pixel. Si le Modèle Numérique de Terrain (MNT) rencontre peu de modifications au cours des années, cela est différent pour le Modèle Numérique de Surface correspondant (MNS), très utile au demeurant pour différencier le sol du sursol. Or, cette donnée n'est pas cohérente *a priori* avec la couverture monodate annuelle de tout le territoire, en plus d'être une donnée onéreuse à l'acquisition, que ce soit par LiDAR ou par photogrammétrie optique.
Des données d'une modalité différente existent et ont été mentionnées, à l'image d'OpenStreetMap. Bien que loin d'être dénuées d'intérêt, ces données n'ayant pas été qualifiées par l'IGN, et ne disposant pas de spécifications établies selon des procédures unifiées (fiabilité relative au crowd-sourcing), nous ne les utilisons pas non plus dans le cadre, qui sert des objectifs futurs opérationnels, qui est le nôtre. D'autant plus que, avec l'idée de détecter éventuellement plus tard des changements, il faut pouvoir comparer des bases de données maîtrisées et existantes entre deux dates. Les travaux présentés ici sont menés par analyse multispectrale uniquement, en utilisant les quatre canaux disponibles sur les images de SPOT 6/7, Rouge, Vert, Bleu et Infra-Rouge.

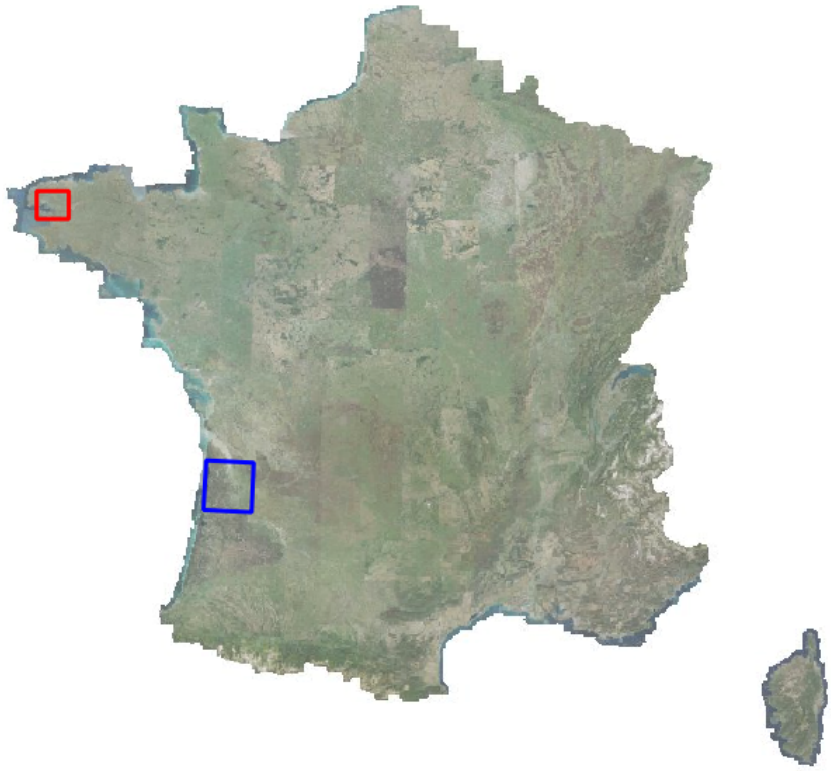


FIGURE III.3. Différentes régions avec différentes thématiques, liées à l'activité humaine et au climat environnant : en rouge, la Bretagne, région côtière dont l'activité est essentiellement liée à la pêche ou à l'agriculture. En bleu, la région bordelaise dont l'activité est portée sur la foresterie, avec des essences d'arbre différentes de la Bretagne. Les bâtiments des deux régions présentent également des différences vu de l'espace.

Dans le but d'obtenir une méthode de traitement pouvant être appliquée dans des conditions matérielles sans prétention, c'est-à-dire sur une machine standard de bureau, dotée d'une carte graphique pour l'apprentissage du réseau et son exécution dans les phases de prédiction, le réseau utilisé présente moins de couches que les réseaux de la littérature. Toutefois, deux autres raisons majeures justifient ce choix. Premièrement, les investigations menées, à l'inverse de l'état de l'art mentionné précédemment, portent davantage sur la dimension télédétection de la thèse. Notamment, les problèmes liés aux diachronies sont un véritable écueil et nécessite des stratégies particulières, deux images adjacentes sur une région pouvant avoir été acquises à des périodes de l'année bien différentes, comme l'illustre la figure III.1. Les régions étudiées étant vastes, il est également nécessaire de résoudre les problèmes de nomenclature variable selon ces régions (figure III.3), et d'aspects variables des objets lorsque l'on observe deux zones géographiques différentes. La figure III.2 met en opposition les apparences des villes de Bordeaux et de Brest, toutes deux très peuplées, mais caractérisées différemment selon l'organisation urbaine, les matériaux de construction utilisés, etc.
Toutefois, afin de conserver une cohérence sémantique dans le chapitre, les classes considérées sont au nombre de 5 : *bâti*, *routes*, *végétation*, *cultures*, *eau*.

Nous rediscuterons plus avant de ce choix dans la section 2.2, détaillant le processus d'extraction des jeux d'apprentissage. Un cas spécifique de nomenclature élargie sera étudié, on en précisera la portée des classes dans le paragraphe qui y est consacré.

L'ensemble des algorithmes développés dans ce chapitre utilisent les bibliothèques Torch puis Pytorch pour les opérations liées aux réseaux de neurones. En terme de *hardware*, la machine était dotée de 16 Go de RAM et d'une carte graphique Nvidia GTX 980 dans un premier temps puis d'une carte graphique Nvidia GTX 1080 Ti.

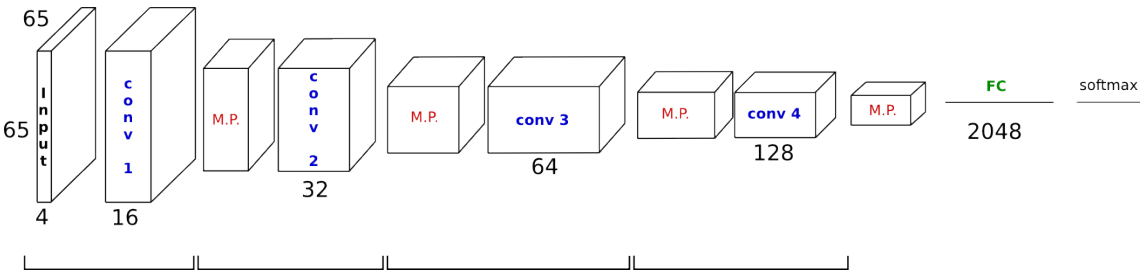


FIGURE III.4. Architecture employée.

2 Architecture et jeu d'apprentissage

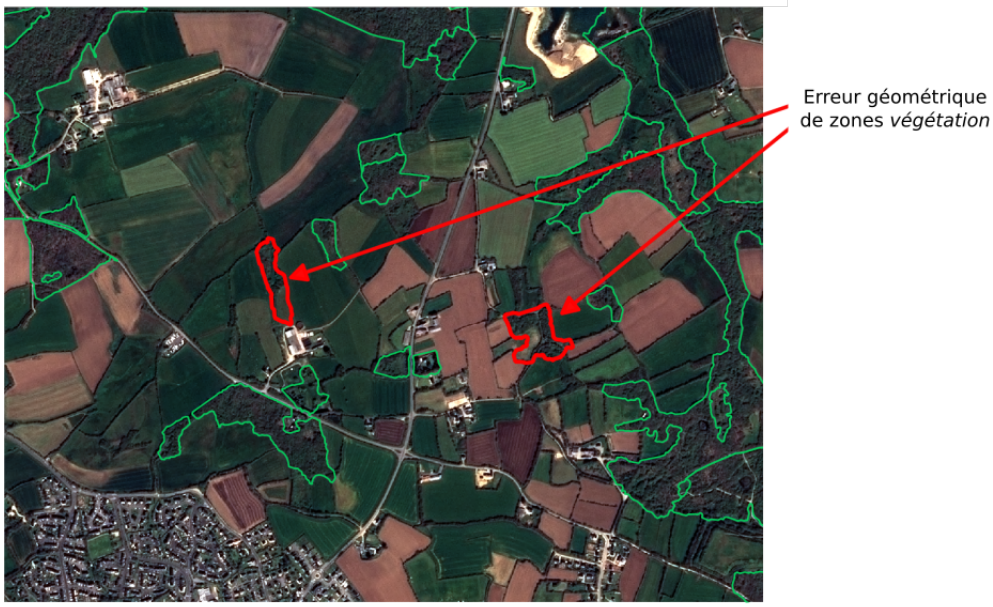
2.1 Réseau de neurones pour de la classification d'OCS

Afin de répondre aux besoins opérationnels de ce travail de thèse, une architecture légère est mise en place. Celle-ci est visible en figure III.4. L'approche au « patch » permet de construire des types de réseaux qui étaient jusqu'à il y a près de trois ou quatre ans très répandus pour la classification d'objets dans les images, et donc pour lesquels de bonnes pratiques existaient alors. Les travaux et résultats de ce chapitre ont été produits sur une période s'étalant d'il y deux à quatre ans. En revanche, classifier des images à très large échelle peut être lourd en temps de calcul, surtout sur une machine dont les spécifications techniques sont mentionnées précédemment. Ceci explique pourquoi nous contraignons le réseau à être moins dense que des architectures déjà existantes.
Les patches en entrée du réseau doivent avoir une taille fixe : pour décider de cette taille, nous avons dû prendre en compte le fait que l'on cherche aussi bien à classer de petits objets (bâtiments) que des objets étendus (champs, forêts). Une fenêtre de $65 \times 65 \times 4$ (on utilise les quatre bandes disponibles) pixels est un compromis qui permet de prendre en compte les classes d'intérêt dont certains objets seraient petits (bâtiment isolé en milieu rural). Cela correspond à une emprise au sol d'un peu moins de 100 m de côté (65×1.5). Successivement, des convolutions opèrent sur les couches précédentes. On accroît alors l'abstraction et la complexité des représentations lorsqu'on s'éloigne de l'image d'entrée. Chaque couche est constituée de filtres de convolution

3×3, auxquels succèdent des fonctions d'activation non linéaires, prenant ainsi en compte le caractère non linéaire du problème (Zhu et al., 2017). La fonction d'activation choisie est celle commune à toutes les architectures récentes, le « Rectified Linear Unit » (ReLU). Les couches de convolution sont séparées par des couches dites de « max-pooling 2×2 » (MP). En fin de réseau, une couche « fully connected » combine l'ensemble des filtres de la couche précédente pour créer des attributs à haut niveau d'abstraction à partir du patch en entrée. Enfin, le critère de « cross-entropy » à minimiser est suivi d'un « softmax » pour rendre les scores homogènes à des probabilités.

Nous l'avons souligné dans le chapitre II, et l'avons explicité dans le paragraphe précédent : le réseau est inédit en le sens que l'**on utilise bien les quatre bandes disponibles des images SPOT 6/7**, pour des raisons de richesses spectrales à défaut d'avoir des informations exogènes (altitude, cartographie crowd-sourcée). Il est donc nécessaire de construire notre propre jeu d'apprentissage, en adéquation avec les caractéristiques du réseau mis en place.

(A) Erreur de sémantique dans la BD TOPO, probablement en raison d'un défrichement au profit de l'étalement urbain.



(B) Erreur de géométrie dans la BD TOPO *végétation*, probablement due à une évolution du paysage et de la zone qui s'est vue amputer d'une partie de sa forêt.

FIGURE III.5. Les bases de données géographiques sont imparfaites.

2.2 Constitution du jeu d'apprentissage

La constitution des données qui permettront l'apprentissage se fait en exploitant les bases de données géographiques nationales existantes. Une stratégie alternative consiste, en l'absence de données de référence suffisantes, à générer synthétiquement ces données (Kemker et al., 2018).
Dans notre cas de figure, ces bases sont très massives et mettent à notre disposition de nombreux polygones appartenant aux classes d'occupation des sols que l'on cherche à discriminer. Elles sont géo-référencées, se superposant aux images SPOT parfaitement. La précision des polygones est de 1 m, ce qui est compatible avec la résolution spatiale de nos images. Une imprécision géométrique sur la base de données aura pour conséquence une erreur, au plus, inférieure au pixel sur la cartographie finale. En revanche, des erreurs sémantiques et/ou géométriques peuvent apparaître car ces bases de données ne sont pas nécessairement à jour sur l'ensemble du territoire, en atteste la figure III.5. L'utilisation de CNNs pour apprendre massivement sur ces bases de données permet de réduire l'influence de ces erreurs lors de la phase d'entraînement. Pour constituer nos patches d'apprentissage ($65 \times 65 \times 4$), on sélectionne régulièrement sur la zone d'apprentissage des pixels autour desquels sont extraits leur voisinage 65×65.
Dans un souci de généralisation, les échantillons subissent aléatoirement des transformations afin (i) de limiter le sur-apprentissage, (ii) d'obliger le réseau à modéliser certaines invariances. Cela correspond au mécanisme d'augmentation de données, abordé dans le chapitre II. Ainsi, les images étant acquises depuis des satellites, une invariance aux rotations par rapport au nadir d'acquisition est nécessaire dans la modélisation de nos classes. Les images subissent également une transformation de symétrie aléatoire. Rotations et symétries sont composées, de manière « online » : cela veut dire que ces transformations sont opérées sur chaque image d'un batch d'apprentissage avant son passage dans le réseau. Cela permet d'accélérer les temps de calcul, évitant les accès disque, et améliore la variété de ces transformations, étant différentes à chaque époque d'apprentisssage.
Ces classes ont été énumérées dans la description de l'objectif de ce chapitre. Au nombre de 5, elles varient toutefois grandement d'une région à une autre, compte tenu des spécificités locales de construction des bâtiments par exemple. Le choix de ces classes a été motivé principalement sur la base de deux arguments :

— ces classes regroupent l'essentiel du territoire d'un point de vue sémantique, à part les régions très minérales (paysage montagneux), sur lesquelles peu d'annotations ont été faites. Or, le but ici est de proposer un outil supervisé par réseau de neurone dans des objectifs de compatibilité avec les bases de données IGN. La classe *route* est une classe habituellement difficile à détecter par réseau de neurones, et même si la BD TOPO à l'IGN présente un squelette du réseau routier maintenu très régulièrement, il est intéressant d'étudier les capacités d'un modèle profond à détecter cette classe sur des images satellites ;

— nous le verrons sur les expérimentations d'élargissement de la nomenclature à une autre classe, une approche hiérarchique peut être conduite afin d'améliorer la granularité sémantique de l'OCS, tout en partant d'un modèle entraîné sur les cinq classes principales. En utilisant ces classes-ci, on peut en dériver des sous-classes, à la manière de Corine Land Cover par exemple, ou du RGE en constitution à l'IGN.

Nous présentons sur la figure III.6 des échantillons d'apprentissage construits selon le protocole décrit précédemment. Pour chaque classe, on constate une variabilité importante des objets représentant ces classes (différents types de bâtiments, végétation à différents stades de croissance, turbidité changeante de l'eau), motivant davantage l'utilisation des CNNs pour l'extraction et la sélection des attributs qui seront le plus à même de discriminer les classes entre elles.

FIGURE III.6. Patch d'apprentissage centré en un pixel de la classe considéré, dans l'ordre de gauche à droite et de bas en haut : *bâti*, *routes*, *végétation (forêt)*, *cultures*, *eau*.

2.3 Stratégies d'entraînement

Le réseau décrit précédemment, en section 2.1, est entraîné sur les données d'apprentissage construites selon section 2.2 sur des zones géographiques spécifiques. Deux stratégies (complémentaires) ont été considérées :

— le réseau RWI pour « Random Weight Initialization », pour laquelle les paramètres des CNNs sont initialisés aléatoirement :

— le réseau FT pour « Fine-Tune », dont les paramètres sont copiés depuis un réseau existant.

Dans les deux cas, l'architecture considère, en pratique, des batchs de 200 échantillons d'apprentissage en entrée, pour accélérer la convergence et les temps de calcul.

Random Weight Initialization (RWI)

Aucun réseau existant n'accepte, à notre connaissance, en entrée, quatre bandes RVB-IR. L'entraînement d'un réseau *from scratch* est donc obligatoire en passant par une initialisation aléatoire des paramètres de notre architecture. Le pré-entraînement de réseau type VGG (Simonyan et Zisserman, 2014) fait sur des images trois bandes peut être facilement transféré vers des images quatre bandes **en terme d'implémentation**. Toutefois, rappelons-le, un réseau convolutif est une succession de convolutions dont chacune est tributaire de la ou les convolution(s) précédente(s). En conséquence, toutes les convolutions d'un réseau sont fortement dépendantes de la première couche, qui elle, varie selon le nombre de bandes de l'image d'entrée.
S'il est facile d'ajouter des poids aléatoires correspondant à l'ajout d'une bande sur l'image d'entrée d'un réseau pré-entraîné sur des images trois bandes, un profond déséquilibre du réseau s'ensuivrait puisque le reste des paramètres sont, quant à eux, bien appris pour ce type d'image. D'un point de vue conceptuel, il est donc déraisonnable et incohérent d'envisager cette approche pour utiliser des images quatre bandes. D'où l'approche *RWI* dans un premier temps.

Fine-Tuning (FT)

De nombreuses architectures ont été largement entraînées sur des bases d'images trois canaux RVB très volumineuses, offrant ainsi l'occasion d'utiliser des paramètres pré-calculés et très robustes. Cependant, des architectures telles que (Simonyan et Zisserman, 2014 ; Szegedy et al., 2015) possèdent de très nombreux hyperparamètres et y sont très sensibles. Elles peuvent rendre leur utilisation difficile lorsqu'elles sont utilisées dans des cas éloignés de leurs cas d'utilisation premiers (Papadomanolaki et al., 2016).
Affiner (ou « fine-tuner ») un réseau existant repose sur le même principe itératif que pour le *RWI*, sauf au moment de l'initialisation, pour laquelle on récupère des paramètres pré-entraînés, que l'on affine sur le nouveau cas à traiter. En effet, de nouvelles données peuvent présenter des spécificités absentes du jeu d'entraînement sur lequel un réseau à été précédemment entraîné. L'idée est d'absorber ces nouvelles spécificités en ajoutant des données d'apprentissage du nouveau problème à un réseau qui convergeait déjà sur un autre jeu. Les couches les plus profondes sont souvent les seules à être *affinées* : le travail de Zeiler et Fergus (2014) a montré que les couches les plus proches de l'image extraient des attributs bas niveaux, tels que des contours, ou coins. Ils sont communs à l'ensemble des images que l'on peut trouver. En revanche, les couches les plus profondes modélisent des caractéristiques de

plus en plus dépendantes du cas d'utilisation. Ce constat permet ainsi, pour des réseaux très profonds, de ne ré-entraîner que les couches les plus profondes, accélérant les temps de calcul, et réduisant grandement le nombre de paramètres.
Le *fine-tuning*, *FT*, facilite grandement la convergence sur de nouvelles zones ou de nouvelles époques d'acquisition, tout en réduisant le nombre d'échantillons nécessaires à l'entraînement et le nombre d'itérations.

3 Classification sur zone géographique étendue

3.1 Choix des régions d'intérêt

Les images ont été acquises par les capteurs des satellites SPOT 6/7. Elles et sont mises à disposition par le pôle de données THEIA. Nous disposons d'une image par année et par zone. Les quatre bandes, à 1.5 m une fois le processus de pan-sharpening effectué, sont utilisées pour extraire le plus d'information des images (ne disposant pas d'information d'altitude, nous souhaitons tirer parti du maximum d'information radiométrique disponible dans les images satellites).
L'ensemble des tests décrits dans la suite ont été effectués sur deux zones du territoire français, présentant des paysages différents. Cette configuration autorise l'étude de la viabilité d'usage d'un modèle d'une zone vers une autre. Les zones étudiées sont les suivantes (localisées par la carte III.3) :

i Une partie du **département du Finistère** a été choisie pour sa diversité paysagère, une image SPOT 6/7 de 2014 sur la figure III.7 est visible, avec, en rouge, les limites de la zone d'étude. Avec des paysages côtiers très découpés, la zone comporte plusieurs types d'écosystèmes, pouvant être très urbanisés comme c'est le cas avec l'agglomération de Brest, ou bien très ruraux.
Deux sous-régions sont d'ailleurs considérées plus particulièrement pour entraîner les réseaux et étudier leur capacité à bien généraliser sur de nouvelles données et paysages : autour de la ville de Brest (notée ROI-1, milieu urbain) et de Le Faou (notée ROI-2, milieu rural), dont les images sont fournies sur la figure III.8. En plus de l'impact d'un changement géographique, différentes dates sont considérées : 2014 et 2016. Ce choix sur l'étude d'un changement temporel permet de jauger le potentiel du modèle proposé dans le cadre de production de cartes d'occupation des sols à diverses dates pour de futures analyses (détection de changement, évolution des forêts par exemple).

ii Le **département de la Gironde** est la seconde région que l'on a désignée d'intérêt car elle présente, à l'instar de la première zone du Finistère, plusieurs aires urbaines, notamment celle de Bordeaux, mais aussi des paysages plus naturels. Toutefois, nous l'avons en introduction de ce chapitre, l'aspect visuel des toits des habitations diffèrent de celui de Bretagne ; en plus de cela, les surfaces naturelles sont en majorité liées à l'exploitation forestière avec de nombreuses pinèdes à l'Ouest

de la zone. Enfin, même si là encore, le département est en littoral de l'Atlantique, l'estuaire de la Gironde présente un aspect de l'eau très particulier en raison des conditions d'écoulement de la Garonne et de la Dordogne. Ces deux cours d'eau appporte de leur côté de nombreux sédiments, tandis que les marées favorise l'apparition de bancs de sable, donnant à l'estuaire cette aspect très turbide en raison des particules en suspension. Cette turbidité donne une texture à l'estuaire qui contraste fortement avec l'eau présente sur le Finistère, essentiellement partie de l'océan.

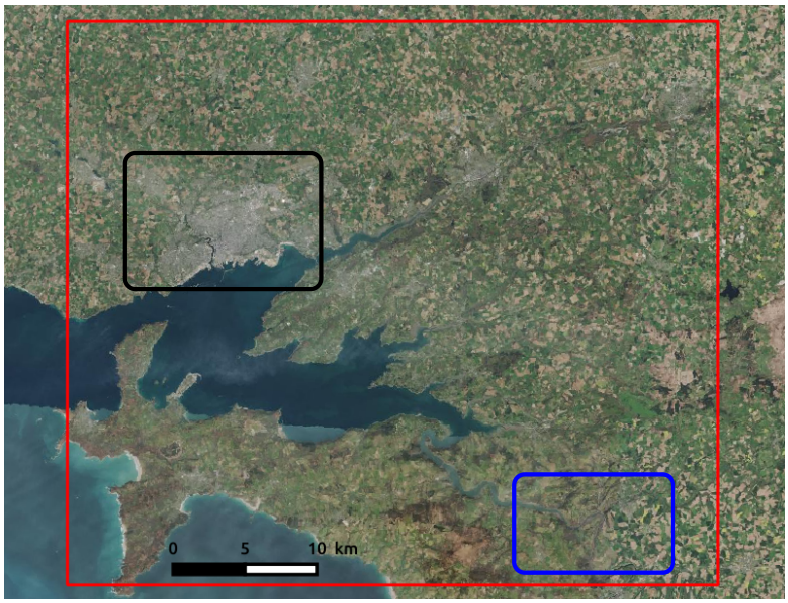
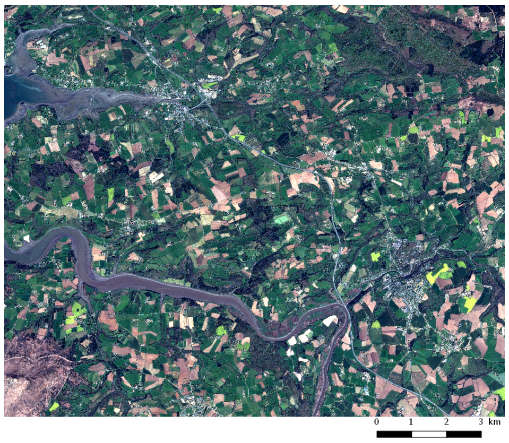


FIGURE III.7. Région proche de Brest, Finistère.
En rouge, l'agglomération de Brest, en bleu la zone du Faou.

(A) Aire urbaine de Brest, à forte densité de population : surface artificielle majoritaire.



(B) Zone rurale autour du Faou, peu peuplée : surface naturelle majoritaire.

FIGURE III.8. Deux paysages d'étude différents en Bretagne.

FIGURE III.9. Région proche de Bordeaux, Gironde (les découpes sur la gauche de l'image correspondent aux limites de la couverture SPOT annuelle sur la France.

La zone autour de Brest permet d'étudier différentes stratégies d'apprentissage par évaluation des performances sur des données décalées dans le

temps, ou sur des écosystèmes opposés. Le département de la Gironde à pour but de décrire une approche pour la classification de zones étendues à partir de réseaux initialement appris sur une autre région, qui a des caractéristiques paysagères très dissemblables.

3.2 Comportement de réseaux *RWI* et *FT* sur des données « non vues »

Motivations : la nécessité de stratégies d'optimisation pour la classification à large échelle géographique

Pour envisager la classification de zones géographiques à l'échelle d'une partie d'un département ou d'un département entier, il faut avoir conscience de ce qui suit :

1. Etudier la capacité d'un réseau à généraliser au-delà de sa zone d'apprentissage ;
2. Pour cela, il est vital de définir une stratégie d'apprentissage efficace minimisant notamment l'extraction de patches d'apprentissage ;
3. les charges de calcul inhérentes à l'utilisation de réseaux de neurones sont importantes ; couplées à l'aspect large échelle de notre problématique, il faut prendre en compte que les temps de calcul peuvent être un réel point bloquant, et donc envisager des schémas d'apprentissage et de prédiction efficaces.

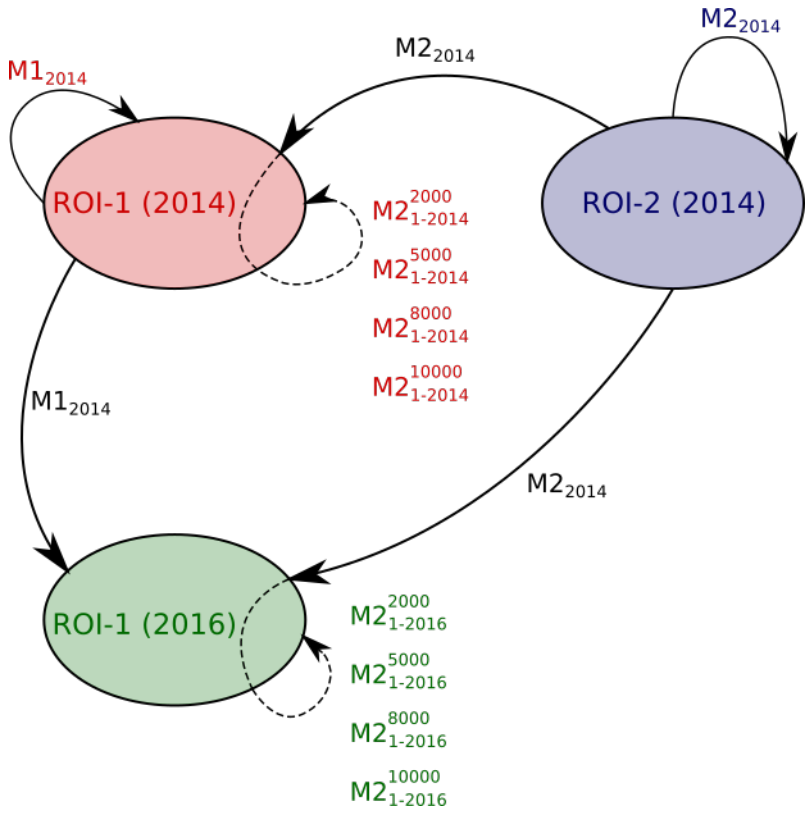


FIGURE III.10. Schéma renvoyant aux différents scenarii d'apprentissage pour l'étude de la capacité de généralisation de notre réseau lorsqu'on lui soumet des données variant en temps, ou en biome, pour une nomenclature donnée.

Pour prendre en considération les points précédents, nous avons mené plusieurs études correspondant à divers scenarii d'entraînement / prédiction.
Nous avons notamment utilisé abondamment du mécanisme de *fine-tuning* permettant de charger un réseau pré-entraîné sur un jeu d'apprentissage A, de l'entraîner sur la base de ce pré-entraînement sur un jeu d'apprentissage B. L'intérêt du fine-tuning est notamment d'affiner un réseau sur une nouvelle donnée en apprenant ses nouvelles spécificités, mais en partant de couches mettant en évidence des *features* similaires entre les deux jeux d'apprentissage, qu'il faut simplement faire converger vers de nouvelles classes ou de nouvelles variétés spectrales par exemple. Typiquement, que ce soit sur des images de Bretagne ou du Sud-Ouest de la France, un contour est décrit de la manière, par l'apparition de haute fréquence ; or, c'est précisément cela qui est appris dans les premières couches d'un réseau de neurones profond. Les *features* sont de plus en plus globales lorsqu'on s'approche du classifieur, et ce sont souvent ces couches qui doivent être ajustées sur les nouvelles données d'apprentissage. Cela permet notamment, nous le verrons d'obtenir un réseau viable sur la nouvelle zone à moindres coûts, que ceux-ci concernent les temps de calcul, ou le nombre d'échantillons nécessaires pour *fine-tuner* le réseau initial.

Fine-tuning temporel et géographique de réseau pré-entraîné par *RWI*

Le schéma III.10 renvoie aux possibles scenarii décrits précédemment. Sur ce schéma, diverses notations sont utilisées :

— on dispose de trois images différentes :
  — ROI-1(2014) renvoie à l'agglomération de Brest, figure III.8a, dont l'image SPOT utilisée a été acquise en 2014 ;
  — ROI-1(2016) renvoie à Brest également, mais pour une image SPOT acquise cette fois en 2016 ;
  — ROI-2(2014) renvoie à l'aire rurale aux alentours de la commune du Faou, pour laquelle nous disposons de l'image uniquement en 2014.

— les stratégies d'entraînements sont notées de la manière suivante :
  — les réseaux *RWI* sont notés MN_{year}, avec N la région concernée, et *year* l'année d'acquisition de l'image sur laquelle a été entraîné le réseau en question ;
  — les réseaux *FT* sont notés MN^{xxxx}_{P-year}, ou N et P désignant, respectivement, la région où le réseau a été pré-entraîné (par *RWI*), et la région où le réseau est affiné à la date *year*, en considérant un nombre $xxxx$ d'échantillons d'apprentissage sur la région P pour l'affiner.
  A titre d'exemple, le modèle $M2^{8000}_{1-2014}$ renvoie au *fine-tuning* sur la région 1 (Brest), avec 8000 échantillons issus d'une image acquise en 2014, d'un réseau pré-entraîné sur la région 2 (Le Faou).

Deux jeux d'entraînement ont été constitués pour les régions ROI-1(2014) et ROI-2(2014). Ils sont constitués, avec une répartition équitable entre les cinq classes d'étude de 10000 échantillons d'apprentissage (patches de tailles 65×65), dont on conserve 10% pour l'étape de validation. Les 90% restant sont utilisés pour mettre itérativement à jour les poids du réseau, et subissent à chaque nouveau batch des transformations aléatoires entre rotation, et symétries, suivant ainsi les stratégies courantes d'augmentation de données. Ces jeux servent à entraîner les deux réseaux par *RWI* $M1(2014)$ et $M2(2014)$ renvoyant respectivement aux zones de Brest et du Faou de 2014.

Si le fine-tuning d'une région géographique vers une autre a un intérêt évident, pour une année fixée, c'est aussi le cas quant à l'utilisation d'un réseau existant entraîné sur une zone à un instant t sur la même zone à un instant t'. La détection de changement est, en effet, un cas d'utilisation direct des cartes d'occupation des sols. Des différences d'apparences de végétation, dues à des saisons différentes, ou des rotations de culture, peuvent induire une variabilité intra-classe très forte d'une époque à une autre. Pour étudier cela, nous avons tout d'abord appliqué le réseau $M1_{2014}$ sur ROI-1(2016), ce qui correspond au test du réseau appris en 2014 sur la **même zone** en 2016. Puis, pour tester un changement géographique et temporel, nous avons appliqué le réseau $M2_{2014}$ sur ROI-1(2016), correspondant à un réseau appris

sur Brest en 2014, puis que l'on utilise pour prédire la zone rurale du Faou à partir d'images acquise en 2016.
Notons que les couvertures SPOT 6/7 annuelles présentent des diachronies comme nous avons pu le voir (le mosaïquage est effectué à partir d'images acquises sur l'ensemble de l'année), l'étude de transférabilité d'un réseau appris d'une année Y vers l'année $Y+1$ par exemple, est tout fait justifiée pour classifier un millésime en particulier, puisqu'on a également ces différences temporelles intra-annuelles.

Les résultats numériques sont exposés dans le tableau III.1 tandis qu'une appréciation visuelle des cas considérés est fournie en figure III.11.

Analysons en premier lieu les performances des réseaux entraînés par *RWI*, c'est-à-dire avec une initialisation aléatoire des poids du réseau. Dans ce cas, seules les trois premières lignes du tableau III.1 nous intéressent. La première ligne retranscrit l'application du réseau $M1_{2014}$ sur ROI-1(2014). On peut constater que le réseau routier et le bâti ressortent très bien. Ce résultat est déjà satisfaisant, les méthodes telles que les Forêts Aléatoires peinant à détecter les routes sans l'utilisation d'une information d'altitude (Gressin et al., 2013). La classe de *culture* obtient un très bon F-score du fait du nombre très important à l'intérieur des polygones de cultures étant correctement détectées. Cela masque ainsi les problème aux frontières. En analysant la seconde ligne, qui correspond à l'application du modèle rural sur la zone urbain dense, on peut constater que toutes les classes subissent une détérioration. Les classes de *bâti* et de *route* sont, comme prévu, les plus impactées car mal caractérisées par le réseau "rural". Enfin, les derniers résultats proviennent du passage du modèle "urbain" vers la même scène urbaine, mais deux ans plus tard. Les scores sont donc très similaires, indiquant une très bonne robustesse du réseau à travers le temps. Les résultats peuvent être appréciés visuellement sur la figure III.11, sur la première colonne à droite des images satellites.
Malgré une architecture légère, une application naïve du réseau permet de détecter des classes habituellement difficiles à trouver. Le réseau routier est spécifiquement très compliqué à récupérer sans MNS. Or, nous arrivons ici à le détecter sans cette donnée d'altitude, généralement quasi systématiquement intégrée dans d'autres méthodes de classification, afin de limiter les confusions avec le bâti.

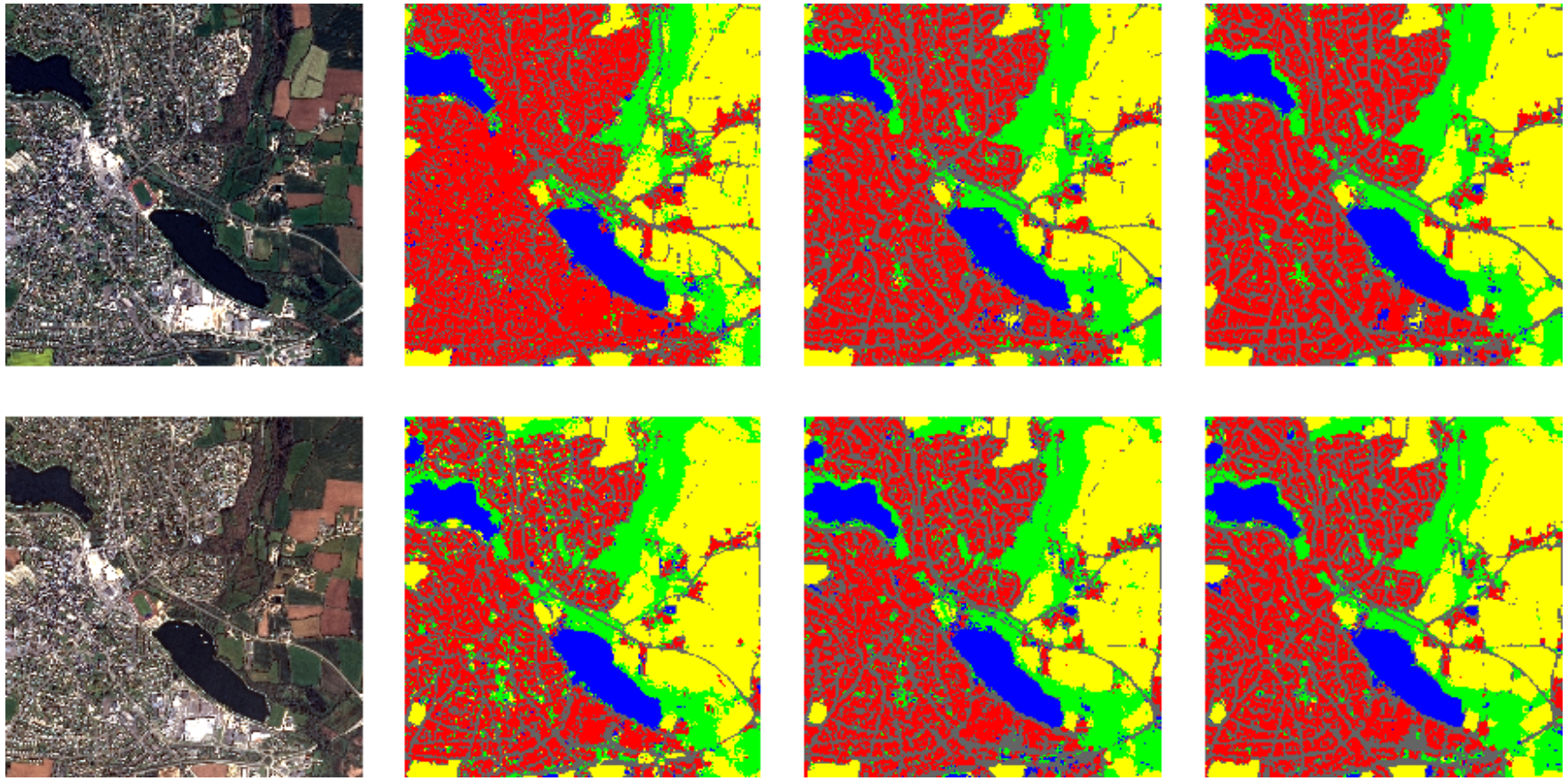


FIGURE III.11. Impact des différentes stratégies d'entraînement sur une zone urbaine.
1ère ligne - de gauche à droite : année 2014 - $M2_{2014}$, $M2^{2000}_{1-2014}$, $M2^{8000}_{1-2014}$.
2ème ligne - de gauche à droite : année 2016 - $M1_{2014}$, $M2^{2000}_{1-2016}$, $M2^{8000}_{1-2014}$.
● *Bâti*, ● *Route*, ● *Culture*, ● *Végétation*, ● *Eau*.

Tests	K	OA	F_{moy}	F_{routes}	$F_{forêt}$	$F_{bâti}$	$F_{culture}$	F_{eau}
$M1_{2014} \rightarrow$ ROI-1 (2014)	0,84	0,92	0,85	0,75	0,85	0,81	0,97	0,86
$M2_{2014} \rightarrow$ ROI-1 (2014)	0,71	0,84	0,73	0,54	0,73	0,68	0,94	0,77
$M1_{2014} \rightarrow$ ROI-1 (2016)	0,83	0,91	0,83	0,75	0,82	0,79	0,97	0,84
$M2^{2000}_{1\text{-}2014}$	0,81	0,90	0,82	0,71	0,84	0,76	0,96	0,82
$M2^{5000}_{1\text{-}2014}$	0,83	0,91	0,84	0,74	0,85	0,80	0,96	0,86
$M2^{8000}_{1\text{-}2014}$	0,85	0,92	0,86	**0,77**	0,87	**0,82**	0,97	0,86
$M2^{10000}_{1\text{-}2014}$	**0,86**	**0,92**	**0,86**	**0,78**	**0,87**	0,82	**0,97**	0,84
$M2^{2000}_{1\text{-}2016}$	0,80	0,89	0,81	0,71	0,83	0,76	0,96	0,78
$M2^{8000}_{1\text{-}2016}$	0,84	0,92	0,85	**0,76**	0,86	**0,81**	0,97	0,85
$M2^{10000}_{1\text{-}2016}$	0,84	0,92	0,85	0,77	0,86	0,81	0,97	0,84

TABLE III.1. Evaluation des résultats issus de l'étude de la capacité de généralisation d'un réseau.

Concernant les tests sur les méthodes employant le procédé de *fine-tuning*, nous avons concentré nos efforts sur la région ROI-1 (Brest) étant donné qu'elle affichait les résultats les moins bons par application directe de réseau pré-entraîné sur la zone du Faou. L'étude comporte aussi l'impact du nombre d'échantillons utilisés pour effectuer ce fine-tuning. En général, même si l'entraînement était programmé pour 300 itérations, environ 100 étaient suffisantes. Dans cette partie, le fine-tuning agit toujours sur le réseau pré-entraîné sur ROI-2(2014), et est appliqué sur ROI-1, aux deux dates.
Lorsque l'on cherche à utiliser un réseau pré-entraîné dans un contexte de classification d'images satellites, dans un objectif de production de cartographies d'occupation des sols, on doit s'assurer de l'homogénéité des résultats sur le territoire. Notamment, l'homogénéité sémantique doit être garantie entre deux images satellites adjacentes et partageant des objets au sol d'intérêt. Cela implique deux choses dans l'analyse de ces images par un modèle de classification quel qu'il soit :

i en faisant l'hypothèse que les deux images adjacentes sont acquises le même jour, et donc que la continuité radiométrique est assurée, on reste tributaire de la nomenclature disponible dans le jeu d'apprentissage puisque celui-ci est défini à partir d'une étendue géographique donnée. Le cas du Faou en est l'illustration : les données disponibles sur cette zone sont très rares pour les classes de bâties, menant aux résultats déjà décrits précédemment.

ii l'hypothèse de continuité radiométrique n'est en pratique pas vérifiée, et renvoie aux diachronies plusieurs fois mentionnées déjà. Pourtant il faut trouver une manière de produire une carte sémantiquement continue à partir de données spectralement complètement décorrélées, les conditions d'illumination pouvant non seulement varier fortement, mais les saisons d'acquisition également.

C'est pourquoi deux stratégies d'ajustement de réseau sont étudiés, l'un permettant de montrer l'efficacité de cette mécanique pour passer d'une zone sémantiquement différente à une autre, pourtant proche géographiquement,

la seconde pour passer d'une date à une autre **et** d'une région à une autre.

1. Abordons en premier lieu le *fine-tuning* purement géographique. Pour lever les ambiguïtés en région urbaine, $M2_{2014}$ a été affiné en utilisant quelques échantillons d'entraînement de ROI-1(2014). Ces tests correspondent aux quatre lignes du milieu du tableau III.1. Des améliorations par rapport à l'application directe sont à noter, notamment pour les deux classes qui avaient été le plus affectées (*bâti* et *route*). Un résultat attendu est également l'augmentation des précisions avec le nombre d'échantillons. On améliore de 23% et 14% les classes de *route* et *bâti* en utilisant 8 000 patchs d'apprentissage de ROI-1. Cette amélioration montre le caractère polyvalent et extrêmement adaptatif des CNNs. On peut constater également visuellement l'amélioration sémantique grâce au fine-tuning sur les deux dernières colonnes de la figure III.11. On n'y affiche que les résultats avec ajustement du réseau à en intégrant 2000 et 8000 échantillons, l'évolution visuelle de la classification variant moins de 2000 à 5000 ou de 5000 à 8000.

2. Intéressons-nous à présent au mécanisme d'ajustement lorsque l'on passe non seulement d'une zone à une autre, mais également avec un décalage dans la date d'acquisition entre ces deux zones. Ainsi, Les trois dernières rangées du tableau concernent le fine-tuning d'un réseau que l'on souhaite appliquer sur une région géographiquement éloignée, et acquise à deux ans d'intervalle. En l'occurrence, $M2_{2014}$ est appliqué sur ROI-1(2016). Les métriques nous donnent encore satisfaction.
L'utilisation de 8 000 échantillons apporte la plus grande amélioration, mais 2 000 échantillons suffisent à gagner 10% au moins sur chaque classe. Il est également intéressant de noter que les durées d'entraînement par ce biais sont biens moindres que pour une approche *RWI*. Elles varient selon le nombre d'échantillons, mais environ 3h suffisent pour obtenir un modèle adéquat. Les conclusions sont donc positives au même titre qu'en *1.* en terme géographique, mais également positif en terme temporel : le *fine-tuning* permet aussi bien de ré-utiliser un réseau pré-entraîné vers une nouvelle zone géographique, mais également vers une temporalité différente.

Fine-tuning sémantique : ajout de la classe *haie*

Nous avons mené jusqu'ici des cas de cartographies d'occupation des sols intégrant cinq classes bien distinctes les unes des autres. Pourtant, nous l'avons vu en chapitre introductif, chaque cartographie existante propose sa propre nomenclature, multipliant celle-ci, en accord avec divers besoins et diverses échelles d'études. De plus, des nomenclatures telles que celles de Corine Land Cover ou de l'OCS-Grande Echelle de l'IGN proposent une hiérarchisation des classes, chaque niveau de hiérarchie correspondant généralement à une lecture de l'information géographique à une résolution donnée. Dans le cas du millésime de 2018 de Corine Land Cover, les territoires imperméabilisés par l'homme sont hiérarchisés selon :

1. Territoire artificialisé

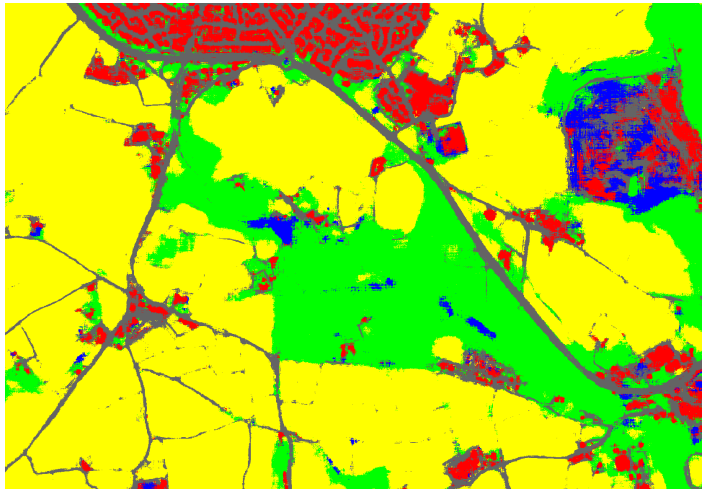
1.1 Zones urbanisées
1.1.1 Tissu urbain continu
1.1.2 Tissu urbain discontinu
1.2 Zones industrielles ou commerciales et réseaux de communication
1.2.1 Zones industrielles ou commerciales
1.2.2 Réseaux routier et ferroviaire et espaces associés
1.2.3 Zones portuaires
1.2.4 Aéroports
1.3 Mines, décharges et chantiers
1.3.1 Extraction de métaux
1.3.2 Décharges
1.3.3 Chantiers
1.4 Espaces verts artificialisés, non agricoles
1.4.1 Espaces verts urbains
1.4.2 Equipements sportifs et de loisirs

La classe globale *Territoire artificalisé* est en adéquation avec des études à grande échelle, européenne ou nationale, pour constater l'étendu des pertes des surfaces agricoles dû à l'étalement urbain. Les niveaux plus précis 1.X ou 1.X.X sont difficiles à percevoir à ces échelles très larges mais deviennent bien plus cohérentes pour des analyses d'agglomérations ou de communes, pour des problématiques de nouveaux chantiers de transports en communs par exemple, qui nécessitent une analyse préalable des densités de population et de la répartition géographique de cette dernière.

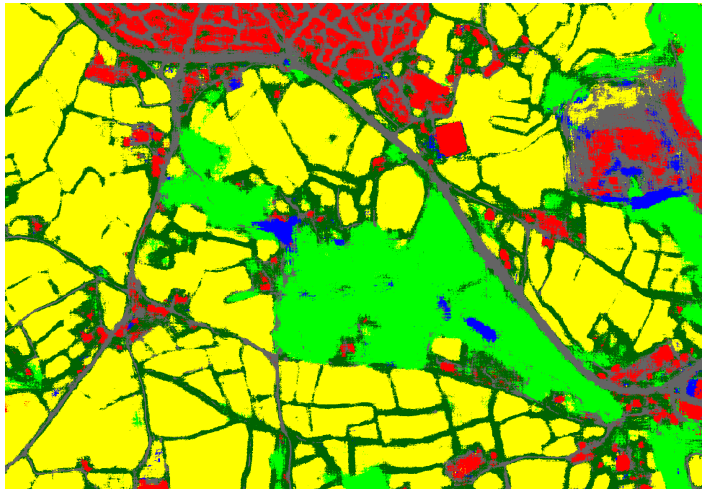
C'est à cette fin de hiérarchisation et d'enrichissement de la nomenclature que nous avons ajouté une nouvelle classe *haies* dérivée de la classe *forêt*. L'étude a été menée sur une partie de notre zone du Finistère. Le réseau entraîné par *RWI* pour détecter cinq classes sur le Finistère a subi un nouvel ajustement, mais cette fois-ci, un **ajustement sémantique**. Comme précédemment, le but recherché est de déterminer l'aptitude d'un réseau à pouvoir être ré-entraîné à moindres coûts pour, dans le cas présent, accroître la granularité de la nomenclature, mais également en gardant à l'esprit les potentiels gains en temps de calcul : réduire le temps d'apprentissage, et le nombre d'entraînements selon la méthode *RWI* à effectuer (il y a potentiellement autant de réseaux à entraîner *from scratch* que de nomenclatures possibles). De même que pour un ajustement temporel / géographique, environ 3h étaient nécessaires.

Le choix de la classe *haies* est dû au fait que nous avons constaté que les classes de *routes* et de *végétation* étaient fortement confondues. La cause majeure de ceci provient de la nature de certains objets de la classe *végétation* qui se trouve être des haies, objets très fins et similaires morphologiquement aux routes. Par ailleurs, on peut voir sur la figure que les haies bordent souvent

(A) Image SPOT



(B) 5 classes :
● *Bâti*, ● *Route*, ● *Culture*,
● *Végétation*, ● *Eau*.



(C) 6 classes :
● *Bâti*, ● *Route*, ● *Culture*,
● *Végétation*, ● *Haie*, ● *Eau*.

FIGURE III.12. Passage d'une occupation des sols à 5 classes vers 6 classes, en affinant un réseau pré-entraîné. On présente le résultat sur une zone de 3,5 km×1,2 km, extraite de ROI-1.

les routes dans les paysages aussi bien urbains que ruraux. Afin d'obtenir de meilleurs résultats sur la classe *route*, nous avons donc subdivisé la classe *végétation* initiale en classes *haies* et *végétation* (qui n'inclue donc plus les haies). A l'instar de ces deux cas d'étude, il a fallu constituer au préalable un jeu d'apprentissage sur lequel affiner le réseau, en incluant la classe *haies* et en retravaillant la classe *végétation* pour que le jeu associée à celle-ci ne comporte plus de haies. Ce jeu, de 5 000 échantillons a été extrait de la zone du Finistère (répartition homogène sur l'ensemble de la zone). Il a également fallu recréer un jeu d'apprentissage pour la nouvelle classe *végétation*, celle-ci n'intégrant plus les haies. Ces deux jeux de données données proviennent de la BD TOPO IGN.
Qualitativement, la classification peut être appréciée sur la figure III.12, qui représente une partie de la zone totale classifiée. Une inspection visuelle permet de confirmer l'efficacité du processus de fine-tuning dans le cas d'ajout de classe supplémentaire étant donné que les haies séparant les parcelles agricoles ainsi que celles longeant le réseau routier sont bien détectées, tout en préservant les classes initiales.
La zone a été choisie également pour l'objet présent à l'Est, qui sur l'image, se révèle être une carrière. Sur les deux classifications, cet objet est labellisé dans l'une des cinq, respectivement six, classes que nous avons fournies à notre classifieur. Il est important de comprendre que le classifieur ne peut pas *inventer* la classe *carrière* et est tributaire de la nomenclature fournie via le jeu d'apprentissage en amont de la tâche de prédiction. Ce cas de figure est intéressant dans la mesure où il représente une classe que le classifieur n'a pas apprise, car absente du jeu d'entraînement. De tels objets n'appartenant à aucune classe sont d'autant plus susceptibles d'apparaître que la zone géographique analysée est étendue. Une solution serait d'intégrer une classe de *carrière* dans notre cas, mais (i) ces objets sont très rares comparés au reste des classes, (ii) d'un point de vue plus global, nous venons de le dire, nous aurons toujours des objets « parasites » pour lesquels il sera impossible de constituer un jeu d'apprentissage, (iii) étant rares, ils n'ont qu'un impact moindre sur les métriques. En revanche, dans un objectif cartographique, la connaissance préalable de leur emplacement (les carrières étant des objets pérennes et peu sujets au changement), permettrait de filtrer a posteriori ces emplacements très particuliers.

Une validation quantitative a également été effectuée, dont les résultats sont fournis dans le tableau III.2. La première ligne correspond aux performances, calculées sous forme de F-Score pour chaque classe, lors de la classification à cinq classes tandis que la seconde indique les F-Scores après ajout de la classe *haie*. L'objectif qui était de séparer les classes *végétation* et *route* ne semble qu'en partie résolue. En effet, si la première obtient un meilleur résultat qu'en ne considérant pas les haies à part entière, on peut constater une décroissance sur les autres classes en terme de qualité, malgré un visuel très satisfaisant. En particulier, la classe *culture* subit la décroissance la plus importante. Cette diminution des métriques, contradictoire avec le résultat visuel s'explique en partie par la présence de la carrière pour les classes

concernées (*eau*, *bâti*, *route* sont sur-détectés sur la carrière) mais une explication supplémentaire s'ajoute.
La qualité des données de référence pour la classe *haie* utilisées pour effectuer notre validation croisée, est très discutable (mais demeure la seule à notre disposition). La donnée est fortement impactée par des erreurs de mise à jour, avec des objets *haies* réellement existants, mais absents de cette vérité terrain. Ces haies, à juste titre classifiées comme telles par le réseau, sont donc considérées à tort comme des parties de parcelles agricoles par la donnée de référence. Par ailleurs, les confusions *route / végétation* sont transposées vers des confusions entre *route* et *haie*. On note beaucoup de haies qui longent les routes, avec des frontières floues entre les deux objets. En ajoutant le fait que les routes sont des objets fins, une erreur de classification à leurs frontières impacte fortement les performances.
En disposant de données de référence à jour pour la classe *haie*, il est certain que les scores augmenteraient significativement pour les classes de *culture* et de *végétation*. En outre, le F-score de la classe *route* n'a qu'en réalité très peu diminué pour un objectif cartographique, tout en offrant une granularité sémantique plus riche.

Bâti	*Route*	*Végétation*	*Culture*	*Eau*	*Haie*
73.30	**80.85**	83.66	**95.84**	67.76	/
72.15	79.94	**84.28**	92.36	**68.98**	43.54

TABLE III.2. Évolution des performances en ajoutant la classe *Haie*. La métrique correspond au F-Score par classe.

Conclusions sur le *fine-tuning*

Après avoir soumis un réseau entraîné par *RWI* aux trois changements de perspectives envisageables lorsque l'on procède à la classification de larges régions géographiques, les possibilités de gains en temps machine, en nombre d'échantillons minimal nécessaires pour couvrir ces régions et classes, sont très prometteuses.
Que l'on mette en œuvre le *fine-tuning* pour prédire une nouvelle zone géographique, à une nouvelle date, avec des classes différentes, les résultats restent cohérents, et homogènes avec la zone de pré-entraînement, garantissant notamment une continuité sémantique que l'on peut qualifier de « forte » dans les deux premiers cas (décalage temporel et géographique), et une continuité sémantique « faible » dans le cas d'un ajustement sémantique puisque la nomenclature est modifiée, mais demeure malgré tout cohérente sur les classes identiques (le *bâti* par exemple dans le cas étudié, est constant que l'on soit dans le cas avec ou sans la classe supplémentaire de *haie*).

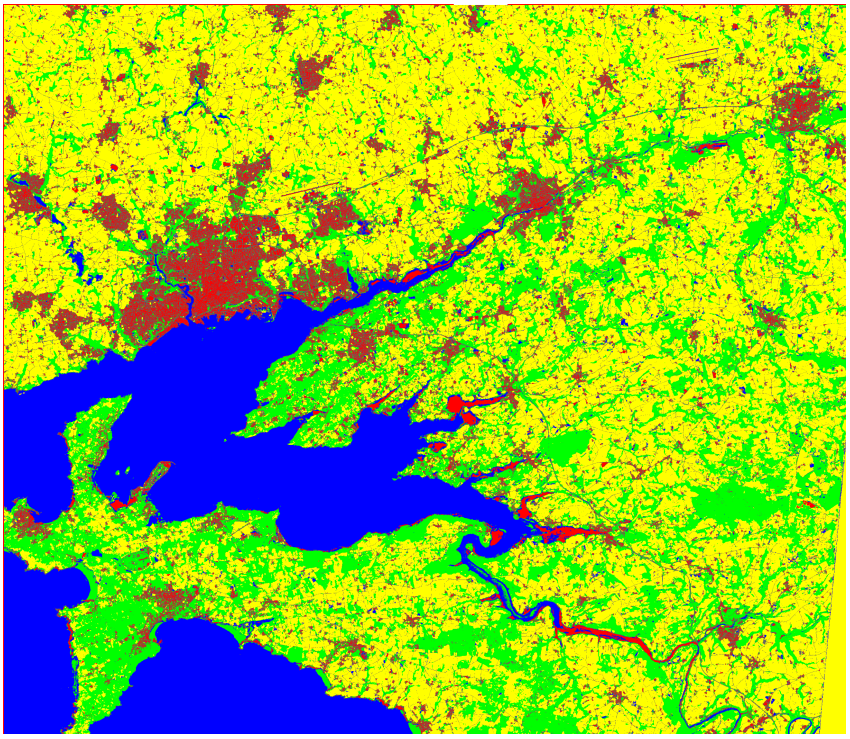
Région	label	S (km^2)	Pixels
Finistère	A	1755	780 millions
Gironde	B	4554	2 milliards

TABLE III.3. Etendues des deux régions classifiées à large échelle (et à la pleine résolution disponible avec SPOT 6/7.

3.3 Classification à large échelle

Les différentes stratégies de *fine-tuning* ayant été éprouvées avec succès dans des cadres de cartographie d'occupation des sols, on se concentre ici sur la classification d'images SPOT 6/7 sur de larges zones géographiques. Nous avons pu décrire dans le paragraphe 3.1 les deux zones d'intérêt étudiées. Si la zone bretonne a fait l'objet de plusieurs analyses au cours des sections précédentes, on s'attache à en fournir une labellisation complète dans cette partie, de même pour le département de la Gironde.
Les images SPOT étant très volumineuses (jusqu'à 20 Go sur la zone de la Gironde), il est inenvisageable de mener une prédiction sur l'image d'origine d'une seule traite. Pour des raisons de mémoire, mais également pour éviter toute perte d'information liée à une éventuelle interruption du processus de classification, les régions sont donc découpées en tuiles de taille 2100×2100 et prédites une par une. Pour garantir une cohérence sémantique entre tuiles successives, un recouvrement entre elles de la moitié de la taille d'un patch est effectué.
Pour plus de facilité dans les études comparatives qui suivent, nous ferons référence dans la suite au Finistère par la région A, la Gironde par la région B. Le tableau III.3 décrit en terme de superficie les deux régions.

(A) Région A, classifiée avec un modèle appris sur la zone délimitée en noire.



(B) Occupation des sols sur la zone du Finistère avec le réseau entraîné sur ROI-1 (la zone de Brest étudiée en section 3.2).
Les dimensions de cette zone sont de 39×45 km.
● *Bâti*, ● *Route*, ● *Culture*, ● *Végétation*, ● *Eau*.

FIGURE III.13. Classification de la région A.

OCS sur la région Finistère

En premier lieu, une classification a été effectuée sur la zone A, représentant le quart du Finistère. La classification s'est faite par application du réseau $M1_{2014}$ qui a servi de baseline en section 3.2 pour l'ensemble des tests de *fine-tuning*. Aucun ajustement n'a été effectué au moment du passage à l'échelle, ce qui pourrait nous porter préjudice car si la région comporte plusieurs aires urbaines, elle n'en reste pas moins majoritairement agricole. Or $M1_{2014}$ correspondant à un réseau entraîné sur l'agglomération de Brest, le modèle caractérise un paysage urbain dense (ainsi que des zones plus naturelles). De plus, les classes d'occupation varient beaucoup spectralement par rapport aux patches qui ont servi d'apprentissage pour $M1_{2014}$. En effet, les estuaires sont des zones humides et les landes (zone à classer en *végétation*), caractéristiques des zones proches de la côte, sont des objets présents sur la région A mais tous deux absents absents du jeu de données d'apprentissage. Rappelons que l'emprise de la zone ainsi classifiée est visible sur la figure III.7, délimitée en rouge.
L'objectif de cette étude sur la région A est de suivre l'évolution des performances d'un réseau entraîné sur une zone puis appliqué sur des régions progressivement plus éloignées de sa zone d'apprentissage. Même si la couverture de cette zone correspond à une unique image SPOT, des différences d'illumination au sol et de changements progressifs du paysage peuvent éventuellement engendrer des difficultés pour le réseau à classer correctement les objets d'une même classe mais situés à l'opposé sur l'image.

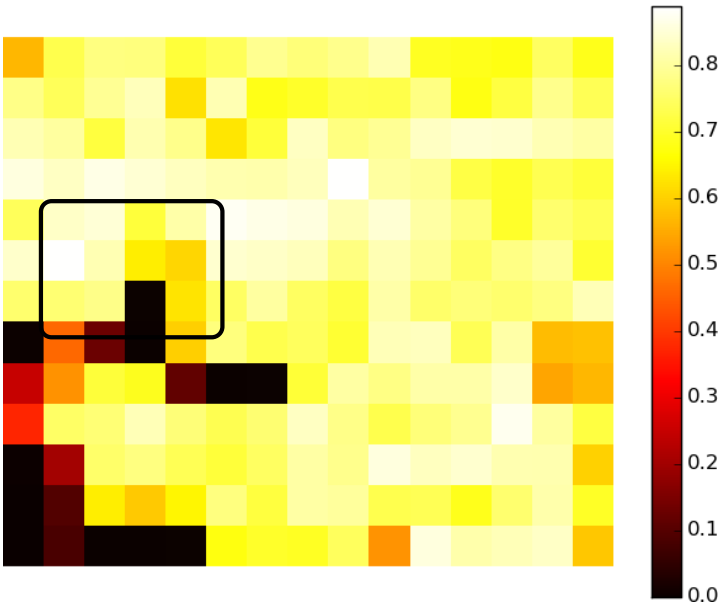
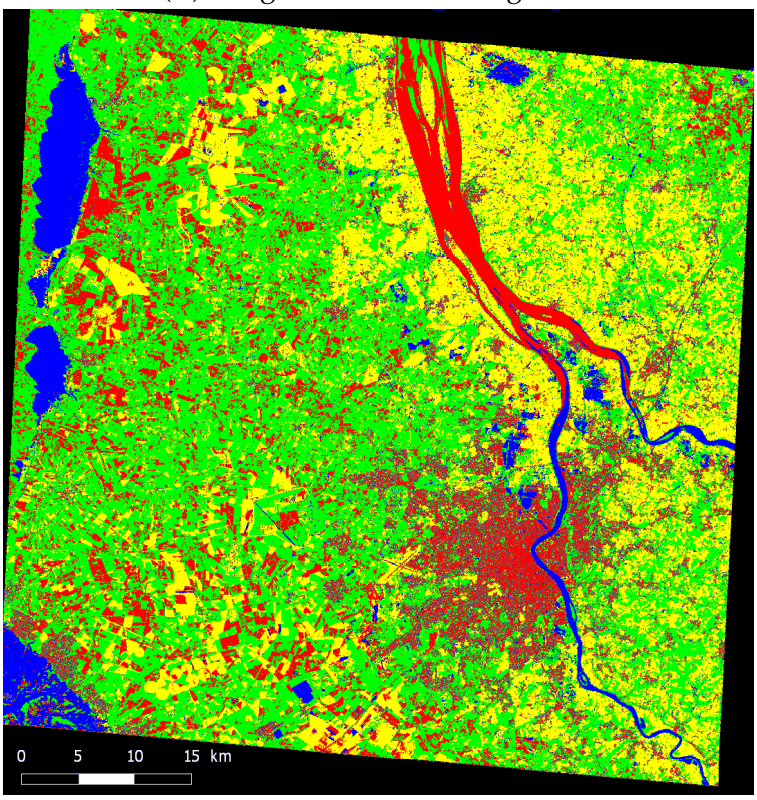


FIGURE III.14. Indice Kappa sur la région A : la qualité ne décroît pas en s'éloignant de la zone d'apprentissage (en noire).
Les pixels noirs correspondent aux régions sans données de référence pour la validation croisée.

La carte d'occupation des sols que nous obtenons en appliquant $M1_{2014}$

sur ROI-3 est visible sur la figure III.13. La classification de cette zone a pris 3,5 jours, pour 780 millions pixels. La zone qui a servi à l'apprentissage est détourée en noir sur la figure III.13a, et correspond à environ 8% de la zone totale. Il est intéressant de regarder la figure III.14 qui montre l'indice Kappa calculé sur chaque tuile. On constate que le résultat est très satisfaisant globalement (l'indice kappa étant global), et **la sémantisation reste homogène en s'éloignant progressivement de la zone d'apprentissage**. Les pixels noirs sur la figure indiquant le kappa correspondent à des zones où il n'y pas de données de référence (zones exclusivement maritimes si on regarde la figure III.13). Les estuaires sont des éléments que l'on souhaiterait classer en *eau* mais aucun estuaire n'étant présent sur la zone d'apprentissage, il est normal que des confusions surviennent sur ces régions. Pour pallier ce problème, procéder par *fine-tuning* sur des échantillons bien choisis représentant cette configuration est une solution à explorer au vu des résultats précédents sur le mécanisme de *fine-tuning*.

(A) Image SPOT 6/7 sur région B.



(B) OCS de la région B avant ajustement.
● *buildings*, ● *roads*, ● *crops*, ● *forest*, ● *water*.

FIGURE III.15. OCS sur la Gironde par prédiction directe avec le modèle appris sur l'agglomération de Brest.

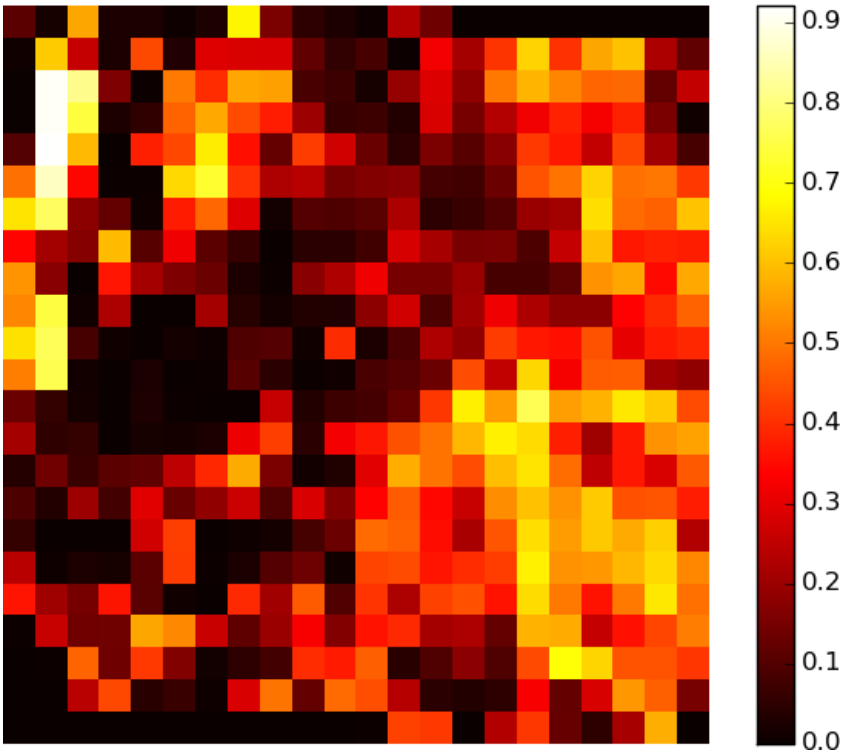


FIGURE III.16. Indice Kappa, par tuile, de la classification avec un modèle « brut » entraîné sur un écosystème différent.

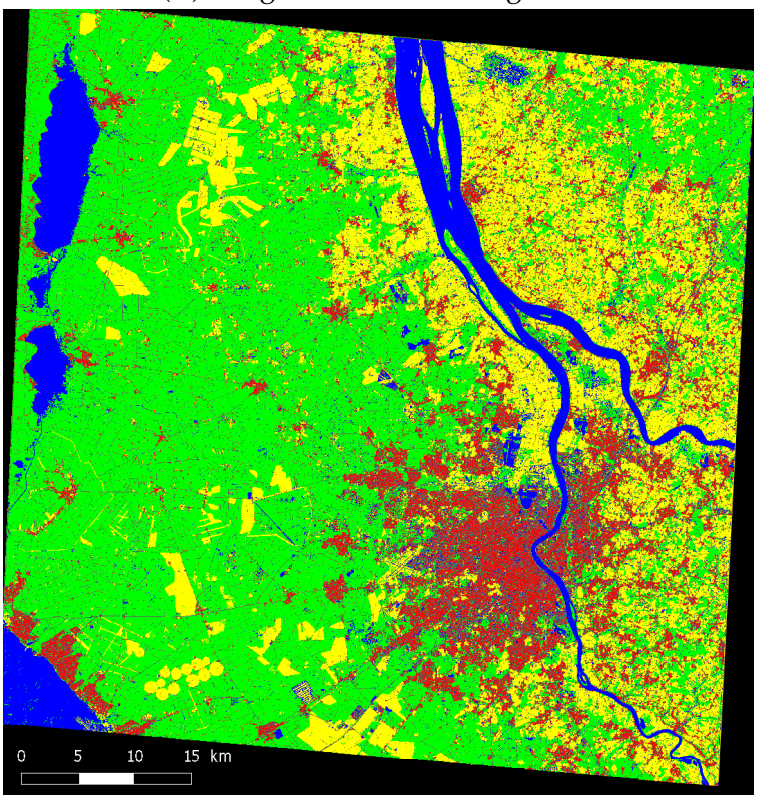
OCS sur le département de la Gironde

La région B est deux fois et demi plus grande que la région A et correspond à la moitié du département girondin, mais offre une variété importante de paysages, et est complètement différente de la région A, avec laquelle le seul point commun est leur proximité du littoral Atlantique.
La classification de cette zone par le modèle $M1(2014)$ rejoint les contraintes explicitées dans le paragraphe 3.2 puisque le changement géographique est évident, et exacerbé par rapport au cas étudié dans la section en question. Mais, la couverture de la région B correspondant à une image acquise en 2016, le changement temporel est lui aussi à prendre en compte, et possiblement la différence de saison entre les deux dates d'acquisition.

Le premier test correspond à la classification de la zone par le modèle sans aucun ajustement pour comprendre l'impact des différences entre région A et B, et voir si elles sont visibles sur l'image de classification résultante. Ce test a déjà été conduit sur le département du Finistère lors du passage de Brest à Le Faou, donc on s'attend à des résultats similaires, voire moins bons étant donné le changement plus drastique de paysage (notamment au nvieau de la végétation). En localisant les pixels mal classifiés dans cette image, on peut ainsi inverser le processus et comprendre les causes qui engendrent ces erreurs, notamment en comparant les paysages des régions A et B. La carte d'occupation que l'on obtient est celle visible en figure III.15, sur laquelle on

peut également voir l'image SPOT classifiée correspondante.
Visuellement, le résultat permet de distinguer les différentes parties de la région, il n'en demeure pas moins qu'il n'est pas celui escompté, avec de fortes erreurs de classification, et beaucoup de confusions entre diverses classes. Ce produit pouvait être attendu, le modèle de prédiction ne connaissant pas le biome en Gironde, et attribuant les classes sur la base des échantillons fournis, issus de l'agglomération de Brest.

(A) Image SPOT 6/7 sur région B.



(B) OCS de la région B après ajustement.
● *buildings*, ● *roads*, ● *crops*, ● *forest*, ● *water*.

FIGURE III.17. OCS après ajustement du modèle.

L'essentiel des confusions provient de la classe *bâti* qui est très sur-détectée sur les régions forestières, notamment les pinèdes, et sur l'estuaire de la Gironde. L'agglomération bordelaise est toutefois bien présente, suggérant un faible taux de sous-détection du bâti. Les confusions *bâti / eau* ont déjà été expliquées en introduction du chapitre : l'eau de l'estuaire diffère complètement de l'eau présente sur la zone d'entraînement. Avec un aspect très turbide et avec une forte réflectance en raison des bancs de sable notamment, l'estuaire présente une texture et une radiométrie très particulièree que le classifieur approche davantage du bâti. Cette explication est appuyée par ailleurs par la bonne classification des surfaces d'eau en bordure de image, à l'Ouest, qui sont, elles, très opaques, à l'image des étendues d'eau de la zone d'apprentissage. Concernant le *bâti* sur-détecté sur les zones de végétation, la raison est un peu moins évidente, et s'explique par deux points principaux :

— les pinèdes, essentiel des zones arborées en Gironde, constituent une essence très peu présente en Bretagne, et appartiennent à la famille des résineux, dont le jeu d'apprentissage a très peu voir pas du tout de représentants. L'agglomération de Brest est principalement un paysage urbain qui regroupe peu de résineux.
Dans ce paysage girondin, les résineux, même s'ils sont de la classe *végétation* en réalité, ont une signature spectrale très différentes de la couverture de végétation connue du modèle. Plus particulièrement, les résineux ont une réponse moins forte que les feuillus dans le spectre infrarouge.

— en raisonnant d'un point de vue morphologique, les pinèdes n'étant pas naturelles (industrie forestière), elles présentent des géométries et les structures d'une grille. Or, les paysage urbain est construit de cette amnière, avec des îlots d'habitations séparés par des routes droites. Cette similarité morphologique conduit à de faux-positifs d'objets *bâti* au détriment de la classe de *végétation*.

A l'image de la classification sur la région A, on peut calculer un indice Kappa par tuile permettant voir la répartition des erreurs sur la zone. La carte de qualité III.16 permet de voir les grandes régions mal classées correspondant effectivement à l'estuaire de la Gironde et aux régions dédiées à l'activité forestière (pinèdes) situés à l'Ouest.

Une application directe du modèle n'étant que peu fructueuse, le modèle est affiné en ré-utilisant le mécanisme de *fine-tuning* ayant déjà fait ses preuves auparavant. Pour cela, 5000 échantillons, toutes classes confondues, sont extraits de la zone, avec une répartition à nouveau régulière sur ladite zone. Les résultats sont visibles sur la figure III.17, avec à nouveau la cartographie du kappa en figure III.18. L'entraînement a été conduit pendant 300 itérations, soit 1h sur les 5000 échantillons. Nous avons également testé en n'affinant qu'avec 2500 échantillons pour une légère décroissance des performances (81.1% de précision sur le jeu de validation contre 82.2% lorsqu'on utilise 5000 échantillons). En revanche, porter le nombre d'échantillons à 7500 n'apportait pas d'amélioration notable. Toutefois, le choix des échantillons est très important, et affiner sur la classe d'eau par exemple permettrait de

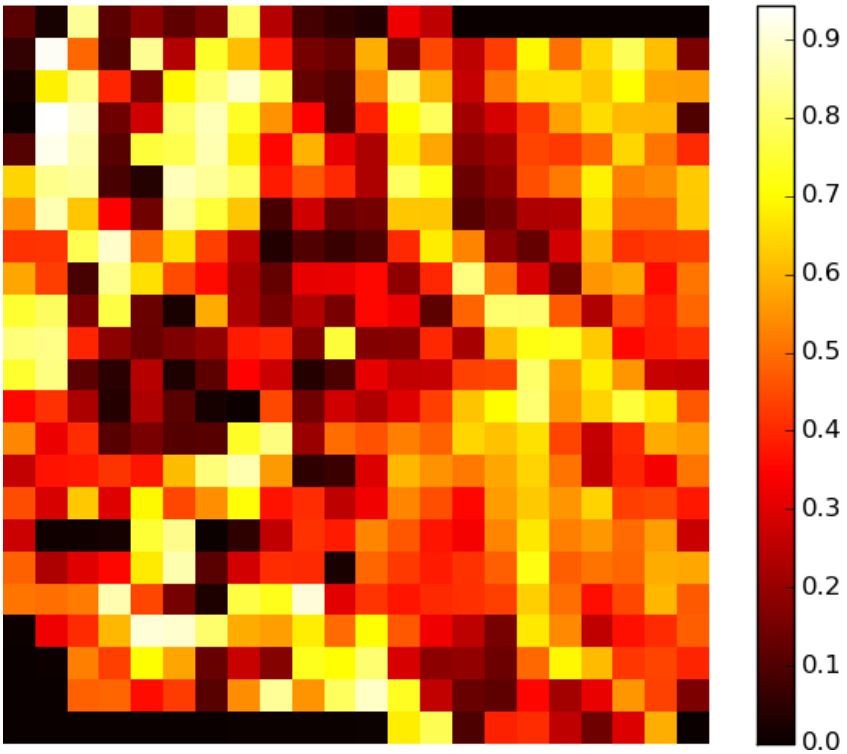


FIGURE III.18. Indice Kappa, par tuile, de la classification avec un modèle « brut » entraîné sur un écosystème différent, **puis ajusté** sur la région B.

lever des ambiguïtés en milieu urbain en particulier, dans lequel les ombres des bâtiments sont sources de sur-détection de la classe *eau*.
L'occupation des sols post-ajustement est très satisfaisante, les confusions majeures ayant été levées pour la classe *bâti* qui se voit contrainte grâce au jeu d'apprentissage local, spécifiant davantage l'aspect des pinèdes et de l'estuaire. Deux focus sont proposés permettant d'illustrer la levée d'ambiguïtés *bâti / eau* et *bâti / végétation* en figure III.19 et III.20 respectivement. L'eau est bien mieux détectée, et le bâti isolé sur la figure III.19 est conservé, indiquant que le réseau ne classe pas mieux l'eau au détriment d'une amputation d'une partie du bâti.

Conclusion

Au travers des études des deux grandes régions d'intérêt que l'on s'est fixé, on a pu retrouver les dynamiques paysagères de chacune d'entre elles, avec des précisions meilleures pour la région A, correspondant au Finistère. La raison principale tient dans le fait que le réseau utilisé est entraîné sur cette région à l'origine, proposant une très bonne classification du reste de la région A. La région B, au demeurant, propose des paysages plus compliqués et est parfois plus difficile à interpréter même visuellement, l'image SPOT étant plus bruitée.
Concernant la région B, davantage d'échantillons auraient très probablement aidé, notamment pour la classe *eau* qui est très sur-détectée, et pour laquelle

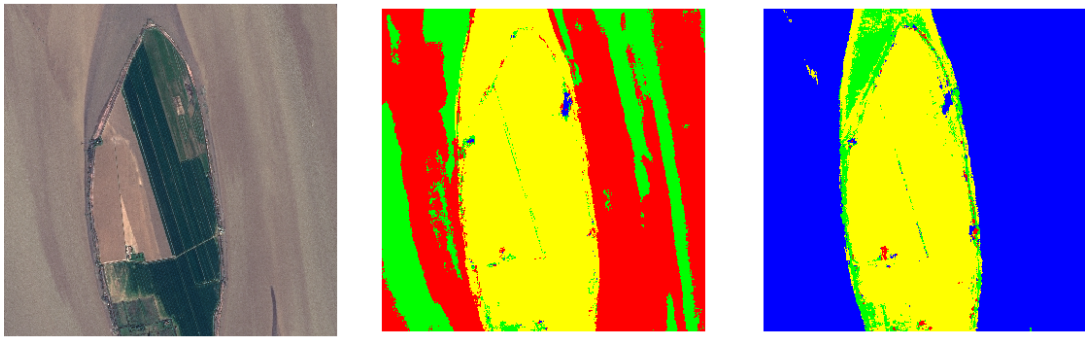


FIGURE III.19. *Fine-tuning* sur l'estuaire de la Gironde.
De gauche à droite : image SPOT - avant ajustement - après ajustement.

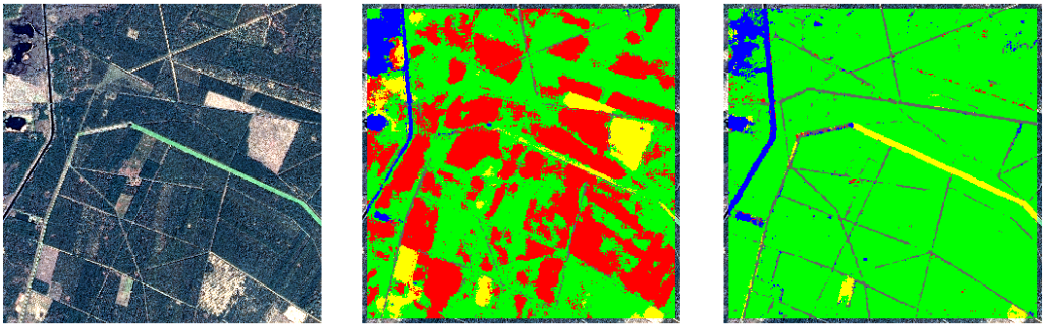


FIGURE III.20. *Fine-tuning* sur les pinèdes.
De gauche à droite : image SPOT - avant ajustement - après ajustement.

il aurait été intéressant de construire un jeu d'apprentissage plus représentatif. De manière générale, 5000 échantillons toutes classes confondues est trop peu pour une zone si étendue : les tests de *fine-tuning* menés en section 3.2 considéraient également autant d'échantillons, mais sur la région A, un peu plus petite que la région B.
Malgré tout, ces travaux montrent l'impact indéniablement positif du *fine-tuning* et que les résultats s'améliorent sensiblement d'un point de vue qualitatif **et** quantitatif entre une cartographie sans et avec ajustement. Cette conclusion permet notamment de valider des stratégies de cartographie d'OCS par des méthodes de réseaux profonds en minimisant les entraînements nécessaires sur l'ensemble de territoire, en gardant un réseau « père » duquel des réseaux « fils » peuvent être dérivés par *fine-tuning*. Avec des résultats satisfaisants sur l'ajustement sémantique, cette stratégie est réellement plausible, tout en gardant à l'esprit que la labellisation et la construction de jeux d'apprentissage en adéquation avec la nomenclature et la variété des paysages est une étape cruciale à mener, cela étant vrai pour n'importe quelle méthode de cartographie automatique envisagée.

3.4 Sur-segmentation de l'image à classifier

Les résultats de prédiction détaillés dans la section précédente 3.3 sont coûteux en temps de calcul, chaque tuile de 2100×2100 étant traitée en 30 minutes environ.
Dans le but de réduire l'impact de la stratégie de l'approche au patch sur les temps de prédiction, nous proposons de **segmenter** l'image à classer en amont de cette classification. Pour cela, deux options sont possibles : (i) adopter une approche visant à extraire des régions correspondant aux objets sémantiquement significatifs de la scène, (ii) produire une sur-segmentation de l'image en segments plus petits que des objets mais homogènes en radiométrie. Le choix s'est plutôt porté sur une sur-segmentation, ou approche **superpixels**. L'appellation *superpixels* renvoie au fait que les algorithmes en question produisent un résultat pour lequel les objets sémantiquement significatifs sont divisés en segments, au lieu d'avoir l'objet en entier pour une approche purement OBIA (Object Based Image Analysis). Toutefois, nous privilégions ici la **pureté** de chaque superpixel : chacun d'entre eux doit rassembler tout ou une partie d'un objet. A l'inverse, une approche purement OBIA est très susceptible de produire des segments incluant des pixels n'appartenant pas à celui-ci (sous-segmentation). Un grand nombre d'algorithmes de sur-segmentation existent, présentant tous des avantages et des inconvénients différents. Par exemple, l'algorithme SLIC (Simple Linear Iterative Clustering), développé par Achanta et al. (2012) permet de produire une grille de superpixels à l'aspect très régulier (mosaïque), en calculant un K-means intégrant les coordonnées dans l'image de chaque pixel en plus de son information radiométrique. Les algorithmes de segmentations produisant de tels résultats (Stutz et al., 2018) ne peuvent être envisagés dans notre configuration car les objets que nous souhaitons classés sont (i) différents en superficie et (ii) présentent des morphologies alternant entre fine (*route*) et

large (*culture*).

L'algorithme utilisé (Felzenszwalb et Huttenlocher, 2004) permet de produire des segments de formes variées et paramétrés pour que ceux-ci restent restreints spatialement. Cette méthode repose sur l'analyse d'un graphe et la comparaison entre deux scores. La premièr représente le score entre deux superpixels adjacents et correspond à la variation minimale de radiométrie entre deux pixels (chacun appartenant à l'un des superpixels). Le second représente l'homogénéité au sein de chaque superpixel. Si celui-ci est supérieur au premier, alors le poids entre ces deux superpixels est important (chaque superpixel définit une région homogène en son sein et différente de l'autre). L'algorithme est paramétré par trois variables, dont un en amont de la segmentation elle-même : il s'agit d'un paramètre lié à la largeur du noyau gaussien utilisé pour lisser l'image. Le résultat de la segmentation elle-même est tributaire de deux variables : m permet de contrôler la taille minimale des superpixels (en pixels) et k définit une échelle d'analyse (k élevé produit de plus larges superpixels). Nous avons ainsi opté pour cette méthode.
La stratégie de classification est la suivante :

1. Segmentation de l'image en superpixels ;
2. Pour chaque superpixel :
   (a) sélectionner régulièrement au sein du superpixel un sous-ensemble inclus dans l'ensemble des pixels ;
   (b) classifier chaque pixel ;
   (c) attribuer au superpixel la classe majoritaire.

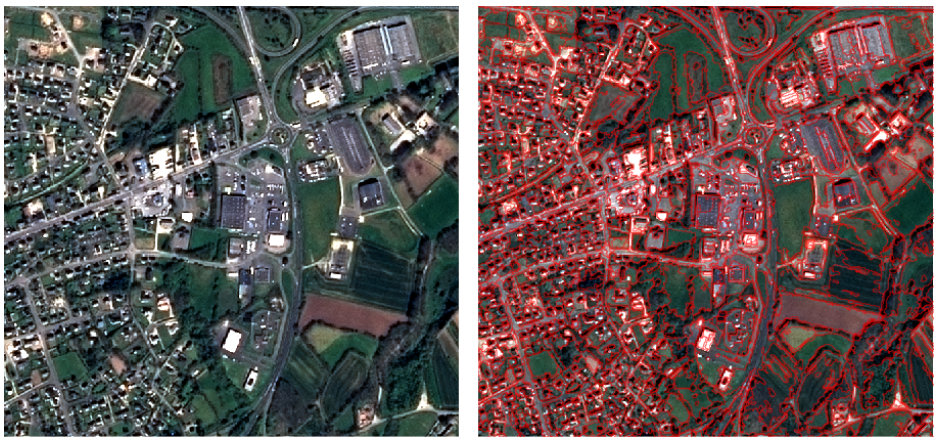


FIGURE III.21. A gauche : Image SPOT - A droite : segmentation utilisant la méthode (Felzenszwalb et Huttenlocher, 2004) avec $k = 30$, $m = 20$.

On peut voir un exemple de segmentation sur la figure III.21, produite pour le paramétrage suivant : $(k, m) = (30, 20)$. On peut noter que des valeurs voisines de celles choisies produisent des cartes de segmentation similaires. Le résultat de segmentation montre des segments cohérents avec la morphologie semi-urbaine de la scène, et ne présente pas de segment mixtes

Stratégie	Kappa	F-score					T
		Bâti	*Route*	*Végétation*	*Culture*	*Eau*	
20%	**75,76**	**68,19**	**76,77**	**81,90**	**92,27**	**14,44**	**4,5'**
« Pixels purs »	73,82	66,61	75,29	79,79	91,64	10,99	28'

TABLE III.4. Comparaison des performances entre une approche « pixels purs » et l'utilisation de superpixels.

contenant des pixels de classes d'occupation différentes.

Afin d'apprécier la pertinence de l'utilisation de superpixels pour réduire les temps de traitement à la phase de prédiction, nous avons comparé une classification effectuée sans segmentation préalable, où chaque pixel est classifié par le réseau de neurones convolutifs, avec une classification de superpixel. Visuellement, le résultat est très satisfaisant comme en atteste la figure III.22. La scène choisie montre que l'approche reste valide que l'on soit en milieu urbain dense, semi-urbain ou encore rural.

Les résultats de validation croisée, visible sur le tableau III.4 indiquent une nette amélioration de la classification en pré-segmentant l'image (ligne 1 du tableau), en comparaison de l'approche utilisée jusqu'ici (ligne 2). Ceci se fait en ne classifiant que 20% des pixels au sein de chaque superpixel. L'amélioration globale sur toutes les classes provient de la nature même de l'approche superpixels qui limite fortement le bruit, au contraire présent sur la classification « pixels purs ». La classe *eau* obtient de faibles performances car une forte sur-détection survient du fait de l'inondation de friches après des précipitations qui ont précédé l'acquisition de cette image. Toutefois, la couverture des BD utilisées pour réaliser cette validation croisée contient des zones vides, or la segmentation peut parfois généraliser des objets en les faisant déborder dans ces zones. Les chiffres indiqués ne traduisent donc pas ce phénomène.
Par ailleurs, un peu moins de 5 minutes suffisent à établir une classification, de qualité supérieure. Cette durée, qui inclut le calcul de la segmentation, est bien plus satisfaisante face aux 30 minutes environ nécessaires pour l'approche précédente. Dans le but de compléter cette étude, nous avons mené ce traitement en utilisant jusqu'à 80% des pixels au sein de chaque superpixel. Toutefois, cela n'améliore les performances au mieux que de 0,6%, pour un temps de calcul considérablement accru (augmentation linéaire avec le nombre de pixels classifiés).

En marge de cette étude sur les superpixels, afin de réduire ces temps de calcul indépendamment de l'utilisation d'une segmentation, nous avons calculé les prédictions par batch de fenêtre 65×65, exploitant ainsi le GPU le plus possible à la prédiction également, en classifiant les fenêtres en parallèle, au lieu de séquentiellement (fenêtre glissante classique). Ainsi, une tuile de taille 2100×2100 pixels est constitué de 1044 « sous-fenêtres » ($\frac{2100 \times 2100}{65 \times 65}$) :

si on prédit par batch de 256 fenêtres en parallèle, quatre batchs sont nécessaires, divisant le temps de prédiction par quatre également. On a ainsi pu réduire le temps de prédiction à moins de 8 minutes pour une tuile.

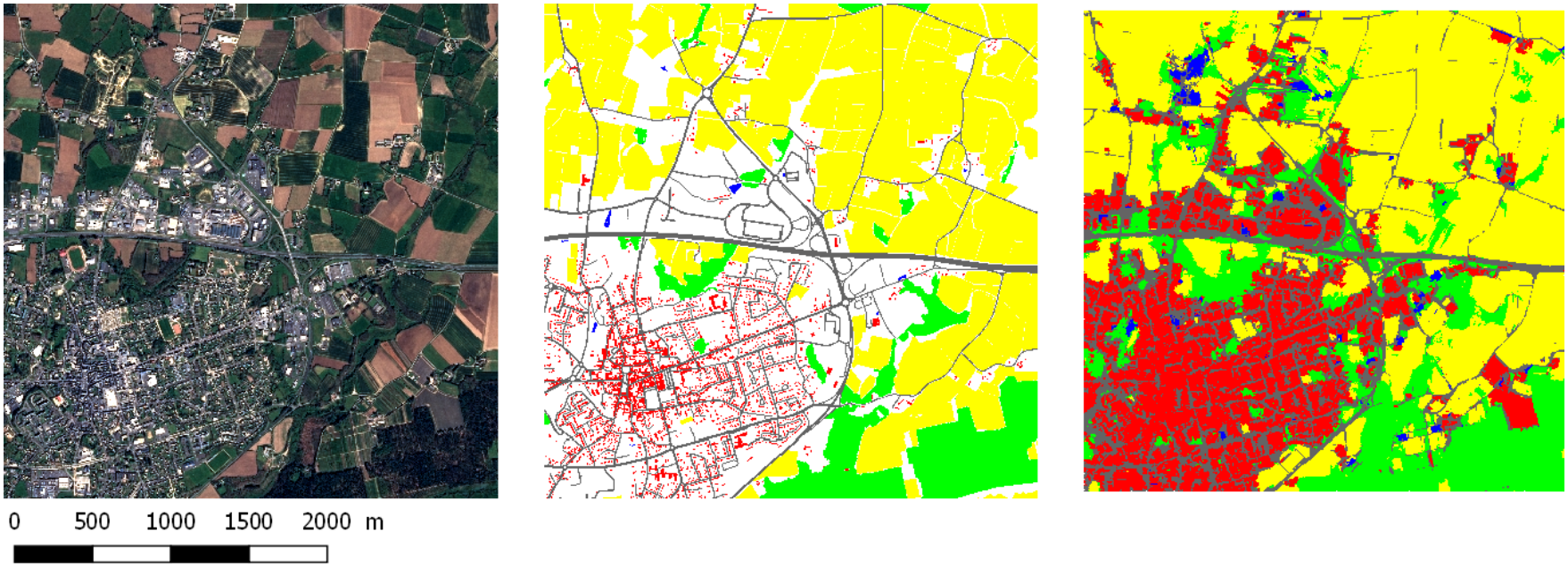


FIGURE III.22. Classification via superpixels
- De gauche à droite : Image SPOT, donnée de référence, classification sur les superpixels :
● *Bâti*, ● *Route*, ● *Culture*, ● *Végétation*, ● *Eau*.

Chapitre IV

Méthodologie vers une mise à jour opérationnelle de bases de données d'Occupation des Sols

1 Genèse de l'étude

Les travaux de cette section sont le fruit d'une réflexion répondant à la problématique du suivi de l'**artificialisation des sols**. Nous avons évoqué ce besoin en chapitre d'introduction, sans détailler de quoi il est question réellement. Pourtant, le problème posé par l'artificialisation est au cœur de l'un des axes stratégiques du Plan Biodiversité mis en place à l'été 2018 par le gouvernement français.
Ce plan vise plus largement à rendre à la Nature sa place sur la planète, et plus spécifiquement à rendre à la faune et à la flore leurs écosystèmes respectifs, qui sont variés selon les espèces, mais aujourd'hui grandement menacés. De tels écosystèmes correspondent entre autres à des espaces aujourd'hui anthropisés, à l'image des zones de végétation déforestées au profit de nouvelles terres agricoles, ou encore de pertes d'espaces naturels au profit de l'étalement urbain et de la densification des agglomérations. Leur destruction et disparition engendrent une perte bien évidemment de la biodiversité, mais impactent également le réchauffement climatique par la destruction de

zones naturelles de stockage de carbone que sont les forêts. L'artificialisation des sols dans un contexte urbain provoque quant à elle une imperméabilisation de ces sols, amplifiant les risques d'inondations.
L'axe du plan Biodiversité en question affiche justement un objectif de « zéro artificialisation nette » : les surfaces nouvellement artificialisées doivent être **en surface** égales aux surfaces réhabilitées vers un habitat 100% naturel. Sans rentrer dans le détail, cette réhabilitation suggère des étapes de déconstruction, dés-imperméabilisation et dé-pollution du sol qui sont des processus très coûteux. Toujours est-il qu'une **mesure quantitative de cette artificialisation** est nécessaire pour répondre à cet axe stratégique. On estime en particulier à 65000 ha la surface nouvellement artificialisée tous les 8 ans, correspondant à un département français en moyenne.
Ainsi, l'action 7 de l'axe stratégique de reconquête de la biodiversité dans les territoires engage le gouvernement à « publier, tous les ans, un état des lieux de la consommation d'espaces et de mettre à la disposition des territoires et des citoyens des données transparentes et comparables à toutes les échelles territoriales » [1]. A cette fin, l'IGN a été sollicité avec l'IRSTEA et le CEREMA par la DGALN (Direction Générale de l'Aménagement, du Logement et de la Nature) pour créer et maintenir cet outil de suivi d'artificialisation des sols à l'échelle de chaque commune.
Le socle sur lequel se construit cet outil est l'OCS GE, recommandé par le rapport d'expertise scientifique établi par l'IFFSTAR (Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux) et l'INRA (Institut National de la Recherche Agronomique) [2], lui-même sollicité conjointement par les ministères en charge de l'Environnement et de l'Agriculture et par l'ADEME (Agence de l'Environnement et de la Maîtrise de l'Energie). Ce choix de l'OCS GE a été porté notamment en raison de (i) sa constitution et mise à jour possible tous les trois ans grâce aux campagnes d'acquisition d'images aériennes par l'IGN à cette même fréquence, (ii) sa nomenclature qui répond d'ores et déjà à la plupart des objets attendus. Toutefois, les coûts de production de l'OCS GE sont élevés à l'échelle nationale. Le défi est donc de rendre le processus le plus automatisé possible.
La définition de l'artificialisation est loin d'être unique et de faire l'objet d'un consensus au sein des diverses communautés impliquées dans l'étude ce phénomène. Afin d'initier ces travaux, l'équipe a décidé de se concentrer en majorité sur le cas du *bâti*, servant donc de vecteur primaire d'étude de l'artificialisation.

Les résultats obtenus en chapitre III sur l'analyse d'images satellites très haute résolution par apprentissage profond ont permis d'orienter cette automatisation vers lesdits algorithmes d'apprentissage profond. En revanche, pour des raisons de précision géométrique (délimitation des objets la plus fine possible), la donnée la plus résolue possible a été choisie pour cette tâche afin de parvenir aux résolutions de l'OCS GE et être ainsi compatible avec le

1. Plan Biodiversité du 4 juillet 2018.
2. Rapport de l'IFFSTAR et de l'INRA sur les processus d'artificialisation des sols et leurs suivis.

RGE. Les images utilisées sont donc les orthophotographies aériennes IGN, acquises sur trois ans sur le territoire, d'une résolution de l'ordre de 20 cm. Ce choix est en outre motivé par la disponibilité du MNS progressivement calculé conjointement avec les orthophotographies, qui, nous le verrons permet d'affiner sensiblement les résultats et notamment les géométries des objets détectés. Ces résultats sont en accord avec des travaux utilisant des classifieurs plus classiques montrant que le MNS permet de mieux discriminer sol et sursol (Rottensteiner, 2007 ; Le Bris et Chehata, 2011 ; Briottet et al., 2017). Dans un contexte où le but ultime est celui de quantifier une surface artificialisée, et sachant que celle-ci varie d'environ 65000 ha tous les ans, ramenée à la surface totale française, cela constitue moins d'1% de changement à détecter. Le pari est donc grand et utiliser un MNS ainsi que les données les plus résolues possibles pour le suivi de l'artificialisation est cohérent. Les travaux sur le sujet sont encore en cours à l'IGN, et nous présentons ici les méthodes mises en place et les résultats obtenus lors des premiers mois de ce projet. Le but final de ce dernier étant de mettre à jour / compléter une base de données existantes (OCS GE) et de suivre l'évolution de l'artificialisation des sols, plusieurs stratégies sont possibles : (i) proposer une solution totalement automatique de détection de changement, (ii) faire intervenir un opérateur dont la charge est de valider ou invalider la classification automatique (levé d'alertes par exemple) ou bien (iii) recalculer une carte remplaçant totalement le précédent millésime.

La genèse du projet explicitée dans les paragraphes précédents font de ce chapitre une étude totalement tournée vers des besoins opérationnels, avec un cahier des charges précis défini entre autres par la nomenclature OCS GE (et ses spécifications), pour lesquels les modèles d'apprentissage profonds ont rarement, voire jamais été utilisés. Le terme « opérationnel » renvoie au besoin de données plus résolue et à la réflexion faite sur la raison des expérimentations menées pour que celles-ci permettent de s'intégrer dans des chaînes de production.

Les résultats présentés dans ce chapitre sont le fruit de réflexions et d'expérimentations menés au sein et par l'équipe du projet dédié à cette action du plan à l'IGN, qui a été nommée TERMOS pour Télédection ExpérRiMentale pour l'Occupation des Sols. Je remercie beaucoup mes collègues de cette équipe, Camille Parisel, Nicolas David, Anthony Wiart et Sylvain Galopin dans ces travaux qui ont constitué une passerelle entre recherche, encore très présente dans ce chapitre, et réflexion pour une mise en production future.

2 Architectures utilisées

Nous avons rapidement évoqué cette nouvelle étude en fin de l'état de l'art en chapitre II. Plus particulièrement, nous y avons fait mention de l'architecture U-Net (Ronneberger et al., 2015).

Comme pour le chapitre précédent, on s'attache à définir des processus d'automatisation de cartographie d'occupation des sols, par le biais de l'utilisation de réseaux de neurones. Nous verrons qu'un travail conséquent de préparation des données est nécessaire afin de constituer des jeux d'apprentissage cohérents. A ce titre, les tests ont pour la plupart visé à qualifier et comparer diverses stratégies d'apprentissage à réseau fixé.
Le choix de U-Net est dû à ses résultats qui font l'unanimité au sein de la communauté de la télédétection, avec plusieurs résultats qui concordent sur ses performances pour l'analyse d'images aériennes (Huang et al., 2018). A l'inverse d'une approche « patch », U-Net s'inscrit dans la famille des réseaux encodeur-décodeurs prenant une image en entrée, et générant également une image de même dimension que l'image initiale en sortie de réseau. Si un réseau encodeur-décodeur suit en partie le même schéma qu'un réseau de classification d'image classique type VGG, avec des couches de « max-pooling » qui réduisent peu à peu les dimensions spatiales de l'image au travers du réseau (partie encodeur), il en diffère par l'ajout d'une étape de sur-échantillonnage progressif qui permet justement de retrouver les dimensions initiales de l'image. Ce sur-échantillonnage est conduit par un réseau, le décodeur, et est progressif car les dimensions spatiales sont accrues par étape, avec autant d'étapes de sur-échantillonage que d'étapes de *max-pooling* dans la partie encodeur.

L'architecture U-Net est visible sur la figure IV.1, et décrit l'implémentation originale du papier.
La première partie du réseau est l'encodeur et est constituée de blocs appelés « contractants » de part la présence de couches de *max-pooling* avec des fenêtres de 2, après deux couches convolutives. L'encodeur se charge d'extraire l'information la plus pertinente de l'image qui servira ensuite à labelliser chaque pixel de la manière la plus adaptée possible. Agissant comme un extracteur d'attributs (révélateurs des objets sur l'image initiale), l'encodeur produit des résultats spatialement très réduits après la succession de plusieurs couches de *pooling*, permettant une prise en compte du contexte accrue. Après chaque couche de *pooling*, le nombre de filtres est doublé pour accroître le pouvoir d'interprétation de l'encodeur.

C'est là qu'intervient le décodeur qui permet de produire en sortie une carte de mêmes dimensions que l'image en entrée de l'encodeur. Cette partie est constituée de blocs cette fois-ci dits « extensifs » avec une étape de sur-échantillonnage à l'issue du bloc. Cette couche de sur-échantillonnage est véritablement le pendant de la couche de *pooling* de l'encodeur puisqu'au lieu de ne conserver qu'un pixel de la seconde couche convolutive d'un bloc contractant sur un voisinage de 2×2, on génère un voisinage de 2×2 à partir d'un pixel à l'issue d'un bloc extensif. Les paramètres permettant de calculer la valeur de ces pixels sont également optimisables et sont appris au même titre que les couches de convolution plus classiques.
L'architecture propose d'adjoindre à chaque première couche couvolutive

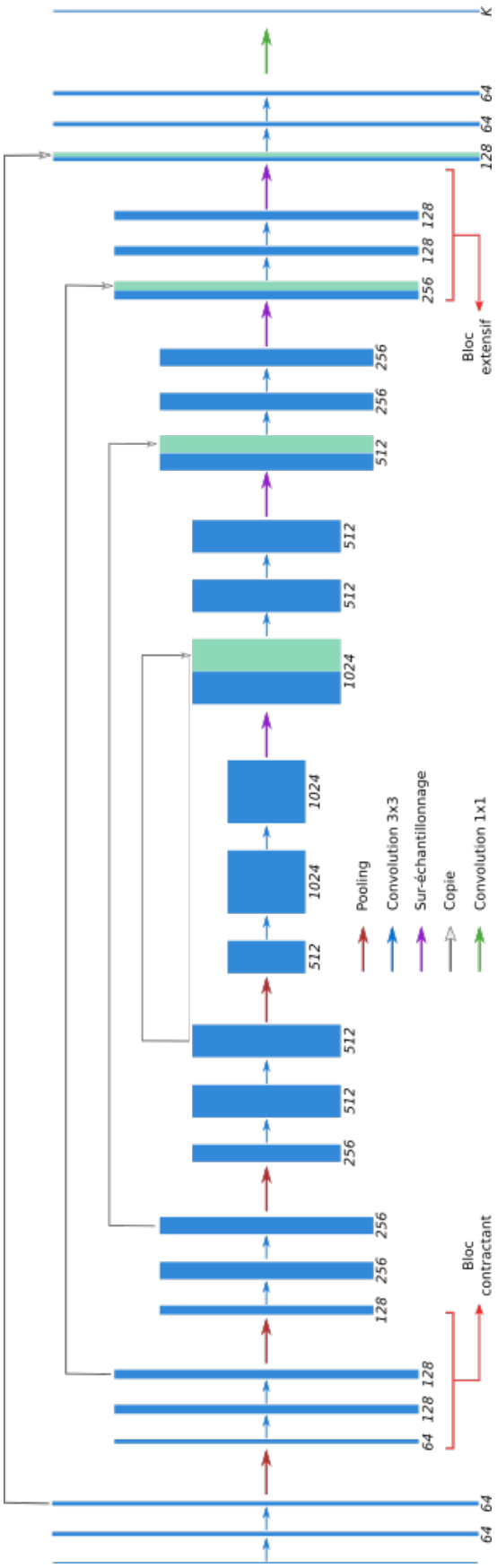


FIGURE IV.1. Architecture U-Net telle que décrite dans Ronneberger et al. (2015).
Un bloc contractant correspond à deux couches de convolution suivies d'une couche de « pooling ».
Un bloc extensif renvoie à deux couches de convolution suivies d'une couche de sur-échantillonnage.

d'un bloc extensif la carte d'attributs issue de chaque seconde couche convolutive du bloc contractant correspondant, par concaténation des cartes d'attributs. Ce procédé donne au décodeur accès aux différentes échelles d'analyse effectuées dans l'encodeur, guidant progressivement et efficacement le réseau pour qu'il produise des cartes de classification cohérentes et avec des contours bien localisés. Ces concaténations de cartes d'attributs dans la partie décodeur est spécifique au réseau U-Net.
Enfin, au terme du dernier bloc extensif de l'encodeur, une couche de convolution 1×1 permet de conserver l'information spatiale de chaque carte d'attributs, et de concaténer ces cartes vers une images à K canaux, chacun étant une carte de chaleur correspondant à l'un des classes de la nomenclature étudiée.
L'avantage de U-Net sur des approches *patch-based* utilisées en chapitre III réside dans l'exploitation intra-patch de l'information sémantique : ainsi, au lieu de fournir au modèle des échantillons dont les étiquettes correspondent au patch global, le réseau a connaissance du label au niveau d'un pixel. Une telle configuration affine spatialement puisque l'on contraint à l'intérieur d'un patch les relations topologiques entre objets géographiques (objets labellisés en *bâti* souvent adjacents à d'autres objets *bâti*).

Il est possible de créer un réseau type U-Net en enchaînant autant de blocs contractants que l'on veut au prix d'un accroissement du nombre de paramètres, des temps de calcul et de prérequis *hardware* accrus.
Ici, l'architecture U-Net est celle de la figure IV.1, mais avec un nombre de filtres initial de huit dans la première couche, réduisant grandement les contraintes citées précédemment, avec une une étape d'extraction d'attributs *a priori* moins efficace. Notons que les problèmes de sur-apprentissage sont également très limités (le réseau a environ 400000 paramètres), et que dans un objectif opérationnel, cartographier la France entière à 20 cm ou 50 cm de résolution doit être effectué dans des temps raisonnables.

L'architecture U-Net a été la plus abondamment utilisée dans les expériences qui suivent, offrant un compromis intéressant entre temps d'entraînement et performances. Toutefois, d'autres architectures de l'état de l'art ont été entraînées. En particulier, nous avons pu mettre à l'épreuve les réseaux MobileNetV2 (Sandler et al., 2018), architecture légère et pouvant être utilisée sur des appareils mobiles, et DeepLabV3+ (Chen et al., 2018d), ce dernier étant une amélioration de plusieurs versions successives du modèles DeepLab (Chen et al., 2014b ; Chen et al., 2017a ; Chen et al., 2017b), en incluant notamment une architecture encodeur-décodeur à la manière de U-Net. Ces architectures, plus profondes, sont également plus coûteuses en temps de calcul et en nombre de paramètres, avec deux millions de paramètres pour MobileNetV2 et vingt millions pour DeepLablV3+.
MobileNetV2 apporte des éléments constitutifs des réseaux de neurones différents de U-Net, notamment par le biais :

1. des *depthwise convolutions*, introduits par Chollet (2017), et réduisant considérablement le nombre de paramètres. Le principe de ce type de

convolution est de décomposer un filtre de convolution classique, par exemple de la forme $3 \times 3 \times N$, avec N le nombre de canaux de la carte d'attributs issue de la couche précédente, en plusieurs filtres de convolution. Les premiers agissent spatialement (filtres *depthwise*) uniquement, avec une taille de $3 \times 3 \times 1$ (pour ne pas changer la profondeur de la carte d'attributs en entrée de cette couche), on en a donc N, un pour chaque canal d'entrée. En empilant les N canaux issus des convolutions précédentes, on obtient une carte d'attributs de profondeur N. Le second type de filtres *pointwise* et agit sur la carte d'attributs précédente, avec une dimension $1 \times 1 \times N$, le filtre *pointwise* est donc unique dans cette décomposition d'un filtre $3 \times 3 \times N$.

2. des couches résiduelles sont également ajoutées, facilitant l'entraînement et limitant fortement le problème de gradients évanescents ou explosifs.

De son côté, DeepLabV3+ est le fruit d'itérations d'une architecture au cours du temps avec notamment l'ajout successif de :

1. *Atrous Spatial Pyramid Pooling* (ASPP), dont l'intérêt est double :
   (i) la dimension *atrous* qui augmente le *receptive field* d'un filtre en effectuant des convolutions dilatées. Par exemple, on garde le nombre de neuf paramètres d'un filtre 3×3, mais en ne prenant qu'un pixel sur deux dans l'ensemble des directions spatiales ; le filtre s'étend alors sur un voisinage 5×5.
   (ii) l'utilisation de pyramides spatiales permet d'analyser un voisinage avec plus ou moins de recul et de capturer des informations de contexte plus riches qu'avace une convolution à une seule échelle spatiale ;
2. du mécanisme de *depthwise convolutions* pour découpler les convolutions spatiales de convolutions en profondeur sur les ASPP, réduisant considérablement le nombre de paramètres comme pour MobileNetV2.

3 Création des jeux d'apprentissage

Les images utilisées sont celles de la BD Ortho IGN, et les bases de données utilisées sont celles déjà passées en revue dans les chapitres I et III. Nous nous intéressons donc dans cette section à fournir des informations sur (i) les orthophotographies elles-mêmes et les zones d'intérêt ainsi que (ii) la stratégie de création de jeux d'apprentissage.

3.1 A propos des orthophotographies et des régions d'étude

Les orthophotographies sont disponibles depuis la BD Ortho IGN à la résolution de 20 cm. Le choix de cette source d'image est double : (i) la résolution qui donne accès à une finesse autrement non accessibles pour détecter des objets fins, et pour obtenir des formes d'objets les plus proches possibles de ceux présents dans les bases de données (un bâti y est souvent représenté par une forme rectangulaire), (ii) l'accès à un MNS pour lever certaines ambiguïtés est coûteux mais très utile, or ce MNS a déjà été dérivé des prises de

vues aériennes utilisées pour produire la couche d'orthophotographies. Notons que le MNS, à l'instar des orthophotographies, sont codés sur 255 bits, et donc pour privilégier la plage d'information pertinente de hauteur sur les zones étudiées, peu élevées, les hauteurs sont ré-étalées sur une échelle logarithmique afin de favoriser les hauteurs basses à moyennes.
La couverture nationale est effectuée en trois ans, ce qui suggère une actualité hétérogène non seulement au sein de la couverture de l'intégralité du territoire, mais également entre la couverture et les bases de données géographiques. C'est un problème auquel nous étions déjà confrontés dans le chapitre III, mais avec un impact plus fort ici comme nous le verrons plus loin dans le paragraphe concernant la constitution de jeux d'apprentissage.

Avec des données respectant ces contraintes de cohérence temporelle (un an d'écart entre les prises de vue aériennes et la mise à jour des bases de données topographiques de l'IGN) sur le département de la Vendée (85), celle-ci a été la zone de départ pour nos tests. La raison supplémentaire tient au fait que l'OCS GE est disponible sur cette zone depuis 2013 (figure IV.2). Ce chapitre s'intéresse à la mise à jour de données pour le suivi de phénomène (artificialisation), plusieurs millésimes sont donc étudiés, en particulier, les couches orthophotographies sont celles de 2013 et 2016.
Outre le département de la Vendée qui constitue notre socle pour cette étude, plusieurs zones sur divers départements ont été ponctuellement classifiées (notamment sur les départements suivants : Bouche-du-Rhône (13), Haute-Garonne (31), Jura (39)).

(A) orthophotographie de 2016.

(B) OCS GE : ● *Zones bâties*, ● *Zones non bâties*, ● *Formations Herbacées*, ● *Feuillus*

FIGURE IV.2. orthophotographie aérienne et OCS GE correspondantes sur une zone semi-urbaine de la Vendée.

3.2 Extraction des échantillons et classes d'intéret

A l'image de tout processus d'apprentissage, l'étape de jeux d'apprentissage est cruciale pour obtenir un modèle cohérent avec l'objectif visé. Cet aspect est exacerbé avec les contraintes que l'on a dans ce cadre d'étude en terme de nomenclature et de spécifications OCS GE. En l'occurrence, les

données d'apprentissage étant extraites des bases de données géographiques déjà évoquées et re-détaillées plus loin, il faut pouvoir faire le lien entre ces données disponibles et la nomenclature voulue en bout de chaîne. A moins de précisions contraires, l'ensemble des expérimentations a été mené en utilisant les orthophotographies et l'information de hauteur correspondante lorsque celle-ci était disponible sur la zone d'intérêt : cette hauteur est générée en soustrayant le MNS au MNT local afin de ne conserver que le sursol. Le MNT provient du RGEAlti. Les images sont utilisées :

— soit à la pleine résolution de 20 cm, pour lesquelles on a construit des imagettes d'entraînement de taille 256×256 pixels. Cette résolution est intéressante pour des cas où la texture est très discrimante ou pour lesquels l'information de sursol (en terme de hauteur) n'est pas fiable ;

— soit à une résolution sous-échantillonnée à 50 cm. Le cas échéant, les imagettes sont dimensionnées à 128×128 pixels afin de conserver une empreinte au sol de même étendue que pour le cas à une résolution de 20 cm. En particulier, la couverture d'images en bande infrarouge sur certains départements n'est pas disponible à la résolution de 20 cm, notamment sur le département du Jura. Un sous-échantillonnage permet de donner à chaque filtre du réseau un contexte spatial plus riche qu'une image à pleine résolution.

Au même titre qu'en chapitre III, les bases de données géographiques qui ont servi d'apprentissage sont :

— la BD Topo pour les classes relevant des classes artificialisées : *bâti*, *pont*, *voies ferrées*, etc ;

— le RPG pour les classes de cultures ;

— le RGFor pour les classes de *ligneux*. Le RGFor est réalisé manuellement à l'IGN depuis 2006, et présente donc une très forte hétérogénéité dans l'actualité des données disponibles. Il demeure cependant l'unique source depuis laquelle on peut obtenir une couche de végétation cohérente sur l'ensemble du territoire.

Les divers jeux d'apprentissage sont constitués dans l'objectif d'utiliser le réseau U-Net, et nécessitent donc des échantillons dont chaque pixel est étiqueté, par opposition aux jeux construits dans le chapitre III qui font correspondre à un échantillon une étiquette unique pour l'ensemble des pixels.

Nous abordons maintenant la question de l'intégration de la hauteur dans nos échantillons d'apprentissage. Il est possible de l'intégrer par concaténation aux canaux radiométriques, ou bien en traitant séparément les deux modalités. Cette dernière méthode permet de donner autant d'importance aux deux sources de données. Hazirbas et al. (2016) propose ainsi une structure appelée FuseNet, à double entrée, chacune accueillant une modalité et offrant également une structure encodeur-décodeur. On a ainsi deux *streams* d'information pour la partie contractante (deux encodeurs, un radiométrique, l'autre pour la hauteur), puis un seul pour la partie extensive. Les attributs appris dans le *stream* dont l'entrée est le MNS sont progressivement ajoutés par addition aux attributs radiométrique. Cette concaténation s'effectue

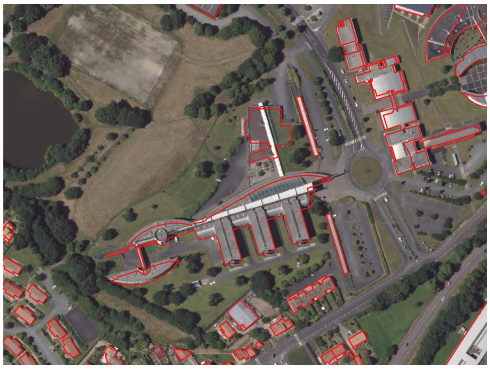
après activation des filtres dans chacun des *streams*, permettant ainsi à chaque modalité de contribuer équitablement. Une expérimentation a été menée avec cette architecture mais les résultats n'étaient que très préliminaires.
L'ensemble des travaux présentés ont donc consisté à construire au préalable des échantillons concaténant radiométrie et hauteur en une seule image.

La thématique principale étant l'artificialisation du sol, une attention particulière a été portée sur la classe *bâti* et une première étude a consisté à produire une segmentation binaire *bâti / non bâti*. Les deux classes ont été constituées ainsi :

1. les échantillons de *bâti* agrègent l'ensemble des types de bâti disponibles dans la BD Topo :
   — *bâti indifférencié* regroupant toute la couverture d'habitations (immeubles, zones pavillonnaires, ...) ;
   — *bâti remarquable* tels que les églises et monuments ;
   — *bâti industriel*
2. la classe *non bâti* regroupe l'ensemble du territoire n'étant pas bâtie. Des objets autres que bâti issus des différentes bases de données ont été échantillonnés afin de représenter au mieux cette classe. Toutefois, présenter une exhaustivité des objets de cette classe est ardu, la définition de *non bâti* pouvant renvoyer à n'importe quel objet. C'est pourquoi nous parlions précédemment de processus itératif : après des tentatives d'entraînement et de classification, les confusions entre *bâti* et *non bâti* ont permis de recenser d'autres objets qu'il n'étaient *a priori* pas évident à première vue d'intégrer à notre apprentissage. Nous creuserons davantage cette direction prise dans la section consacrée à la détection du bâti qui suit.
   Il faut toutefois souligner le problème lié à cette classe résidant dans le fait qu'elle contient des types d'objets absents des bases de données, à l'image des plages.

4 Expérimentations

L'essentiel des travaux effectués dans le cadre de ce chapitre concerne le *bâti* pour les raisons citées précédemment. En revanche, dans un objectif d'accroissement de la nomenclature, cette dernière a été étendue à un cas multiclasses sur le département de la Vendée. Cela permet notamment de comparer et d'évaluer l'enrichissement d'une nomenclature à de nouvelles classes par rapport à un cas de classification purement binaire *bâti / non bâti*. Un cas de détection de vignes sur plusieurs départements est également présenté.
Les résultats sont présentés sous formes de cartes de chaleur dans les cas binaires, représentant le score d'appartenance à la classe d'intérêt, score issu de la couche de softmax du réseau utilisé. Pour les cas multiclasses, on propose comme en chapitre III, des cartes de classification, avec la classe représentée correspondant à celle dont le score est maximal en sortie du réseau.



(A) Dévers sur les bâtiments les plus éloignés du nadir d'acquisition.

(B) Dévers nul au centre, mais plus prononcé pour les objets élevés.

(C) La cohérence temporelle entre base de données et image n'est jamais parfaitement vérifiée.

(D) Le bâti sous couvert forestier n'est pas vu par un capteur optique opérant depuis le ciel.

FIGURE IV.3. Contraintes liées à l'utilisation de bases de données géographiques : conséquence d'une acquisition aérienne et d'erreurs de la base elle-même.
L'apprentissage du modèle est tributaire des conditions d'acquisition et de la cohérence géométrique et temporelle entre les deux sources de données.

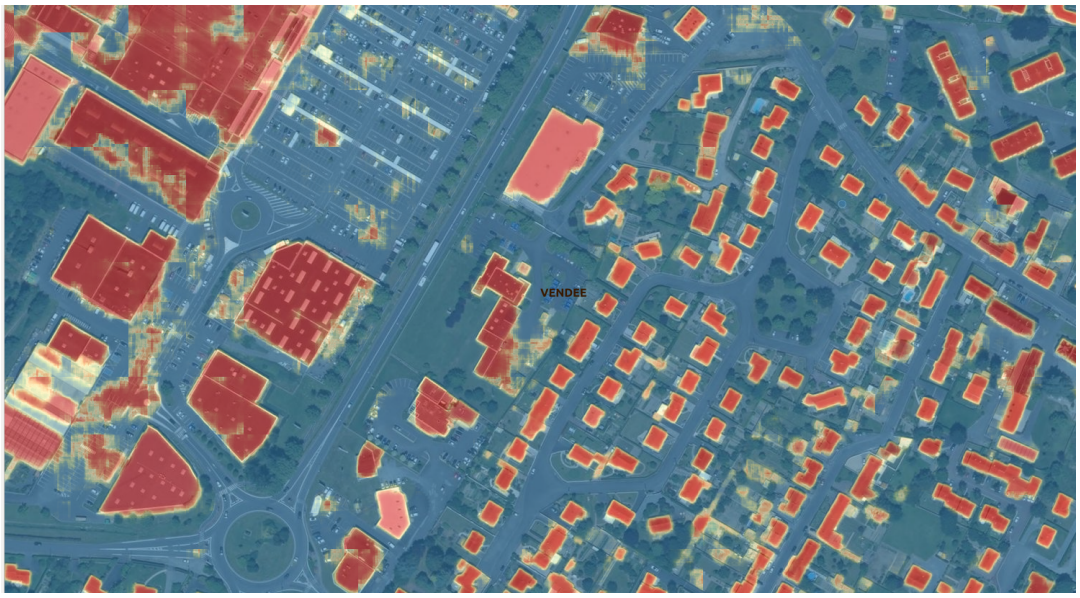
4.1 Classification du bâti

Les résultats affichés avec les images SPOT sont prometteurs, motivant ce travail plus approfondi dans un contexte de mise à jour de bases de données topographiques existantes. A cette fin, pour des soucis de compatibilité avec ces bases de données, on préfère ici les images aériennes offrant un niveau de détail bien plus fin que les images satellites, et pour lesquelles on peut disposer d'un MNS cohérent temporellement avec ces dernières.
Ne nous attachant qu'à la classe *bâti*, les images aériennes sont échantillonnées à la pleine résolution de 20 cm. Sur l'analyse de la compatibilité entre bases de données et images aériennes, la figure IV.3 donne une illustration des difficultés que l'on rencontre lorsque l'on cherche à constituer un jeu de données d'apprentissage. Les réseaux de neurones nécessitant un volume important de ce type de données, il est impensable de valider la cohérence et l'actualité sémantique de chaque base de données géographiques en la

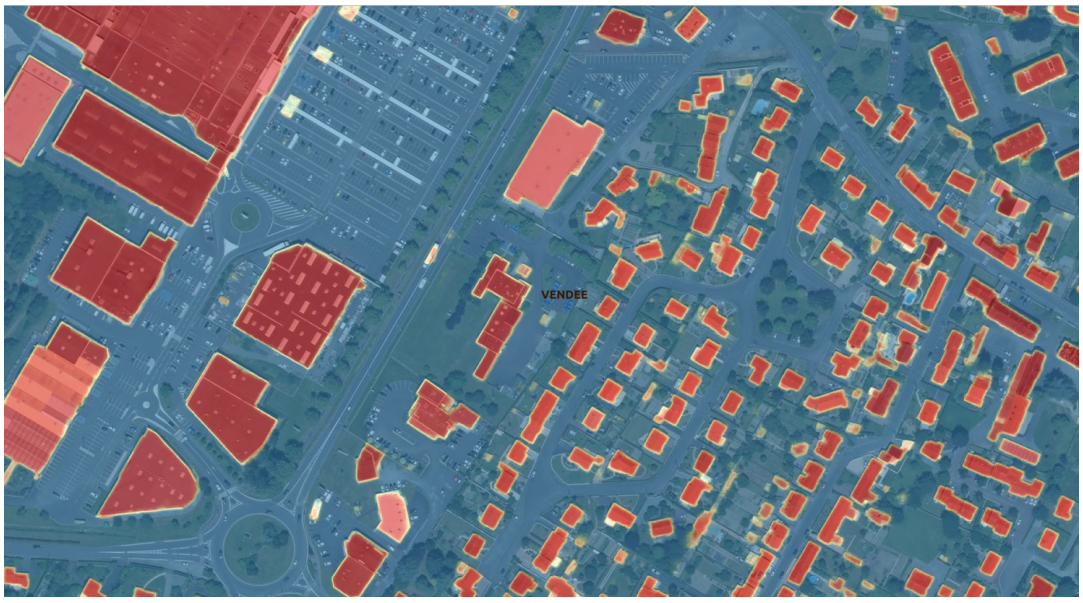
comparant aux orthophotographies, sur l'ensemble du département. La qualité des échantillons d'apprentissage, de taille 256×256, est donc tributaire de ces soucis. Ainsi, d'un point de vue géométrique, les bâtiments (et l'ensemble des objets, mais nous nous intéressons ici au bâti plus particulièrement) sont sujets au phénomène de dévers caractérisant les prises de vue aériennes par les capteurs matriciels, qui en donnent une représentation en projection perspective. Cet effet de dévers est d'autant plus important que les bâtiments sont (i) élevés et (ii) éloignés du centre de l'image. En effet, les objets au nadir d'acquisition d'une image présente un dévers nul (les orthophotographies « vraies » sont équivalentes à des images dont chaque pixel est vue au nadir d'acquisition, par redressement selon le MNS). De son côté, la BD Topo, intégrant la couche de la classe *bâti*, délimite les bâtiments selon leurs empreintes au sol. Le dévers peut donc mener à une représentation biaisée du bâti au moment de la création des patchs d'apprentissage, avec des pixels labellisés comme du *non bâti* sur des pixels de l'image qui représentent des bâtiments sujets au dévers. Cela crée un décalage entre l'objet présent sur l'image et l'objet décrit par la BD Topo.
Les problèmes d'actualité des données peuvent également conduire à des étiquettes erronées, comme pour la destruction du bâti sur l'image IV.3c. Enfin, implicitement, l'imagerie optique aérienne (au même titre que l'imagerie spatiale) ne représente que les objets visibles en premier lieu, et le couvert forestier peut masquer d'éventuels objets situés en deçà, c'est le cas de la situation IV.3d.

Pour en revenir à la stratégie de création d'un jeu de données sur la classe *non bâti*, les objets qui pouvaient poser problème lors de nos entraînements et prédictions successifs sont : *laisse de mer, voies ferrées, ponts, routes, champs cultivés* et *surface d'eau*.
En particulier, les ponts, qui présentaient de fortes confusions avec le bâti ont été échantillonnés pour lever et contraindre le problème. Cette confusion initiale s'explique par la conjugaison de l'aspect sombre de la route sur les ponts qui se rapproche de la radiométrie et de la texture d'un toit, et de l'information de hauteur (MNS-MNT) qui est non nulle (à l'inverse d'une hauteur de route) et qui renvoie donc à un objet élevé. Nous avons procédé de même avec les voies ferrées, ajoutées également au jeu d'apprentissage et de la laisse (correspondant à l'espace découvert entre marée haute et marée basse).

Le jeu de données total est constitué de 50000 échantillons (patchs de 256×256 pixels), répartis géographiquement sur l'ensemble du département avec 10000 échantillons pour la classe *bâti* et 40000 divisés sur l'ensemble des sous-classes constituant la classe *non bâti*. La répartition parmi ces classes est relativement équilibrées, mis à part pour les ponts pour lesquels on n'a moins de candidats. Cet échantillonnage régulier sur le département, à l'instar de la stratégie adopté en chapitre III, aide à construire un jeu représentatif d'un maximum d'objets de chaque classe. L'entraînement a été conduit pendant 10h environ.



(A) Sans utilisation du MNS.



(B) Avec utilisation du MNS.

FIGURE IV.4. Détection du bâti en environnement urbain sur la Vendée. Le réseau a été entraîné sur des échantillons issus de l'ensemble du département. L'utilisation du MNS dans l'apprentissage et la prédiction permet de s'affranchir de la plupart des fausses détection.

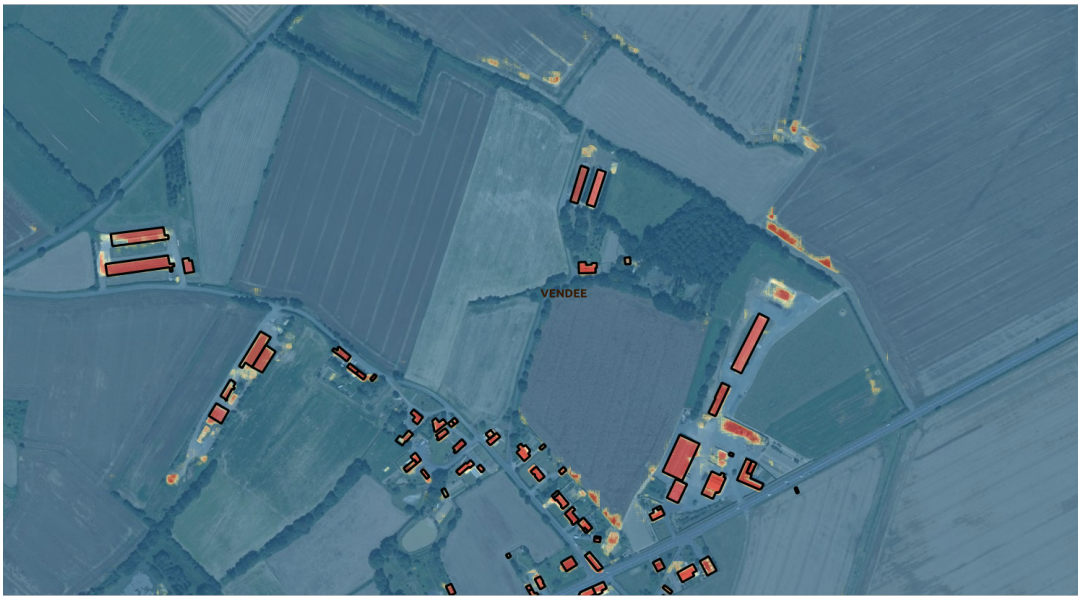
La figure IV.4 illustre les premiers résultats que l'on a eus avec notre réseau ainsi entraîné sur zone urbaine dense. On peut voir l'effet bénéfique de l'ajout du MNS, très discriminant pour lever les ambiguïtés sur les zones dont la texture et la radiométrie s'approchent de celles des échantillons de la classe *bâti* dans le jeu d'apprentissage. Le résultat, sous forme de carte de chaleur, est très vraisemblable et semble adapté à fournir une représentation cartographique de la place du *bâti* sur le territoire. Chaque bâtiment étant individualisé, il est possible d'inventorier précisément, dans ce cas et cette configuration, chaque objet, et de dresser son historique (en terme de date de construction) par rapport à une base de données topographiques datant d'un autre millésime.
Sur des cas moins favorables, tels que ceux représentés sur la figure IV.5, la bâti est soit sur-détecté, soit sous-détecté. Toutefois, ces situations étaient très prévisibles dans le sens où elles renvoient à des comportements classiques de détection d'objets non inventoriés lors de l'apprentissage ou non visibles. On a par exemple, le cas du couvert forestier qui est un obstacle évident à la détection du bâtiment situé en dessous, limitation qui est inhérente à une acquisition en vue du dessus. Par ailleurs, le cas de sur-détection sur les campings relève d'une spécification liée à la BD Topo qui considère que ces constructions ne sont pas du bâti. Ce choix peut être motivé pour des raison d'usages différents de ceux d'habitats plus courants (pavillons, immeubles), mais dans un contexte d'artificialisation, on peut se poser la question de la pertinence de détecter ce type d'objet. Enfin, la sur-détection sur les gravats de l'image de gauche est symptomatique du problème lié aux données manipulées, puisque ce sont des objets non pérennes, et voués à disparaître à court terme mais qui apparaissaient à la date d'acquisition. Avec des séries temporelles, ce type d'objet peut être par exemple filtré, sauf que de telles résolutions ne sont pas disponibles à plusieurs dates rapprochées. Il est en outre impensable de chercher à constituer une base de données pour ce type d'objets, justement parce qu'ils sont non pérennes et donc les inventorier est impossible. Toutefois, pour d'autres classes, cette fois-ci pérennes, mais potentiellement sur-détectées comme du bâti, il est possible de lancer une campagne de labellisation si celle-ci n'existe pas à l'heure actuelle, mais avec un coût humain non négligeable. A l'image du constat observé sur les tas de gravats sur-détectés, plusieurs types d'objets *non bâti* (par exemple les ponts) sur-détectés, mais cette fois-ci présents dans les bases de données existantes, ont été ajoutés dans cette classe après avoir vu la confusion que ces objets apportaient. Notons que le bâti a, lui, été bien détecté, non seulement avec peu de sur et sous-détection, mais également avec un détourage des objets très satisfaisant, dont on fournit une représentation vectorisée.

Au même titre qu'en milieu urbain, l'exploitation du MNS par le réseau conduit à une différenciation très efficace entre le bâti et d'autres classes non bâties en milieu rural. L'utilisation de la modalité radiométrique seule sur la figure IV.6 produit des confusions importantes au niveau de pixels correspondant à une partie de parcelles de champs. Les délimitations noires correspondent aux objets bâtis référencées dans la BD Topo IGN.
Toutefois, nous avons pu constater parfois des défauts du MNS sur certaines

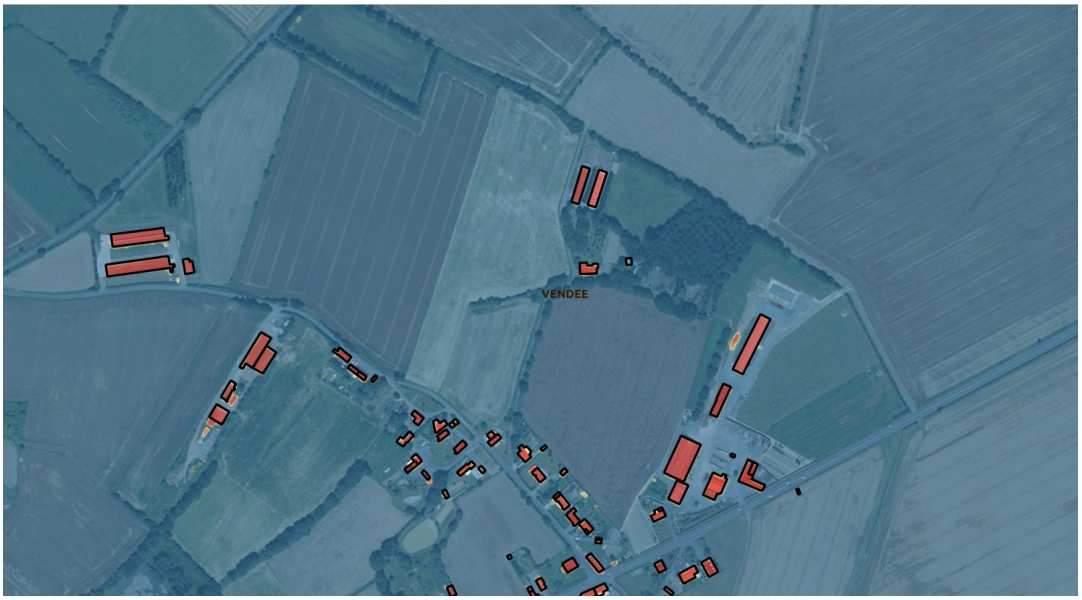
FIGURE IV.5. De gauche à droite : sur-détection sur des tas de gravats sur un chantier, sur-détection de mobil-home/bungalow dans un camping, sous-détection de bâti sous couvert forestier.

régions en milieu rural. La raison principale réside dans la corrélation erronée sur des parcelles de champs non ensemencées : celles-ci présentent des textures très répétitives le long des sillons par exemple. Or, tenter de corréler deux images sur ce type d'objet nous place dans un cas très défavorables, pouvant produire un MNS incohérent avec la topographie réelle du territoire (la différence entre MNS et MNT devant être quasi-nulle sur des parcelles nues).

Les résultats affichés sur le tableau IV.1 correspondent aux pourcentages de sous-détection et de sur-détection du bâti. Environ 25000 patchs ont été soumis au réseau pour réaliser cette analyse quantitative, correspondant à une surface de 65 km^2. Le réseau de neurones produisant des cartes de chaleur, avec pour chaque pixel sa probabilité d'appartenance au bâti, on peut faire varier le seuil de détection pour pénaliser plus ou moins les objets détectés comme du bâti. Ce tableau souligne deux faits. Le premier, évident et en accord avec les résultats visuels, est l'impact du MNS sur les performances globales de détection, les deux dernières lignes renvoyant à une prédiction utilisant le MNS, les deux premières sans MNS : que l'on raisonne en terme de bâti non détectés ou sur-détectés, le MNS améliore dans les deux cas les métriques. Par ailleurs, avec ou sans MNS, on observe une logique (i) croissance du taux de sous-détection (ii) une décroissance du taux de sur-détection, en augmentant le seuil en niveau. Cette observation est logique, puisque qu'on garde de moins en moins de bâti lorsque ce seuil augmente : les objets qui ont un score inférieur au seuil sont écartés, avec parmi eux des sur-détection (baissant donc ce taux), mais aussi des bâtiments réels (qui passent donc en sous-détection). Ce que l'on constate est que ces évolutions inverses entre taux de sous-détection et taux de sur-détection indiquent qu'il



(A) Modèle sans utilisation du MNS.



(B) Modèle avec utilisation du MNS.

FIGURE IV.6. Détection du bâti en environnement rural sur la Vendée. Les ambiguïtés en bordure de champ et de forêt sont levées grâce à l'information de hauteur très discriminante pour la classe *bâti*.

Seuil de détection		25	50	100	150	200
Sans MNS	Sur-détection (%)	2	1,7	1,15	0,95	0,8
	Sous-détection (%)	0,95	1,45	2,15	3,15	4,8
Avec MNS	Sur-détection (%)	0,5	0,4	0,3	0,25	0,2
	Sous-détection (%)	0,75	0,9	1,05	1,45	2,05

TABLE IV.1. Taux de sous-détection et sur-détection du bâti sur le département de la Vendée, en fonction du seuil (niveau de gris) et de l'utilisation ou non d'un MNS.

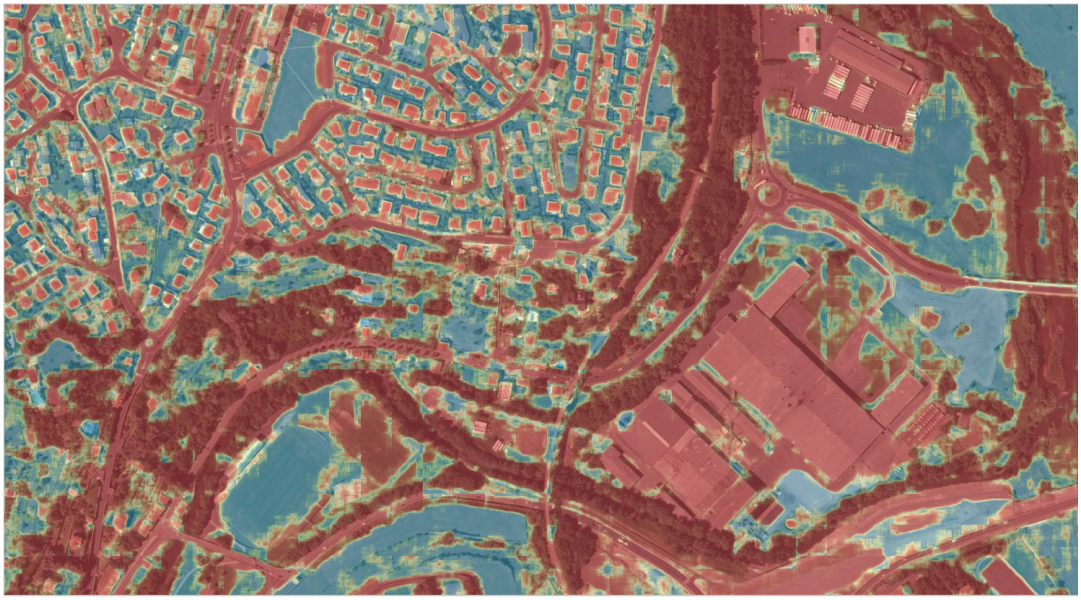
est difficile de limiter les sur-détections tout en limitant les omissions de bâtis. Cela renvoie au compromis qu'il faut souvent faire entre privilégier le rappel et la précision dans une classification automatique.
On le rappelle, ces taux sont également tributaires de la qualité des bases de données topographiques (destruction, et surtout construction de bâtiments qui n'apparaissent pas sur la BD, géométrie et spécifications absentes de la BD, telles que les cabanes ou camping), et des conditions d'acquisition avec des bâtiments occultés par de la végétation par exemple.

Enfin, pour mettre à l'épreuve le réseau appris sur la Vendée sur des zones totalement nouvelles, des classifications ont été effectuées sur la ville du Puy-en-Velay (Haute-Loire) et de Marseille (Bouches-du-Rhône). Le réseau U-Net entraîné sur la Vendée y a été appliqué directement, sans phase de réapprentissage. Le Puy-en-Velay présente des structures urbaines similaires à celles présentes sur le territoire vendéen, avec un développement davantage horizontal que vertical (bâtis pavillonaires, peu élevés), là où Marseille est très dense et intégrant des immeubles élevés et rapprochés (dont les toitures peuvent être de nature différente que de la tuile, très présente en Vendée), donc très différent de la Vendée.
La zone test du Puy-en-Velay a fait l'objet d'une étude double :

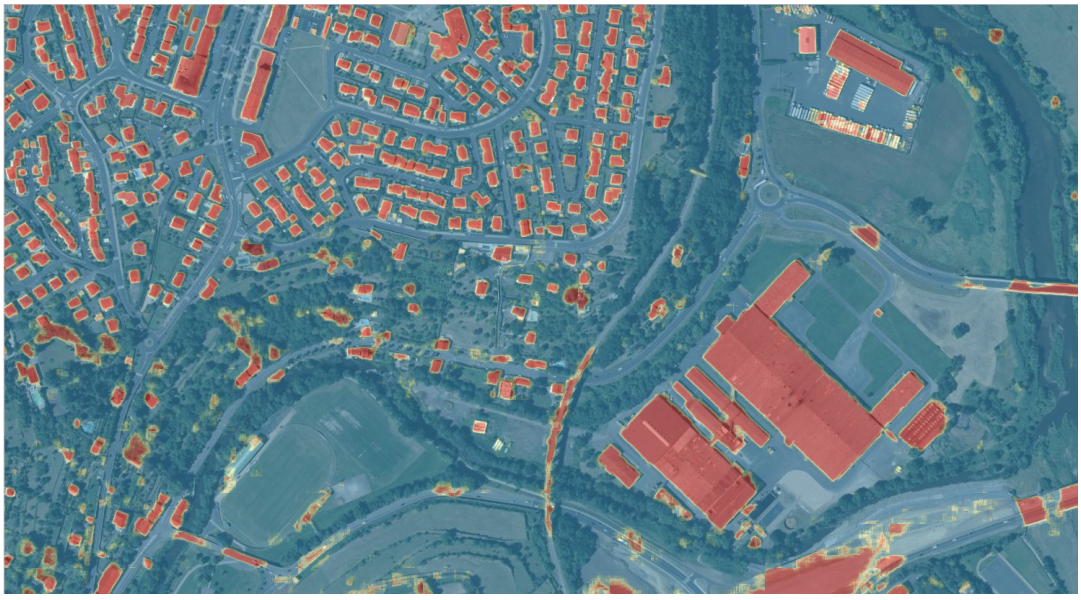
1. la transférabilité du modèle dans un biome différent ; la Vendée étant en milieu côtier et relativement plat topographiquement, on s'éloigne de ce cas sur cette nouvelle zone plus proches des montagnes. Toutefois, ne considérant que la classe *bâti* dans notre nomenclature, on s'affranchit dans une certaine mesure des soucis de végétation différente par exemple. « Dans une certaine mesure » seulement puisque les périodes d'acquistion d'images peuvent varier d'une région à une autre. Ainsi, même si la végétation n'est pas étudiée ici, les conditions d'éclairement ou l'état de la végétation elle-même peut présenter des caractéristiques nouvelles au réseau pré-entraîné et donc induire un comportement inattendu de la part de celui-ci (on renvoie au cas de prédiction sur la Gironde depuis un réseau entraîné en Bretagne en chapitre III). De plus, les toitures en Vendée et en Haute-Loire sont, pour la plupart, en tuiles, permettant *a priori* d'établir un lien entre ces deux régions en terme de bâti.
2. disposant d'un MNS sur la Haute-Loire, deux classifications ont été menées sur le Puy-en-Velay afin de confirmer si l'information de hauteur

d'objet est encore une fois cruciale au même titre que sur la région initiale de la Vendée.
Le MNS sur la ville de Marseille n'était en revanche pas disponible.

Les résultats sur le Puy-en-Velay et Marseille correspondent aux figures IV.7 et IV.8.
A l'instar de la classification sur la Vendée, les erreurs correspondent à des sur-détections sur les deux nouvelles zones. La classification sur le Puy-en-Velay présente initialement de fortes confusions *bâti / non bâti*, plus particulièrement avec la végétation (en bord de bâtiment ou également ailleurs, pour la raison évoquée dans le point 1. du paragraphe précédant immédiatement celui-ci) et les routes. On constate que l'ajout du MNS lève de nombreuses ambiguïtés sans résoudre toutefois l'ensemble des soucis. Les ponts demeurent mal classés ainsi que certains objets particuliers à l'image de camions regroupés sur un parking (Nord-Est de l'image IV.7b, qui présentent une texture proche de bâtis adjacents, avec une différence MNS-MNT non nulle. En revanche, les objets relevant de la classe *bâti* sont à nouveau bien détectés d'un point de vue délimitation géométrique et en terme de sur et sous-détection.
La détection sur la ville de Marseille présente les mêmes confusions avec beaucoup de sur-détections sur des routes et des parkings. Par ailleurs, le voisinage de végétation haute aux abords immédiats des habitations est source de mauvaises classifications. Les résultats étant sous forme de carte de chaleur sur la classe *bâti*, avec un seuil fixe il est possible de filtrer un certain nombre de pixels dont le score d'appartenance est faible. Cela apporterait également une meilleure délimitation des bâtis proches les uns des autres. Toutefois, il y a aussi des sur-détections pour lesquelles le réseau produit des scores d'appartenance élevés. Sans MNS sur la ville de Marseille, nous ne pouvons mener une étude comparable à celle faite sur la zone du Puy-en-Velay, mais au vu des résultats sur cette zone et sur la Vendée, il est avéré que le MNS améliorerait la détection.



(A) Sans utilisation du MNS.



(B) Avec utilisation du MNS.

FIGURE IV.7. Détection du bâti sur la ville du Puy-en-Velay, dans le département de la Haute-Loire (43), avec le modèle appris sur la Vendée.

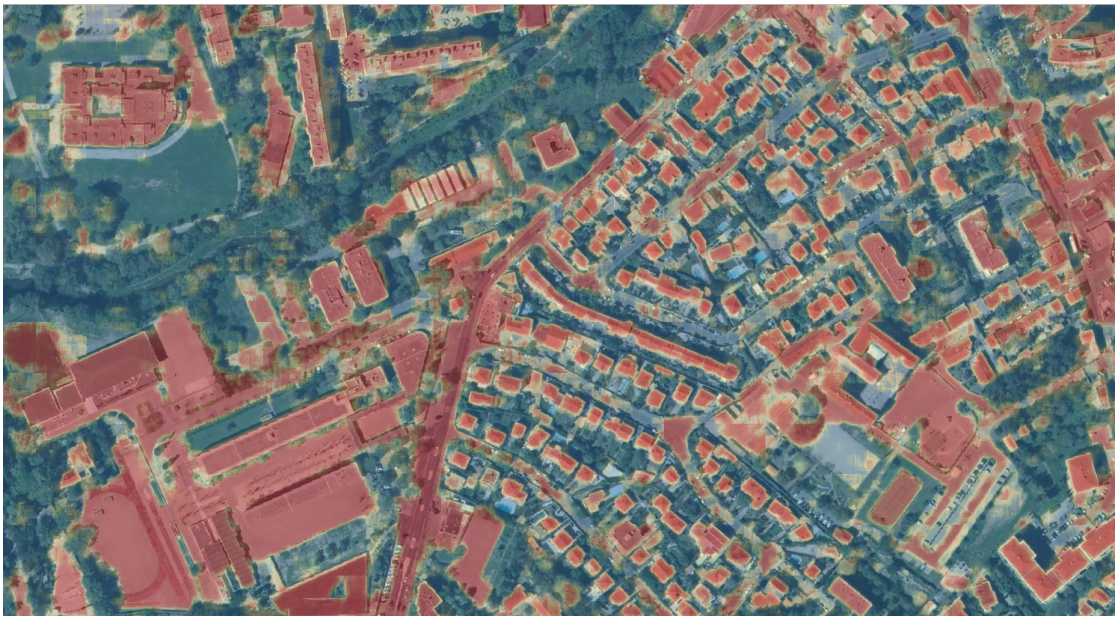


FIGURE IV.8. Détection du bâti sur Marseille avec le modèle appris sur la Vendée, sans MNS (non disponible).

Les études sur une détection du bâti uniquement montrent des résultats très prometteurs, une fois que l'on a ciblé un certain nombre d'objets pouvant être sources de confusion (ponts). Dans le but d'approcher au mieux un aspect cartographique, l'utilisation de MNS est indispensable. Enfin, les résultats sur le *fine-tuning* du chapitre III, même si issus d'une étude sur images satellites, indiquent que cette technique aurait un impact très clairement bénéfique lorsque l'on en vient à classifier des zones dont les biomes diffèrent de la zone initiale d'entraînement, bien que les erreurs soient essentiellement du type sur-détection plutôt que sous-détection.

4.2 Détection multiclasses

Afin d'étendre la nomenclature, cette section s'attache à une classification non plus binaire mais sur quatre classes d'intérêt et une classe correspondant au reste du paysage : *bâti, ligneux, pont, eau, autre*, avec la classe *autre* regroupant les sous-classes *laisse, champs cultivés, voies ferrées, routes*. Cette nomenclature n'est pas courante, avec une fusion des classes *routes* et *champs*, pourtant différentes, et la classe de *pont*, peu fréquente, que l'on garde à part entière. On étudie ici la possibilité du réseau à modéliser un cas multiclasses, et on dispose déjà de jeux d'apprentissage pour celles de la nomenclature établie. De plus, on rappelle que les ponts constituaient une source importante de confusion : les classifier à part permet de renforcer leur discrimination. La classe de végétation est notamment importante pour mesurer la présence de celle-ci en milieu urbain, qui peut se traduire par des parcs, allées d'arbres ou même jardin. Toutefois, comme dans le chapitre III nous ne disposons que du RGFor, dont les jardins sont absents, de même que les arbres individuels en ville, ces objets ne rentrant pas dans les spécifications d'inventaire de la

végétation.
Plusieurs expérimentations ont été conduites pour le cas multiclasses, chacune avec plusieurs architectures :

1. U-Net uniquement avec des images à 50 cm ;
2. U-Net, MobileNetV2, DeepLabV3+ avec des images à 20 cm.

Cela permet de jauger l'impact de la profondeur du réseau et du nombre de paramètres dans ce dernier sur la qualité de la classification. L'ensemble des jeux d'apprentissage utilisent des données du département de la Vendée et de la Haute-Garonne, pour diversifier les instances de chaque classe du point de vue spectral et géométrique. Les divers modèles peuvent ainsi être plus robustes aux changements d'environnement.

Pour la première expérimentation, le jeu d'apprentissage est constitué de 50000 images dont 36000 sont utilisés à l'entraînement, le reste pour valider le modèle.
Nous avons présenté la méthode d'extraction d'échantillons dans la partie 3.2. Ici les patchs sont générés à la résolution de 50 cm : la raison principale est l'étendue géographique très variable des classes à détecter. encore une fois, ce choix a été déjà justifié en chapitre III, dans lequel nous évoquons le fait que, comme dans ce paragraphe, on cherche aussi bien à retrouver la végétation que des bâtis isolés. Si cette résolution peut laisser penser que l'on dégrade la détection du bâti, il faut modérer cette croyance.
En effet, les bâtiments sont des objets dont l'organisation n'est pas aléatoire sur le territoire, à l'inverse d'objets naturels. Leur contextualisation pour un classifieur est donc très importante ; un bâtiment est susceptibles d'être proche d'une second bâtiment ou même d'une route. Le réseau répondant à un objectif contraint en terme matériel et de temps de calcul, le nombre de filtres est réduit, on compense donc cela par un enrichissement du *receptive field*, évoqué dans le chapitre II, dès la création des patchs. De plus, une autre raison justifie l'impact réduit d'une résolution de 50 cm ; le calcul de la hauteur repose sur la qualité du MNS, qui est construit par corrélation dense d'images stéréoscopiques, consistant à retrouver corréler les points d'un même objet physique au sol entre deux images successives sur la trajectoire de vol de l'avion. Or, cette corrélation est souvent mal déterminée sur les ruptures franches de pentes, à l'image des bords de bâtiments. Le MNS étant moins précis sur ces zones, il n'est donc pas gênant de le sous-échantillonner à 50 cm, la perte d'information étant très faible, et la rupture entre toit et sol étant toujours très bien visible.

Les figures IV.9 et IV.10 permettent de mesurer l'évolution entre deux détections à nomenclatures différentes, la première correspondant à la classification du bâti uniquement tandis que la seconde figure s'intéresse également aux classes *ligneux*, *eau*, *pont*. Sur cette dernière, la classe *autre* n'est pas représentée, pour ne pas polluer visuellement le résultat. Du point de vue de la classe *bâti*, on voit que l'intérêt principal est de contraindre le classifieur à détecter les ponts, évitant des sur-détections du bâti par rapport à ce type d'objet.

(A) orthophotographie sur la Vendée.

(B) Classification en utilisant le MNS.

FIGURE IV.9. Détection du **bâti uniquement** sur la Vendée avec le réseau U-Net : le pont est classé à tord comme un objet de la classe *bâti* par le réseau « binaire ».

Réseau	Nb patchs	Nb paramètres	Durée entraînement
U-Net	18000	400k	< 1 jour
MobileNetV2	18000	2M	< 2 jours
DeepLabV3+	36000	20M	< 1 semaine

TABLE IV.2. Configurations des différents réseaux testés dans la seconde expérimentation du cas multiclasses (images à 20 cm de résolution).

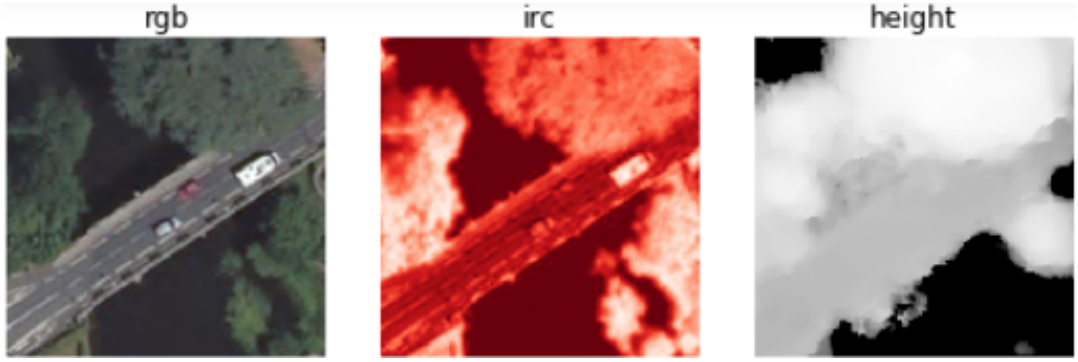
L'enrichissement du nombre de classes est en outre intrinsèquement un effet bénéfique. La détection de la végétation est intéressante du point de vue environnemental et constitue un enjeu pour l'ensemble des communes sur le territoire qui cherchent à recenser la végétation en milieu urbain. La figure IV.10 montre que cela est possible pour les surfaces de végétations relativement étendues mais, pour les arbres isolés, cela reste un défi à relever dans le futur. La sous-détection de ces arbres isolés est due aux données d'apprentissage issues du RGFor qui ne recense que les surfaces de végétation suffisamment grande (plusieurs hectares au minimum). Le réseau n'a donc pas de référence pour détecter correctement les arbres isolés, ceux-ci ayant une géométrie bien plus étroite et fine spatialement qu'une forêt.

Pour les expérimentations avec les trois réseaux, U-Net, MobileNetV2 et DeepLabV3+, nous utilisons des tailles de jeu de données différentes, les réseaux étant eux-mêmes dimensionnés différemment. Cette fois-ci, les images des différents jeu d'apprentissage sont à la résolution de 20 cm. Les trois configurations sont visibles sur le tableau IV.2, par ordre croissante de complexité des réseaux.

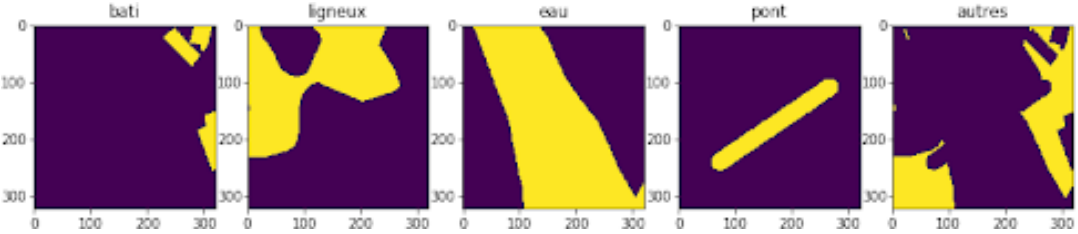
(A) Vérité terrain.

(B) Classification en utilisant le MNS : l'ambiguïté sur le pont est levée et on enrichit sémantiquement la scène par rapport à une classification monoclasse.

FIGURE IV.10. Classification **multiclasses** sur la Vendée avec le réseau U-Net :
● *Zones bâties*, ● *Ligneux*, ● *Eau*, ● *Pont* .



(A) Canaux utilisés : rouge, vert, bleu, infrarouge en spectral et un canal (MNS-MNT) pour l'information de hauteur.



(B) Masques correspondant aux diverses classes d'intérêt sur les bases de données géographiques utilisées.

FIGURE IV.11. Patch d'apprentissage construit sur la BD Ortho pour la détection multiclasses. On observe un problème d'actualité et de géométrie des bases de données sur les classes de *ligneux* et de *pont* en particulier.

Un exemple de correspondance entre patch d'apprentissage et vérité terrain est illustré en figure IV.11. La première partie indique la nature des canaux utilisés, à savoir l'empilement des canaux rouge, vert et infrarouge pour l'information radiométrique, ainsi que la hauteur. La seconde partie renvoie aux masques générés pour chacune des classes qui nous intéresse, à partir des bases de données topographiques. Ce qu'illustre cette figure est le décalage réel qu'il peut y avoir entre la réalité physique affichée par l'image et l'information contenue dans les bases de données que l'on considère comme la vérité. La classe *ligneux* n'est pas à jour par exemple, avec la végétation au sud-est absente du masque et donc du RGFor. De même, le pont issu de la BD Topo est géométriquement incorrect par rapport à la réalité, avec une largeur inférieure à ce qu'elle devrait être.

Le tableau IV.3 recense les résultats pour les trois modèles entraînés. Les métriques exposées sont l'« intersection over union (IoU) » offrant une performance géométrique globale pour chaque réseau, ainsi que les deux indicateurs complémentaires de Précision et Rappel. On y expose ces indicateurs calculés pour la classe *autre*, mais à titre indicatif, celle-ci englobant la partie du territoire complémentaire de l'ensemble des quatre autres classes d'intérêt.
Par comparaison entre modèle, le premier constat est l'accroissement des métriques en passant du modèle U-Net vers des modèles comportant davantage

Classes		*bâti*	*ligneux*	*eau*	*pont*	*autres*
U-Net	IoU	43,76	40,63	49,91	28,44	67,88
	Rappel	55,95	63,45	62,99	49,53	83,99
	Précision	73,70	57,44	74,46	51,33	77,15
MobileNetV2	IoU	**62,33**	**43,98**	55,97	**44,44**	**71,95**
	Rappel	**73,33**	**59,15**	**71,85**	**61,32**	84,44
	Précision	**82,34**	**68,07**	74,32	67,85	**82,10**
DeepLabV3+	IoU	59,25	42,12	**57,66**	44,05	70,84
	Rappel	72,59	58,74	67,06	60,71	**85,96**
	Précision	78,78	64,36	**83,16**	**68,99**	78,75

TABLE IV.3. Performances des différents réseaux entraînés sur les départements 85 (Vendée) et 31 (Haute-Garonne) avec des images à 20 cm de résolution.

de paramètres, accroissement non négligeables puisque si l'on regarde le gain moyen en IoU entre U-Net et MobileNetV2, celui-ci grimpe de 20.8%. Les raisons de cette amélioration notable réside dans la complexité du derniers réseaux qui permet de caractériser un grand nombre de spécificités au sein de chaque classe. Des filtres peuvent être dédiés à la caractérisation de chaque variété possible d'objets pour une classe donnée, tandis qu'un modèle avec moins de paramètres ne permet que de dessiner un comportement global à adopter pour une classe, ne pouvant attribuer des paramètres à la caractérisation précise d'une occurrence d'objet particulière de cette classe. Toutefois, augmenter drastiquement le nombre de paramètres n'apportent pas d'amélioration, avec au contraire, une détérioration de la quasi majorité des classes : DeepLabV3+ est ainsi meilleur uniquement sur la classe *eau*. Si cette étude est très préliminaire, on peut émettre l'explication suivante : le nombre de paramètres étant très élevé pour ce dernier réseau, on ne fait en revanche « que » doubler le nombre de patchs d'apprentissage. Si nous n'avons pas constaté de sur-apprentissage *a priori* sur le jeu de validation lors de l'entraînement, celui-ci n'était pas suffisamment représentatif. En outre, d'un point de vue algorithmique, entraîner un tel réseau n'est pas évident, les degrés de liberté étant nombreux, il faut souvent procéder itérativement (fixer les paramètres des dernières couches par exemple pour n'entraîner que les premières, puis relâcher progressivement l'ensemble des paramètres) pour obtenir un réseau adéquat *in fine*.

Du point de vue de la cartographie d'occupation des sols, on obtient systématiquement de meilleures valeurs de précision que de valeurs de rappel : les objets des différentes classes détectées sont donc relativement bien classés, mais avec des erreurs d'omission importantes. On rappelle encore une fois que le calcul de ces métriques est complètement dépendant de la qualité de chaque base de données, dont la qualité a été discutée plusieurs fois déjà au cours de ce manuscrit. La classe de *ligneux* et de semble particulièrement affectée, avec le RGFor dont l'actualité est discutable. La classe *pont* est difficile à détecter pour des raisons de représentativité, le nombre de ponts dans les jeux d'apprentissage étant plus faibles que pour le reste des classes.

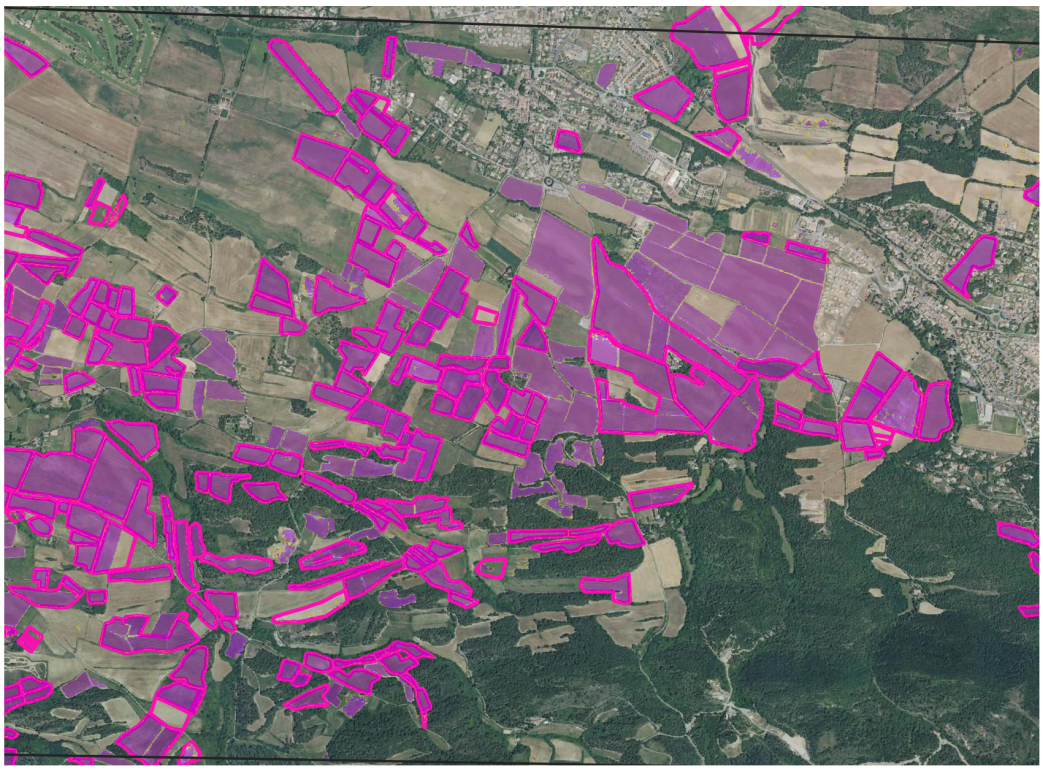
4.3 Détection de vignes

Enfin, nous exposons ici la détection de *vignes* avec des résultats préliminaires uniquement qualitatifs. En effet, si l'on dispose de parcelles de vignes grâce au RPG, celui-ci ne couvre pas l'ensemble des vignes du territoire, ceci pour des raisons économiques : chaque agriculteur annote les parcelles agricoles qu'il détient sur le portail de l'AFP, mais seulement si celles-ci conduisent à une rémunération au titre de la PAC. Or, les vignes ne sont pas source d'indemnités dans le cadre de la PAC, leur recensement est donc incomplet.

Les patchs d'apprentissage sont construits selon la méthodologie employée pour la détection *bâti / non bâti*, à 20 cm, pour le plus d'information géométrique et de texture possibles. La classe *non vigne* est enrichie de classes agricoles telles que les légumineuses et les céréales, en conservant les classes de bâti, ligneux, eau, etc. Les départements 11, 31 et 85 sont tous les trois utilisés pour le jeu d'apprentissage, le premier présentant le plus de terrains viticoles (8,8% du département environ, les deux autres étant inférieurs à 0,3%). Le nombre de patchs est de 1500 pour les départements 31 et 85, de 4000 pour le département 11. A l'inverse des études précédentes de ce chapitre, le MNS n'est pas utilisé, étant de qualité médiocre sur les vignes (bruit ou valeur nulle du MNS par rapport au MNT). Le modèle entraîné est U-Net.

La détection est ensuite menée sur des zones tests sur les trois départements ainsi qu'un quatrième, le département 49 (Maine-et-Loire) pour y avoir une analyse qualité du modèle sur une région sur laquelle il n'a pas de connaissance *a priori*.
Correspondant aux quatre départements en question, les figures IV.12, IV.13, IV.14 et IV.15 proposent une vision globale de chaque zone test relative à un département ainsi qu'un visuel à une échelle plus petite. Les vignes détectées sont colorées avec une transparence permettant de voir l'image aérienne en arrière-plan. Les polygones présents sur chaque image correspondent à l'emprise d'une parcelle présente dans le RPG. On constate sur les quatre zones un écart très important entre ce que l'on détecte et les données du RPG, pour la raison liée à la saisie par les agriculteurs expliquée précédemment.
Les résultats sont qualitativement très intéressants. L'ensemble des parcelles saisies dans le RPG semble être justement détecté par le modèle. Les difficultés les plus importantes se situent sur certaines parcelles du département de la Vendée, pour laquelle certaines d'entre elles ne présentent aucune structure d'alignement telles qu'on peut les observer sur la plupart des vignes, rendant la détection impossible (parcelle nord-est de la figure IV.13b) ou incertaine, avec un phénomène de mitage. En revanche, le modèle caractérise très bien le reste des vignes sur la Vendée, l'Aude et la Haute-Garonne, qui viennent compléter de manière très satisfaisante les données du RPG, sémantiquement et géométriquement. L'absence de l'information de hauteur ne nuit aucunement à la distinction entre forêt et vigne comme on peut le voir sur la figure IV.14b où une parcelle est très bien reconnue au milieu de la forêt. Enfin, sur le département du Maine-et-Loire, non échantillonné pour

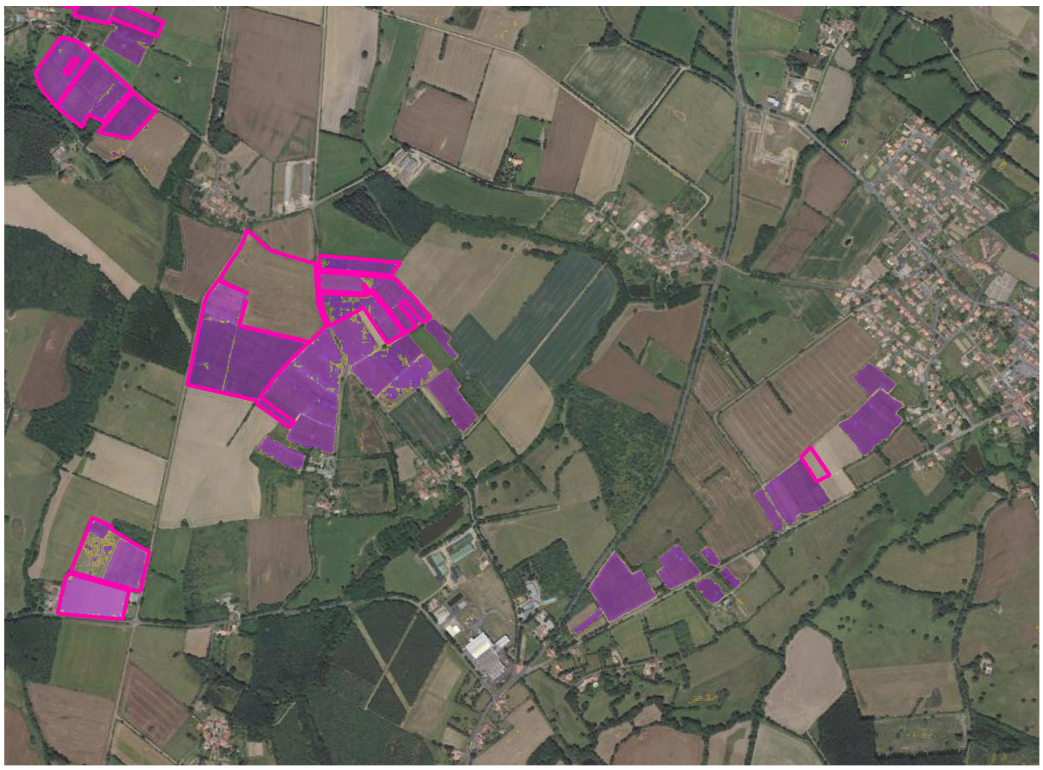
constituer la base d'apprentissage, le modèle généralise très bien ce que sont les vignes, là encore complétant le RPG avec un peu de mitage sur certaines parcelles.



(A) Détection globale.

(B) Zoom sur des parcelles de vignes non recensées dans le RPG, mais correctement détectées.

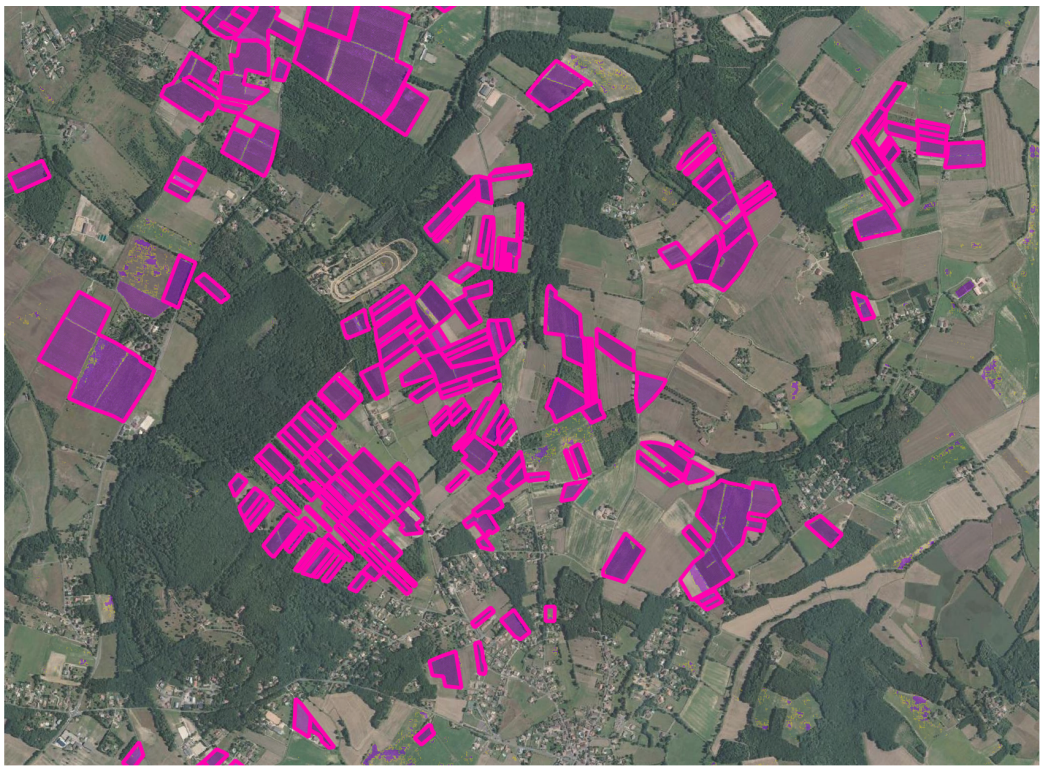
FIGURE IV.12. Détection des vignes sur le département de l'Aude (11).



(A) Détection globale.

(B) Zoom sur des parcelles de vignes non détectées à raison ?

FIGURE IV.13. Détection des vignes sur le département de la Vendée (85).



(A) Détection globale.

(B) Zoom sur des parcelles difficiles à reconnaître (pas de structure d'alignement propres aux vignes).

FIGURE IV.14. Détection des vignes sur le département de Haute-Garonne (31).

(A) Détection globale.

(B) Zoom sur un grand nombres de parcelles non recensées par le RPG mais correctement détectées.

FIGURE IV.15. Détection des vignes sur le département du Maine-et-Loire (49).

Chapitre V

Conclusion et perspectives

1 Conclusion

Lorsque l'on aborde la problématique de classification automatique d'images satellites ou aériennes pour cartographier l'occupation des sols, deux volets d'étude sont possibles et complémentaires. Tout d'abord, du point de vue de la vision par ordinateur, ou *computer vision*, beaucoup de verrous méthodologiques existent en terme (i) d'ingénierie (différentes bibliothèques existent, beaucoup d'hyperparamètres à contrôler lors de l'apprentissge) et (ii) de recherche de méthodes de classification / segmentation toujours plus performantes. L'apprentissage profond est ce premier volet algorithmique. Un historique et un état de l'art des réseaux de neurones ont été proposés dans le chapitre II de cette thèse, et qui montrent qu'aujourd'hui, les recherches liées à ce domaine sont multiples et l'un d'entre eux est la télédétection. Avec l'appui des entreprises du secteur tertiaire (on pense aux GAFA) qui investissent énormément dans ces technologies, celles-ci sont maintenant incontournables, et leur accès est facilité par l'existence de plusieurs librairies également pensées pour des objectifs à long terme de production et d'industrialisation. Si l'infrastructure physique peut être bloquante, les calculs sur réseaux nécessitant des cartes graphiques, aussi bien à leur entraînement qu'à leur utilisation au moment de la prédiction, des solutions de « cloud computing » sont disponibles, même si cela soulève des questions en terme de gouvernance, et de sécurité des données puisqu'il est nécessaire de charger les données sur le « cloud » pour profiter pleinement de la puissance de celui-ci.

La seconde étude relève de la thématique liée à l'usage de la télédétection, et des enjeux et contraintes liés à ce champ de recherche, qui reste, à l'heure actuelle, cantonné au domaine académique si l'on parle de classification automatique d'objets sur des images vues du ciel ou de l'espace. Pourtant, nous l'avons vue en chapitre introductif, la connaissance de l'occupation des sols est cruciale, à plusieurs niveaux. Avec l'impulsion d'établissements publics décisionnaires en matière d'écologie, cette cartographie est au centre des préoccupations. Du point de vue de la télédétection, avec une multitude de capteurs disponibles, nous avons fait le pari de nous orienter en priorité vers les capteurs embarqués sur SPOT 6 et 7, offrant un compromis entre couverture globale et résolution spatiale. Leur utilisation autorise une granularité géométrique plus fine que des capteurs spatiaux tels que Sentinel du programme européen Copernicus, avec une description du territoire sur des classes génériques très satisfaisante. Sur des problématiques de mise à jour de bases de données topographiques, la télédétection spatiale trouve ses limites pour des raisons de spécifications existantes sur ces bases déjà existantes (précision métrique / submétrique de l'information), d'où l'utilisation par la suite d'orthophotographies aériennes et de MNS, conséquence d'une demande directe de Ministères sur l'état de l'artificialisation des sols, en s'orientant vers la nomenclature de l'OCS GE.

L'objet de cette thèse était donc de mettre un type d'algorithme particulier, les réseaux de neurones profonds, au service d'une thématique spécifique, la cartographie d'occupation des sols, en gardant à l'esprit la nécessité de ne pas rester dans un cadre purement académique pour s'inscrire dans un projet plus large à visée opérationnelle. Cela se traduit par l'utilisation de bases de données propres à notre problématique et imparfaites, à l'inverse des jeux de données courants en télédétection servant à confronter deux méthodes entre elles (mais qui demeurent nécessaires pour les améliorations méthodologiques).
L'évaluation de l'apport de méthodes d'apprentissage profond sur des cas d'étude réaliste de cartographie d'occupation des sols a été double, avec la volonté d'une part de valoriser les images satellites monoscopiques SPOT 6 et 7, à très haute résolution spatiale, en montrant leur utilité même dans des situations difficiles pour ce type d'image, et d'autre part d'utiliser ces mêmes méthodes pour la classification d'images aériennes, en réponse à une commande réelle sujette à des spécifications existantes.

Nous avons proposé en chapitre III une méthode de classification par réseau de neurones sur images satellites monoscopiques à très haute résolution spatiale, sans données annexes, en apportant une stratégie de classification sur une zone homogène thématiquement dans un premier temps, à savoir le Finistère. Les résultats ont été probants, avec une cohérence sémantique du paysage breton décrit par les cartes produites en milieux urbain et rural. Les données images étant composées de quatre bandes, le modèle a dû être entraîné initialement de zéro. Confrontés à une couverture nationale

monodate, les problèmes liés aux variations saisonnières et différences d'apparence d'objets appartenant à une même classe d'une région à une autre sont à prendre en compte. Pour cela, à une échelle locale, plusieurs tests de ré-entraînement de ce réseau ont été conduits, chacun renvoyant à un scenario parmi plusieurs. Ainsi, l'effet d'un changement géographique, temporel ou sémantique est réel sur les performances du réseau initial, avec des dégradations plus ou moins importantes, la principale dégradation survenant lorsque la zone cartographiée diffère grandement de la zone d'apprentissage. Malgré tout, au prix de temps de calcul et de volume d'apprentissage fortement réduits, le *fine-tuning* a prouvé son efficacité dans chacun de ces cas. Plus encore, lorsqu'il a fallu passer à l'échelle sur une zone étendue, à savoir le département de la Gironde, le *fine-tuning* a permis de lever un nombre considérable d'ambiguïtés malgré un changement géographique conséquent ainsi qu'une date d'acquisition éloignée en absolu de celle de l'image ayant servi à l'entraînement initial.
Enfin, une approche fondée sur les superpixels calculés par sur-segmentation de l'image a mené à une étape de prédiction bien plus rapide, sans dégrader le résultat.
L'enseignement principal de ce chapitre, a été le fait que les résultats obtenus étaient meilleurs que ceux que l'on pouvait produire par des méthodes non profondes, motivant des efforts supplémentaires dans cette direction, notamment au travers du chapitre suivant lorsque l'on observe peu d'ambiguïté entre sol et sursol sur des images satellites monoscopiques, mais sans MNS, de la part du réseau de neurones profond.

Le chapitre IV a permis d'orienter la problématique vers une direction plus opérationnelle encore, avec un enjeu réel, celui de cartographier le territoire à une fréquence triennale, et un socle existant, l'OCS GE. Celle-ci couvrant de nombreux thèmes à une granularité plus ou moins fine, une série de tests sur divers postes de la nomenclature a été menée. Les classes choisies peuvent l'avoir été pour des raisons thématiques, c'est le cas du bâti pour la quantification de l'artificialisation des sols, ou pour apporter des réponses à certains défis, avec l'exemple des vignes pour lesquelles la texture est déterminante dans leur détection : les réseaux apprenant des attributs très performants et discrimants liés aux textures, ce type de méthode donne un résultat quasi-parfait et opérationnel pour cette classe. Pour obtenir une résolution de la cartographie finale compatible avec les spécifications OCS GE, les images aériennes sont utilisées. De plus, le calcul des orthophotographies permet d'obtenir également un MNS, qui a montré son utilité à plus d'un titre.
En s'appuyant sur une architecture U-Net légère par rapport au reste de l'état de l'art, et des images aériennes, la tâche de détection du bâti a été très satisfaisante, à l'inverse des méthodes classiques qui présentent une forte sensibilité au bruit et à la qualité du MNS et du MNT (l'information retenue étant la hauteur, la soustraction du second au premier). L'information de hauteur s'est révélée capitale, levant des confusions bâti / parkings ou bâti

/ route. Lors du passage sur une classification multiclasses, en plus de l'effet intrinsèque d'avoir un nombre accru de classes, et donc de permettre au réseau d'envisager de nouveaux cas de figure auparavant non connus, les contraintes entre classes établies à l'apprentissage peuvent permettre d'aller vers un meilleur modèle et de diminuer les fausses détections pour les classes existantes. Cela s'est vérifié avec les confusions entre classes *bâti* et *pont*, mais aussi dans le chapitre III entre les classes de *haie* et de *route*. La détection de vignes est uniquement visuelle, une analyse quantitative avec une base de données de référence incomplète n'ayant pas beaucoup de sens. En revanche, les cartes obtenues sur la thématique de la vigne sont cohérentes avec les observations et offre une complétion satisfaisante des parcelles disponibles dans le RPG.
Le constat récurrent pendant l'ensemble de ces tests et le travail itératif de constitution d'une base de connaissance pour entraîner un modèle, les faux positifs servant à inclure de nouvelles classes dans cette base. Toutefois, l'analyse n'étant que monodate, des objets non pérennes n'ont pas pu être filtré, à l'image des monticules de débris affichant un MNS non nul et une radiométrie proche de celles des bâtiments, conduisant à des confusions entre ces classes.

2 Perspectives

2.1 Axes d'amélioration

Les vecteurs d'améliorations des travaux compilés dans ce manuscrit peuvent relever du domaine de la télédétection ou de l'apprentissage automatique. Nous les présentons successivement.

Perspectives en télédétection

Rapidement mentionnée dans le paragraphe précédent, l'exploitation de données monodate trouve ses limitations à divers endroits, en particulier pour les objets non pérennes dont les comportements géométrique et radiométrique sont proches d'objets d'intérêt. Pour filtrer ces artefacts, l'utilisation de séries temporelles est difficile, leurs tailles pouvant être inférieures en surface aux résolutions disponibles pour les capteurs de données multi-temporelles. En revanche, il est possible de constituer des bases d'apprentissage pour ces objets difficiles, ce qui viendrait à un prix conséquent, mais une fois la base de connaissance pour ces objets construites, elle est ré-utilisables dans l'ensemble des cas d'étude. De plus, des méthodes d'*active learning* permettent de distiller progressivement une base de connaissance réduite mais avec des performances intéressantes (Sener et Savarese, 2017).
Malgré tout, la multiplicité des capteurs multi-temporels, dont certains sont décrits en introduction, aurait un impact positif sur certaines classes que l'information monodate ne suffit pas à caractériser. En particulier, sur les classes naturelles étendues, forêts et cultures, il serait possible d'affiner la nomenclature que l'on a proposée dans le chapitre III en intégrant les cartes d'OCS

établies par Inglada et al. (2017) par exemple, dont un millésime annuel est calculé par la chaîne *iota2* à partie de séries temporelles Sentinel-2. De plus, nous avons vu que la détection d'une classe générique *culture* est possible avec des images SPOT 6 et 7, à partir de laquelle il serait donc intéressant de réaliser un masque pour utiliser la chaîne *iota2*, afin d'affiner sémantiquement ce masque.
D'un point de vue global, nous l'avons également constaté en chapitre IV, partir sur une approche hiérarchique de la détection des classes est une pratique intéressante, en s'intéressant dans un premier temps à des classes génériques puis en affinant la granularité sémantique de chacune de ces classes.

Perspectives du point de vue de l'apprentissage profond

Par rapport au chapitre III, les temps de prédiction étant longs, un moyen de réduire ces temps serait de ne prédire qu'un pixel sur deux ou trois, et interpoler entre chaque position prédite la classe des pixels intermédiaires. Cela se justifie puisque les régions détectées sont peu bruitées en leur sein, le processus d'interpolation ne produirait alors pas beaucoup d'artefact.
Concernant l'approche superpixel, si celle-ci est intéressante déjà en l'état, elle ne sert qu'à réduire des temps de calcul, sans remise en question de la manière d'aborder le sujet. Or, un moyen intéressant d'exploiter pleinement ces superpixels seraient de rectifier les superpixels de manière à avoir des carrées ou rectangles de dimensions fixes, et d'entraîner un réseau de neurones sur ces superpixels rectifiés. Il faudrait tout de même s'assurer d'une certaine compacité des segments issus de la sur-segmentation, pour que la rectification de ceux-ci demeure cohérente spatialement. Les patchs d'apprentissage devrait alors subir le même type de traitement pour obtenir un réseau adapté à cette tâche. Procéder par convolution directe sur les superpixels aurait l'intérêt de modéliser directement les relations topologiques, parfois complexes, entre objets segmentés.
Même si cette tâche mêle les deux points de vue, télédétection, et apprentissage profond, l'investigation plus poussée de l'intérêt de la constitution d'un jeu d'apprentissage plus transverse est un aspect que l'on a à peine traité dans le chapitre IV. On suggère ici d'entraîner par exemple un modèle sur des données issues du jeu de données de la zone Finistère mais aussi de la zone Gironde, puis prédire sur ces deux régions afin de voir si on peut encore limiter le nombre de modèles qui seraient nécessaires pour classifier le territoire national. L'échantillonnage sur plusieurs départements a été effectués en chapitre IV mais sur des nomenclature à classe unique.
Par ailleurs, si le *fine-tuning* a pu être abondamment étudié et expérimenté, avec des issues favorables, certaines classes ne possèdent que très peu d'annotation pour des raisons d'absence de bases de données, ou parce que les classes elles-mêmes sont très rares. Envisager des processus d'*active learning* pour apprendre de telles classes sur la base d'un réseau pré-entraîné est champ d'investigation dont la portée est intéressante d'un point de vue recherche, mais aussi pour des visées applicatives réelles, puisque certaines

classes d'occupations répondent justement à l'un de ces deux critères (carrière par exemple). De même, aborder ces problèmes par l'utilisation de réseaux types *few-shot* (Snell et al., 2017 ; Gidaris et Komodakis, 2018) aurait un intérêt particulier, ces architectures ayant justement vocation à traiter de nouvelles classes par transfert d'apprentissage de classes existantes, pour lesquelles on possède beaucoup d'annotations, vers des classes rares. Ce transfert s'opère généralement par critère de similarité : le réseau est d'abord entraîné sur les classes suffisamment annotées, puis, par similarité (cosinus) entre ces classes et les nouvelles classes, les poids du réseau sont modifiés dynamiquement. Des modèles type *siamese network* (Koch et al., 2015) répondent également à ce type de problème en affectant deux *streams* parallèles au sein du réseau, qui sont des branches identiques constituées de couches convolutives, et dont les paramètres sont les mêmes pour les deux branches. Ces deux *streams* se comportent comme des extracteurs d'attributs et prennent en entrée chacun une image d'une paire d'images d'un jeu de données d'apprentissage. Le but de ce réseau est de définir l'appartenance ou non de ces deux images à une même classe, forçant notamment des classe ambiguës à être séparées.

Du point de vue de la qualité cartographique des résultats du chapitre IV, ceux-ci mériterait un travail d'amélioration visuelle afin de satisfaire au mieux des spécifications de rendus cartographique, avec des polygones vectorisés plus rectilignes, et présentant moins d'arêtes. L'utilisation de GANs est une réponse plausible pour cette tâche, pour modéliser la fonction de transfert entre le rendu brut issu des CNNs utilisés vers un rendu cartographique affiné. Cela nécessite la constitution d'un jeu d'apprentissage avec par exemple, pour le bâti, l'empreinte de la BD TOPO couche bâtie comme vérité terrain, et le rendu correspondant prédit par le réseau.

Bibliographie

[1] Radhakrishna Achanta et al. “SLIC Superpixels Compared to State-of-the-Art Superpixel Methods”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), p. 2274-2282. ISSN : 0162-8828 (cf. p. 98).

[2] D. H. Ackley, G. E. Hinton et T. J. Sejnowski. “A learning algorithm for Boltzmann machines”. In : *Cognitive science* 9.1 (1985), p. 147-169 (cf. p. 39).

[3] Naomi S Altman. “An introduction to kernel and nearest-neighbor nonparametric regression”. In : *The American Statistician* 46.3 (1992), p. 175-185 (cf. p. 26).

[4] Antreas Antoniou, Amos Storkey et Harrison Edwards. “Data augmentation generative adversarial networks”. In : *arXiv preprint arXiv :1711.04340* (2017) (cf. p. 50).

[5] N. Audebert, B. Le Saux et S. Lefèvre. “Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks”. In : *Asian Conferenceon Computer Vision, 20-24 November, Taipei, Taiwan*. 2016 (cf. p. 57).

[6] Nicolas Audebert, Bertrand Le Saux et Sébastien Lefèvre. “Beyond RGB : Very high resolution urban remote sensing with multimodal deep networks”. In : *ISPRS Journal of Photogrammetry and Remote Sensing* 140 (2018), p. 20 -32 (cf. p. 58).

[7] Vijay Badrinarayanan, Alex Kendall et Roberto Cipolla. “Segnet : A deep convolutional encoder-decoder architecture for image segmentation”. In : *arXiv :1511.00561* (2015) (cf. p. 57).

[8] G Ball et Isodata Hall Dj. *A novel method of data analysis and pattern classification. Isodata, A novel method of data analysis and pattern classification. Tch. Report 5RI, Project 5533*. 1965 (cf. p. 23).

[9] A Bannari et al. “Effets de la couleur et de la brillance du sol sur les indices de végétation”. In : *International Journal of Remote Sensing* 17.10 (1996), p. 1885-1906 (cf. p. 31).

[10] Yoshua Bengio. “A connectionist approach to speech recognition”. In : *Advances in Pattern Recognition Systems Using Neural Network Technologies*. World Scientific, 1993, p. 3-23 (cf. p. 38).

[11] Yoshua Bengio et Paolo Frasconi. “Credit assignment through time : Alternatives to backpropagation”. In : *Advances in Neural Information Processing Systems*. 1994, p. 75-82 (cf. p. 38).

[12] Yoshua Bengio, Patrice Simard, Paolo Frasconi et al. “Learning long-term dependencies with gradient descent is difficult”. In : *IEEE transactions on neural networks* 5.2 (1994), p. 157-166 (cf. p. 38).

[13] A. Blum et T. Mitchell. “Combining labeled and unlabeled data with co-training”. In : *Proceedings of the eleventh annual conference on Computational learning theory*. ACM. 1998, p. 92-100 (cf. p. 22).

[14] Léon Bottou et Olivier Bousquet. “The tradeoffs of large scale learning”. In : *Advances in neural information processing systems*. 2008, p. 161-168 (cf. p. 48).

[15] Léon Bottou et Yann L Cun. “Large scale online learning”. In : *Advances in neural information processing systems*. 2004, p. 217-224 (cf. p. 48).

[16] Y-Lan Boureau, Jean Ponce et Yann LeCun. “A theoretical analysis of feature pooling in visual recognition”. In : *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, p. 111-118 (cf. p. 54).

[17] Leo Breiman. “Bagging predictors”. In : *Machine learning* 24.2 (1996), p. 123-140 (cf. p. 29).

[18] Leo Breiman. “Random forests”. In : *Machine learning* 45.1 (2001), p. 5-32 (cf. p. 26).

[19] Xavier Briottet et al. “Application de l’optique aux milieux urbains”. In : *Observation des surfaces continentales par télédétection III - Urbain et zones côtières*. 2017. URL : https://hal.archives-ouvertes.fr/hal-02118996 (cf. p. 105).

[20] Greg Brockman et al. *OpenAI Gym*. 2016. eprint : arXiv:1606.01540 (cf. p. 40).

[21] L. Bruzzone, M. Chi et M. Marconcini. “A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images”. In : *IEEE Transactions on Geoscience and Remote Sensing* 44.11 (2006), p. 3363-3373 (cf. p. 22).

[22] L. Bruzzone, M. Chi et M. Marconcini. “Advanced Semisupervised SVM Approaches to Classification of Hyperspectral Data”. In : *2006 IEEE International Symposium on Geoscience and Remote Sensing*. 2006, p. 3887-3890 (cf. p. 22).

[23] Gustavo Camps-Valls et al. “Robust support vector method for hyperspectral data classification and knowledge discovery”. In : *IEEE Transactions on Geoscience and Remote sensing* 42.7 (2004), p. 1530-1542 (cf. p. 26).

[24] Nesrine Chehata, Li Guo et Clément Mallet. “Airborne lidar feature selection for urban classification using random forests”. In : *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 38.Part 3 (2009), W8 (cf. p. 30).

[25] Jun Chen et al. “Global land cover mapping at 30m resolution : A POK-based operational approach”. In : *ISPRS Journal of Photogrammetry and Remote Sensing* 103 (2015), p. 7 -27 (cf. p. 17).

[26] K. Chen et al. “Residual shuffling convolutional neural networks for deep semantic image segmentation using multi-modal data”. In : *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2 (2018), p. 65-72 (cf. p. 58).

[27] K. Chen et al. “Semantic Segmentation of Aerial Images With Shuffling Convolutional Neural Networks”. In : *IEEE Geoscience and Remote Sensing Letters* 15.2 (2018), p. 173-177 (cf. p. 58).

[28] Liang-Chieh Chen et al. “Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In : *IEEE TPAMI* 40.4 (2017), p. 834-848 (cf. p. 108).

[29] Liang-Chieh Chen et al. “Deeplab : Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs”. In : *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2018), p. 834-848 (cf. p. 58).

[30] Liang-Chieh Chen et al. “Encoder-decoder with atrous separable convolution for semantic image segmentation”. In : *Proceedings of the European conference on computer vision*. 2018, p. 801-818 (cf. p. 108).

[31] Liang-Chieh Chen et al. “Rethinking atrous convolution for semantic image segmentation”. In : *arXiv preprint arXiv :1706.05587* (2017) (cf. p. 108).

[32] Liang-Chieh Chen et al. “Semantic image segmentation with deep convolutional nets and fully connected CRFs”. In : *arXiv :1412.7062* (2014) (cf. p. 57).

[33] Liang-Chieh Chen et al. “Semantic image segmentation with deep convolutional nets and fully connected crfs”. In : *arXiv preprint arXiv :1412.7062* (2014) (cf. p. 108).

[34] Kyunghyun Cho et al. “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In : *arXiv preprint arXiv :1406.1078* (2014) (cf. p. 38).

[35] François Chollet. “Xception : Deep learning with depthwise separable convolutions”. In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 1251-1258 (cf. p. 108).

[36] Dan Claudiu Ciresan et al. “Deep big simple neural nets excel on handwritten digit recognition”. In : *arXiv preprint arXiv :1003.0358* (2010) (cf. p. 41).

[37] Corinna Cortes et Vladimir Vapnik. “Support-vector networks”. In : *Machine learning* 20.3 (1995), p. 273-297 (cf. p. 26).

[38] Cour des Comptes. *La politique d’aide aux biocarburants*. 2012, p. 118 (cf. p. 2).

[39] Nello Cristianini, John Shawe-Taylor et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000 (cf. p. 28).

[40] Ekin D Cubuk et al. “AutoAugment : Learning Augmentation Strategies From Data”. In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, p. 113-123 (cf. p. 50).

[41] C. Dechesne et al. “Semantic segmentation of forest stands of pure species combining airborne lidar data and very high resolution multispectral imagery”. In : *ISPRS Journal of Photogrammetry and Remote Sensing* 126 (2017), p. 129 -145 (cf. p. 58).

[42] J. Deng et al. “ImageNet : A Large-Scale Hierarchical Image Database”. In : *CVPR09*. 2009 (cf. p. 55).

[43] Li Deng. “The MNIST database of handwritten digit images for machine learning research”. In : *IEEE Signal Processing Magazine* 29.6 (2012), p. 141-142 (cf. p. 55).

[44] C. Donlon et al. “The Global Monitoring for Environment and Security (GMES) Sentinel-3 mission”. In : *Remote Sensing of Environment* 120 (2012). The Sentinel Missions - New Opportunities for Science, p. 37 -57. ISSN : 0034-4257. DOI : https://doi.org/10.1016/j.rse.2011.07.024. URL : http://www.sciencedirect.com/science/article/pii/S0034425712000685 (cf. p. 14).

[45] M. Drusch et al. “Sentinel-2 : ESA’s Optical High-Resolution Mission for GMES Operational Services”. In : *Remote Sensing of Environment* 120 (2012). The Sentinel Missions - New Opportunities for Science, p. 25 -36. ISSN : 0034-4257. DOI : https://doi.org/10.1016/j.rse.2011.11.026. URL : http://www.sciencedirect.com/science/article/pii/S0034425712000636 (cf. p. 13).

[46] H. D. Eva et al. “A land cover map of South America”. In : *Global Change Biology* 10.5 (2004), p. 731-744 (cf. p. 22).

[47] Mark Everingham et al. “The pascal visual object classes (voc) challenge”. In : *International journal of computer vision* 88.2 (2010), p. 303-338 (cf. p. 55).

[48] P. F. Felzenszwalb et D. P. Huttenlocher. “Efficient Graph-Based Image Segmentation”. In : *International Journal of Computer Vision* 59.2 (2004), p. 167-181. ISSN : 0920-5691 (cf. p. 99).

[49] G. M. Foody et A. Mathur. “A relative evaluation of multiclass image classification by support vector machines”. In : *IEEE Transactions on Geoscience and Remote Sensing* 42.6 (2004), p. 1335-1343 (cf. p. 25).

[50] Giles M Foody. “Assessing the accuracy of land cover change with imperfect ground reference data”. In : *Remote Sensing of Environment* 114.10 (2010), p. 2271-2285 (cf. p. 59).

[51] Giles M Foody. “Status of land cover classification accuracy assessment”. In : *Remote sensing of environment* 80.1 (2002), p. 185-201 (cf. p. 26).

[52] Giles M Foody et Ajay Mathur. “A relative evaluation of multiclass image classification by support vector machines”. In : *IEEE Transactions on geoscience and remote sensing* 42.6 (2004), p. 1335-1343 (cf. p. 25).

[53] K.o Fukushima. "Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In : *Biological cybernetics* 36.4 (1980), p. 193-202 (cf. p. 35).

[54] Spyros Gidaris et Nikos Komodakis. "Dynamic few-shot visual learning without forgetting". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, p. 4367-4375 (cf. p. 140).

[55] Xavier Glorot, Antoine Bordes et Yoshua Bengio. "Deep sparse rectifier neural networks". In : *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011, p. 315-323 (cf. p. 43).

[56] P. Gong et al. "Finer resolution observation and monitoring of global land cover : first mapping results with Landsat TM and ETM+ data". In : *International Journal of Remote Sensing* 34.7 (2013), p. 2607-2654 (cf. p. 17).

[57] Yunchao Gong et al. "Multi-scale orderless pooling of deep convolutional activation features". In : *European conference on computer vision*. Springer. 2014, p. 392-407 (cf. p. 54).

[58] Ian Goodfellow et al. "Generative adversarial nets". In : *Advances in neural information processing systems*. 2014, p. 2672-2680 (cf. p. 50).

[59] A. Gressin et al. "Updating land cover databases using a single very high resolution satellite image". In : *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W2 (2013), p. 13-18 (cf. p. 58, 79).

[60] Robert M Haralick, Karthikeyan Shanmugam et al. "Textural features for image classification". In : *IEEE Transactions on systems, man, and cybernetics* 6 (1973), p. 610-621 (cf. p. 31).

[61] Paweł Hawryło et al. "Estimating defoliation of Scots pine stands using machine learning methods and vegetation indices of Sentinel-2". In : *European Journal of Remote Sensing* 51.1 (2018), p. 194-204 (cf. p. 26).

[62] Caner Hazirbas et al. "Fusenet : Incorporating depth into semantic segmentation via fusion-based cnn architecture". In : *Asian conference on computer vision*. Springer. 2016, p. 213-228 (cf. p. 111).

[63] G. E. Hinton et R. S. Zemel. "Autoencoders, minimum description length and Helmholtz free energy". In : *Advances in neural information processing systems*. 1994, p. 3-10 (cf. p. 39).

[64] Geoffrey E Hinton, Simon Osindero et Yee-Whye Teh. "A fast learning algorithm for deep belief nets". In : *Neural computation* 18.7 (2006), p. 1527-1554 (cf. p. 39, 40).

[65] Geoffrey E Hinton et Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks". In : *science* 313.5786 (2006), p. 504-507 (cf. p. 39).

[66] Sepp Hochreiter. "Untersuchungen zu dynamischen neuronalen Netzen". In : *Diploma, Technische Universität München* 91.1 (1991) (cf. p. 38).

[67] Sepp Hochreiter et Jürgen Schmidhuber. "Long short-term memory". In : *Neural computation* 9.8 (1997), p. 1735-1780 (cf. p. 38).

[68] K. Hornik. "Approximation capabilities of multilayer feedforward networks". In : *Neural networks* 4.2 (1991), p. 251-257 (cf. p. 34, 44).

[69] Chih-Wei Hsu et Chih-Jen Lin. "A comparison of methods for multiclass support vector machines". In : *IEEE transactions on Neural Networks* 13.2 (2002), p. 415-425 (cf. p. 29).

[70] Bohao Huang et al. "Large-scale semantic classification : outcome of the first year of inria aerial image labeling benchmark". In : *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE. 2018, p. 6947-6950 (cf. p. 58, 106).

[71] Chengquan Huang, LS Davis et JRG Townshend. "An assessment of support vector machines for land cover classification". In : *International Journal of remote sensing* 23.4 (2002), p. 725-749 (cf. p. 25).

[72] D. H. Hubel et T. N. Wiesel. "Receptive fields of single neurones in the cat's striate cortex". In : *The Journal of physiology* 148.3 (1959), p. 574-591 (cf. p. 35).

[73] G. Hughes. "On the Mean Accuracy of Statistical Pattern Recognizers". In : *IEEE Trans. Inf. Theor.* 14.1 (sept. 2006), p. 55-63. ISSN : 0018-9448 (cf. p. 25).

[74] Markus Immitzer, Francesco Vuolo et Clement Atzberger. "First experience with Sentinel-2 data for crop and tree species classifications in central Europe". In : *Remote Sensing* 8.3 (2016), p. 166 (cf. p. 30).

[75] Jordi Inglada et al. "Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery". In : *Remote Sensing* 7.9 (2015), p. 12356-12379 (cf. p. 58).

[76] Jordi Inglada et al. "Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series". In : *Remote Sensing* 9 (2017) (cf. p. 7, 17, 30, 139).

[77] Sergey Ioffe et Christian Szegedy. "Batch normalization : Accelerating deep network training by reducing internal covariate shift". In : *arXiv :1502.03167* (2015) (cf. p. 51).

[78] Kun Jia et al. "Land cover classification of Landsat data with phenological features extracted from time series MODIS NDVI data". In : *Remote sensing* 6.11 (2014), p. 11518-11532 (cf. p. 26).

[79] Bela Julesz. "Textons, the elements of texture perception, and their interactions". In : *Nature* 290.5802 (1981), p. 91 (cf. p. 31).

[80] P. Kaiser et al. "Learning Aerial Image Segmentation from Online Maps". In : *IEEE Transactions on Geoscience and Remote Sensing* (2017) (cf. p. 58).

[81] Leonard Kaufman et Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987 (cf. p. 23).

[82] Ronald Kemker, Carl Salvaggio et Christopher Kanan. "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning". In : *ISPRS Journal of Photogrammetry and Remote Sensing* (2018) (cf. p. 70).

[83] Diederik P Kingma et Jimmy Ba. "Adam : A method for stochastic optimization". In : *arXiv preprint arXiv :1412.6980* (2014) (cf. p. 45).

[84] Anders Knudby, Ellsworth LeDrew et Alexander Brenning. "Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques". In : *Remote Sensing of Environment* 114.6 (2010), p. 1230-1241 (cf. p. 26).

[85] Gregory Koch, Richard Zemel et Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition". In : *ICML deep learning workshop*. T. 2. 2015 (cf. p. 140).

[86] T. Kohonen. "Self-organized formation of topologically correct feature maps". In : *Biological cybernetics* 43.1 (1982), p. 59-69 (cf. p. 39).

[87] Alex Krizhevsky, Geoffrey Hinton et al. *Learning multiple layers of features from tiny images*. Rapp. tech. Citeseer, 2009 (cf. p. 55).

[88] Alex Krizhevsky, Ilya Sutskever et Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In : *Neural Information Processing Systems, 3-8 December, Lake Tahoe, USA*. 2012, p. 1097-1105 (cf. p. 41, 43).

[89] Camille Kurtz et al. "Multi-resolution region-based clustering for urban analysis". In : *International Journal of Remote Sensing* 31.22 (2010), p. 5941-5973 (cf. p. 23).

[90] Arnaud Le Bris et Nesrine Chehata. "Change detection in a topographic building database using submetric satellite images". In : *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 38.3/W22 (2011) (cf. p. 105).

[91] Y. LeCun et al. "Backpropagation applied to handwritten zip code recognition". In : *Neural computation* 1.4 (1989), p. 541-551 (cf. p. 35, 52).

[92] Jonathan Long, Evan Shelhamer et Trevor Darrell. "Fully convolutional networks for semantic segmentation". In : *IEEE Conference on Computer Vision and Pattern Recognition, IEEE/CVF, 7-12 June, Boston, USA*. 2015 (cf. p. 56).

[93] T. R. Loveland et al. "Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data". In : *International Journal of Remote Sensing* 21.6-7 (2000), p. 1303-1330 (cf. p. 22).

[94] A. Maas, F. Rottensteiner et C. Heipke. "Using label noise robust logistic regression for automated updating of topographic geospatial databases". In : *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* III-7 (2016), p. 133-140 (cf. p. 58).

[95] Alina E Maas et al. “Multitemporal Classification Under Label Noise Based on Outdated Maps”. In : *Photogrammetric Engineering & Remote Sensing* 84.5 (2018), p. 263-277 (cf. p. 26).

[96] Andrew L Maas, Awni Y Hannun et Andrew Y Ng. “Rectifier nonlinearities improve neural network acoustic models”. In : *Proc. icml*. T. 30. 1. 2013, p. 3 (cf. p. 43).

[97] J. MacQueen. “Some methods for classification and analysis of multivariate observations”. In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*. Berkeley, Calif. : University of California Press, 1967, p. 281-297 (cf. p. 22).

[98] D. Marmanis et al. “Classification with an edge : Improving semantic image segmentation with boundary detection”. In : *ISPRS Journal of Photogrammetry and Remote Sensing* 135 (2018), p. 158 -172 (cf. p. 57).

[99] Dimitrios Marmanis et al. “Classification With an Edge : Improving Semantic Image Segmentation with Boundary Detection”. In : *arXiv :1612.01337* (2016) (cf. p. 57).

[100] Paul M Mather et Magaly Koch. *Computer processing of remotely-sensed images : an introduction*. John Wiley & Sons, 2011 (cf. p. 25).

[101] Warren S McCulloch et Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In : *The bulletin of mathematical biophysics* 5.4 (1943), p. 115-133 (cf. p. 33).

[102] Farid Melgani et Lorenzo Bruzzone. “Classification of hyperspectral remote sensing images with support vector machines”. In : *IEEE Transactions on geoscience and remote sensing* 42.8 (2004), p. 1778-1790 (cf. p. 29).

[103] Marvin L Minsky et Seymour Papert. “Perceptrons : an introduction to computational geometry”. In : (1969) (cf. p. 34).

[104] Giorgos Mountrakis, Jungho Im et Caesar Ogole. “Support vector machines in remote sensing : A review”. In : *ISPRS Journal of Photogrammetry and Remote Sensing* 66.3 (2011), p. 247-259 (cf. p. 26).

[105] Klaus-Robert Müller et al. “An introduction to kernel-based learning algorithms”. In : *IEEE transactions on neural networks* 12.2 (2001) (cf. p. 28).

[106] M. Papadomanolaki et al. “Benchmarking deep learning frameworks for the classification of very high resolution satellite multispectral data”. In : *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* III-7 (2016), p. 83-88 (cf. p. 72).

[107] Charlotte Pelletier et al. “Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas”. In : *Remote Sensing of Environment* 187 (2016), p. 156-168 (cf. p. 58).

[108] Luis Perez et Jason Wang. “The effectiveness of data augmentation in image classification using deep learning”. In : *arXiv preprint arXiv :1712.04621* (2017) (cf. p. 50).

[109] Robert Gilmore Pontius Jr et Marco Millones. "Death to Kappa : birth of quantity disagreement and allocation disagreement for accuracy assessment". In : *International Journal of Remote Sensing* 32.15 (2011), p. 4407-4429 (cf. p. 62).

[110] Ning Qian. "On the momentum term in gradient descent learning algorithms". In : *Neural networks* 12.1 (1999), p. 145-151 (cf. p. 50).

[111] Rajat Raina, Anand Madhavan et Andrew Y Ng. "Large-scale deep unsupervised learning using graphics processors". In : *Proceedings of the 26th annual international conference on machine learning*. ACM. 2009, p. 873-880 (cf. p. 40).

[112] Herbert Robbins et Sutton Monro. "A stochastic approximation method". In : *The annals of mathematical statistics* (1951), p. 400-407 (cf. p. 48).

[113] Olaf Ronneberger, Philipp Fischer et Thomas Brox. "U-net : Convolutional networks for biomedical image segmentation". In : *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, p. 234-241 (cf. p. 58, 105, 107).

[114] Frank Rosenblatt. "The perceptron : a probabilistic model for information storage and organization in the brain." In : *Psychological review* 65.6 (1958), p. 386 (cf. p. 34).

[115] Franz Rottensteiner. "Building change detection from digital surface models and multi-spectral images". In : *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (IAPRS)* 36.3 (2007), p. 145-150 (cf. p. 105).

[116] J_W Rouse Jr et al. "Monitoring vegetation systems in the Great Plains with ERTS". In : (1974) (cf. p. 31).

[117] D. E. Rumelhart, G. E. Hinton et R. J. Williams. "Parallel Distributed Processing : Explorations in the Microstructure of Cognition, Vol. 1". In : MIT Press, 1986. Chap. Learning Internal Representations by Error Propagation, p. 318-362 (cf. p. 34, 37, 39, 44).

[118] R Saini et SK Ghosh. "Crop classification on single date Sentinel-2 imagery using random forest and support-vector machines". In : *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* (2018) (cf. p. 26).

[119] Mark Sandler et al. "Mobilenetv2 : Inverted residuals and linear bottlenecks". In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, p. 4510-4520 (cf. p. 108).

[120] Ozan Sener et Silvio Savarese. "Active learning for convolutional neural networks : A core-set approach". In : *arXiv preprint arXiv :1708.00489* (2017) (cf. p. 138).

[121] Jamie Sherrah. "Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery". In : *arxiv :1606.02585* (2016) (cf. p. 57).

[122] Wenzhe Shi et al. “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network”. In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, p. 1874-1883 (cf. p. 58).

[123] David Silver et al. “Mastering chess and shogi by self-play with a general reinforcement learning algorithm”. In : *arXiv preprint arXiv :1712.01815* (2017) (cf. p. 40).

[124] David Silver et al. “Mastering the game of Go with deep neural networks and tree search”. In : *nature* 529.7587 (2016), p. 484 (cf. p. 40).

[125] Patrice Y Simard, David Steinkraus, John C Platt et al. “Best practices for convolutional neural networks applied to visual document analysis.” In : *Icdar*. T. 3. 2003. 2003 (cf. p. 50).

[126] Karen Simonyan et Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In : *arXiv :1409.1556* (2014) (cf. p. 54, 55, 58, 72).

[127] Jake Snell, Kevin Swersky et Richard Zemel. “Prototypical networks for few-shot learning”. In : *Advances in Neural Information Processing Systems*. 2017, p. 4077-4087 (cf. p. 140).

[128] Nitish Srivastava et al. “Dropout : a simple way to prevent neural networks from overfitting.” In : *JMLR* 15.1 (2014), p. 1929-1958 (cf. p. 50).

[129] H. Steinhaus. “Sur la division des corps matériels en parties”. In : *Bull. Acad. Polon. Sci* 1 (1956), p. 801-804 (cf. p. 22).

[130] D. Stutz, A. Hermans et B. Leibe. “Superpixels : An evaluation of the state-of-the-art”. In : *Computer Vision and Image Understanding* 166 (2018), p. 1 -27 (cf. p. 98).

[131] Hao Su, Jia Deng et Li Fei-Fei. “Crowdsourcing annotations for visual object detection”. In : *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012 (cf. p. 55).

[132] Jérémie Sublime, Andrés Troya-Galvis et Anne Puissant. “Multi-Scale Analysis of Very High Resolution Satellite Images Using Unsupervised Techniques”. In : *Remote Sensing* 9.5 (2017) (cf. p. 23).

[133] Ilya Sutskever et al. “On the importance of initialization and momentum in deep learning”. In : *International conference on machine learning*. 2013, p. 1139-1147 (cf. p. 50).

[134] Christian Szegedy et al. “Going deeper with convolutions”. In : *IEEE Conference on Computer Vision and Pattern Recognition, IEEE/CVF, 7-12 June, Boston, USA*. 2015 (cf. p. 58, 72).

[135] R. Torres et al. “GMES Sentinel-1 mission”. In : *Remote Sensing of Environment* 120 (2012). The Sentinel Missions - New Opportunities for Science, p. 9 -24. ISSN : 0034-4257. DOI : https://doi.org/10.1016/j.rse.2011.05.028. URL : http://www.sciencedirect.com/science/article/pii/S0034425712000600 (cf. p. 13).

[136] Mihran Tuceryan et Anil K Jain. "Texture analysis". In : *Handbook of pattern recognition and computer vision*. World Scientific, 1993, p. 235-276 (cf. p. 31).

[137] A. M. Turing. "I.—COMPUTING MACHINERY AND INTELLIGENCE". In : *Mind* LIX.236 (oct. 1950), p. 433-460 (cf. p. 33).

[138] M. Volpi et D. Tuia. "Dense semantic labeling of sub-decimeter resolution images with convolutional neural networks". In : *IEEE Transactions on Geoscience and Remote Sensing* 55.2 (2017), p. 881-893 (cf. p. 57).

[139] Alexander Waibel et al. "Phoneme recognition using time-delay neural networks". In : *Backpropagation : Theory, Architectures and Applications* (1995), p. 35-61 (cf. p. 37).

[140] François Waldner, Guadalupe Sepulcre Canto et Pierre Defourny. "Automated annual cropland mapping using knowledge-based temporal features". In : *ISPRS Journal of Photogrammetry and Remote Sensing* 110 (2015), p. 1-13 (cf. p. 26).

[141] Li Wan et al. "Regularization of neural networks using dropconnect". In : *International conference on machine learning*. 2013, p. 1058-1066 (cf. p. 50).

[142] C. Wang et al. "A snow-free vegetation index for improved monitoring of vegetation spring green-up date in deciduous ecosystems". In : 196 (juil. 2017), p. 1-12 (cf. p. 15).

[143] P. J. Werbos. "Backwards differentiation in AD and neural nets : Past links and new opportunities". In : *Automatic differentiation : Applications, theory, and implementations*. Springer, 2006, p. 15-34 (cf. p. 34).

[144] P. J. Werbos. "Beyond Regression : New Tools for Prediction and Analysis in the Behavioral Sciences". Thèse de doct. 1974 (cf. p. 34, 44).

[145] Paul J. Werbos. "Generalization of backpropagation with application to a recurrent gas market model". In : *Neural Networks* 1.4 (1988), p. 339 -356 (cf. p. 37).

[146] Ronald J. Williams et David Zipser. "Backpropagation". In : sous la dir. d'Yves Chauvin et David E. Rumelhart. 1995. Chap. Gradient-based Learning Algorithms for Recurrent Networks and Their Computational Complexity, p. 433-486 (cf. p. 37).

[147] Fisher Yu et Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions". In : *arXiv preprint arXiv :1511.07122* (2015) (cf. p. 57).

[148] Matthew D Zeiler. "ADADELTA : an adaptive learning rate method". In : *arXiv preprint arXiv :1212.5701* (2012) (cf. p. 45).

[149] Matthew D Zeiler et Rob Fergus. "Visualizing and understanding convolutional networks". In : *European Conference on Computer Vision, 6-12 September, Zurich, Switzerland*. 2014, p. 818-833 (cf. p. 72).

[150] J. Zhang et J. Kerekes. “Unsupervised urban land-cover classification using WorldView-2 data and self-organizing maps”. In : *2011 IEEE International Geoscience and Remote Sensing Symposium*. 2011, p. 150-153 (cf. p. 23).

[151] X. X. Zhu et al. “Deep Learning in Remote Sensing : A Comprehensive Review and List of Resources”. In : *IEEE Geoscience and Remote Sensing Magazine* 5.4 (2017), p. 8-36 (cf. p. 68).