Prédiction d'indicateurs démographiques à partir d'images satellites

Basile Rousse

Université Paris Cité
Laboratoire d'Informatique Paris Descartes (EA 2517)
Institut National d'Etudes Démographiques

Prédiction d'indicateurs démographiques à partir d'images satellites

Présentée par Basile ROUSSE

dirigée par Valérie Golaz et Laurent Wendling

dans la spécialité Traitement du signal et des images

21 novembre 2024

Membres du jury

Dino Ienco	Directeur de recherche, INRAE	Rapporteur
Kathryn Grace	Professeure, University of Minnesota	Rapportrice
Nicolas Audebert	Directeur de recherche junior, IGN	Examinateur
Ronan Fablet	Professeur, LabSTICC, IMT Atlantique	Examinateur
Géraldine Duthé	Directrice de recherche, INED	Membre invité
Sylvain Lobry	Maître de Conférences, Université Paris Cité	Membre invité
Valérie Golaz	Directrice de rercherche, INED	Directrice de thèse
Laurent Wendling	Professeur, Université Paris Cité	Directeur de thèse

Remerciements

Pour commencer ce manuscrit, j'aimerais tout d'abord remercier la petite (petite ?) équipe d'encadrement, avec d'un côté Sylvain Lobry et Laurent Wendling qui m'ont guidé sur le chemin de la télédétection, et de l'autre côté Géraldine Duthé et Valérie Golaz qui ont réussi à faire de moi un démographe en herbe en un temps record (souvenez-vous de ma première présentation à l'INED, c'était pas gagné) !

J'aimerais aussi remercier chaleureusement Ronan Fablet d'avoir présidé mon jury, Kathryn Grace et Dino Ienco pour leurs relectures de ce manuscrit, ainsi que Nicolas Audebert d'avoir participé à l'évaluation de mes travaux pendant ces 3 ans. Je tiens également à remercier Guillaume Tochon, qui a participé au suivi de cette thèse.

J'ai passé ces trois dernières années entre mes deux labos, dont les sujets n'ont pas grand-chose à voir si ce n'est les images satellites, mais ce qui m'a permis de rencontrer plein de gens de plein de domaines différents. Ça me fait un sacré paquet de personnes avec qui les discussions étaient plus intéressantes les unes que les autres, et je tiens à tous vous remercier.

D'abord, je souhaiterais remercier les Shadocks Robin, Mohammed, Amine et Olivier Gadget pour leurs conseils de vieux briscards auxquels je ne faisais pas attention au début, mais en fait c'était pas déconnant. Une pensée spéciale pour Nathan, mon mentor, qui m'a guidé tout droit sous l'emprise des radars. Et comment oublier les apéros sur la terrasse de l'université avec probablement la plus belle vue de Paris, suivis de parties de Smash endiablées avec Hichem, Swann, Godzilla et le seigneur des coquillettes. LowGAN, en vrai après la thèse, la deuxième chose dont je suis le plus fier c'est de t'avoir battu au moins une fois (let's gooo). Rebecca, qui aurait cru qu'on finirait nos thèses dans les temps, franchement c'est carré. Je suis content qu'on n'ait pas fini dans les choux ! Je souhaite aussi le meilleur aux nouveaux, Nicolas et Marion, qui reprendront le flambeau du traitement d'images satellites au LIPADE ! Je souhaite également laisser un petit mot pour tous les autres de l'équipe SIP, Lazhar, Jérôme, Camille, Nicolas, Zhuxian, Xiaoyang, Yusin et Florence qui ont tous participé à ce cadre agréable que j'ai pu avoir pendant ma thèse.

Quant à vous, doctorants de l'INED, vous étiez un peu comme mes satellites Sentinel-2, on ne se voyait qu'une fois par semaine mais vous étiez toujours au rendez-vous (haha). Bon, d'accord je venais souvent les jours des petit-déj' gratuits, mais c'était quand même principalement pour vous voir. C'est aussi en vous écoutant parler de vos sujets avec passion que j'ai autant accroché avec la démographie, même si parfois c'était un peu glauque (merci Ariane, mais je préfère quand tu me parles de Solal et des lapins). Lucie, tu es un peu la maman ultime du bureau, au sens littéral et au sens figuré du terme (merci pour les gâteaux). J'espère que tu vas prendre soin de tes enfants par adoption, en particulier de Julie. Promis, un jour on trouvera pourquoi ils ont tué le chien ! On peut ouvrir un centre de recherche à Carqueiranne avec une immense bibliothèque, t'en penses quoi ? Pour la « Remote Sensing Crew », la plus fine des équipes, j'espère que vous garderez la tête vers les étoiles. Mais mes désirs secrets resteront de te retrouver, Léo, à la ligne d'arrivée d'un semi, et Ankit de savoir faire des posters aussi beaux que les tiens. Ritu, you may not be part of the crew, but thank you for all the happiness you gave me, from the beginning when I was only the shy guy to the time when you unleashed my gossiping powers. "Computer guy" has done it! Nestor, Margaux et moi attendons toujours les pains au chocolat. Je souhaite aussi remercier toutes les autres personnes que j'ai pu côtoyer à l'INED, Arlette, Heini, Andrea, Paul, Mariam, Narovana, Pierre, Adriana et tous les autres,

qui m'ont permis de développer mon sens scientifique à travers des discussions à la pause café.

Et puis il y a tous les autres qui m'ont soutenu alors que tout le monde ne savait même pas ce que je faisais. Olivier, en cherchant comment je pouvais te remercier, je me suis rappelé de comment on s'est rencontré. Kigali, tu n'avais ni vêtement ni eau chaude dans un hôtel relativement douteux. Malgré ça, tu as gardé ta bonne humeur, et finalement ta chemise est arrivée pour ta super présentation (oublie pas d'utiliser les mains pour parler). C'est une belle image de ta mentalité de guerrier, qui arrive à son paroxysme lorsqu'il s'agit d'aller chercher un amorti ou de mettre des coups de raquette. On se fait un squash bientôt ? Mais si je ne suis pas dispo, c'est parce qu'Arthur m'a emmené courir, pour me réconforter de m'avoir spoilé l'Attaque des Titans. C'est ta faute si je ne peux plus m'arrêter de courir, et bientôt le vélo (tout sauf un Canyon). T'as intérêt à revenir de NY t'échouer une nouvelle fois dans la côte des Gardes. Léa, si jamais tu ne cours pas avec nous, on t'attend avec la plus belle des pancartes ! Je pense que M. Halberstadt serait particulièrement fier de nous. Il faudra que tu rajoutes Virginie (ou Gollum) sur ta pancarte, qui a probablement la progression la plus fulgurante que j'ai vue jusqu'ici ! Tu as étonnamment bien compris toutes mes interrogations et angoisses, en particulier dans cette dernière phase de la thèse, mais quand j'y pense c'est pas si étonnant vu qu'on est pareil... A quand la thèse du coup ? Puisqu'on est en Terre du Milieu, je remercie Johanne de m'avoir accueilli dans sa maison de Hobbit après que j'ai décalé mon train de retour un certain nombre de fois (encore désolé pour le spam Adriana). Mais bon, Ganon n'a qu'à bien se tenir. Je suis content de n'avoir que bu des bières avec notre Hobbit déboiteur national Maxime, qui a eu la gentillesse de ne pas envoyer mon épaule rejoindre le programme Copernicus. Le petit Gaspard, a aussi été un solide soutien dans la mêlée (enfin vous l'aurez compris, plus solide que son épaule). Question solidité on peut aussi parler de Koby (et de son foie) qui est apparue sans prévenir personne pour me donner un coup de boost monumental dans l'amélioration de mon endurance fondamentale. En parlant de soutien, Emma at the top ! C'est à moi de te remotiver maintenant pour cette dernière ligne droite, je crois en toi ! Avec Jordan, soyons honnêtes, vous êtes tous les deux mes *numero uno*. Jordan, n'oublie pas qu'on a un DM de maths à rendre pour la rentrée. J'ai aussi une petite pensée pour Alinache avec qui je me suis lancé dans la recherche, et pour Parth le meilleur chanteur de karaoké de Corée, Lisa, Virginie, Arnaud, Julie, et tous les autres avec qui j'ai passé des moments d'exceptions. *Oh Jaja, Réréneujaja, j'ai fini ma thèse Jaja, mais, j'espère qu'elle te plaira même si des paillettes y'en a pas.*

Et bien sûr, comment ne pas laisser un petit mot pour Lucrezia. Non ci sono parole per esprimere la mia gratitudine. Sei stata presente nei momenti belli e (soprattutto) in quelli difficili, trovando sempre le parole giuste per motivarmi. Meriti una montagna di waffle (don't kill me). Ti prometto che un giorno finiremo The Office in cima alle piste con un cappuccino, o quando raggiungerai Nathan e me! Remember, the Force will always be with you, bitch.

J'aimerais réserver ces derniers mots pour remercier ma famille de m'avoir soutenu pendant ces trois années de thèse, mais aussi pendant toutes mes études. Vous m'avez donné le cadre idéal pour que je puisse m'éclater dans ce que je fais, je ne pouvais pas rêver mieux ! Tout le monde, big up à mon papa, Hana, Judith, papy, mamie et à mon petit frère Jules, très certainement le meilleur supporter que l'on puisse avoir, et qui m'a tant laissé gagner à Mario Kart pour me remonter le moral.

Sommaire

Liste des Figures

Liste des Tableaux

Résumé

En Afrique subsaharienne, les recherches sur le lien entre la population et l'environnement sont souvent limitées par la disponibilité des données dans ces deux domaines, qui sont souvent rares, parcellaires et datées. Les données satellites, en revanche, offrent une opportunité précieuse pour caractériser précisément l'environnement grâce à leur haute résolution et à leur fréquence de mise à jour. Ces informations environnementales peuvent ensuite être intégrées aux données démographiques, permettant ainsi une analyse plus complète et actualisée des dynamiques de population en tenant compte des facteurs environnementaux. Cependant, le traitement de ces données nécessite des méthodes adaptées, telles que l'apprentissage profond, pour répondre aux contraintes spatiales et temporelles qu'imposent l'exploitation d'enquêtes démographiques spécifiques. Cette recherche aborde deux échelles d'analyse : celle d'une ville et celle d'un pays. Chaque étude se divise en deux parties principales. D'abord, une méthode d'adaptation de domaine utilisant l'apprentissage profond pour la classification des Local Climate Zones (LCZ) est développée, afin de caractériser précisément l'environnement de la zone d'étude. Ensuite, cette caractérisation est utilisée pour analyser les données de population tout en tenant compte des facteurs socio-économiques, afin de garantir que les phénomènes observés résultent d'un effet de l'environnement et non des données de population elles-mêmes. Le premier cas d'étude concerne la mortalité à Antananarivo la capitale de Madagascar. La méthode d'adaptation utilise les définitions physiques et géométriques des LCZ comme base universelle. Nous établissons ensuite des liens entre plusieurs caractéristiques environnementales, dont les LCZ, et la mortalité par cause estimée dans les différents quartiers de la ville. Le deuxième cas d'étude porte sur le paludisme au Burkina Faso. La méthode d'adaptation de domaine des LCZ prend ici en compte des variations saisonnières pour extraire les informations pertinentes des images satellites et classifier les LCZ. Nous mettons en évidence un lien significatif entre certains profils de zones classés selon les LCZ qui les composent et la présence de paludisme parmi les enfants au sein des ménages. Ces résultats soulignent l'intérêt d'une caractérisation fine de l'environnement pour l'analyse des données de population. Ils suggèrent que cette approche pourrait améliorer la modélisation des données démographiques et permettre l'estimation de certains indicateurs.

Mots-clés : Apprentissage profond, adaptation de domaine, télédétection, démographie, Local Climate Zones, Afrique subsaharienne.

Asbtract

In Sub-Saharan Africa, research on the link between population and the environment is often constrained by the availability of data in these two areas, which are frequently scarce, fragmented, and outdated. Remote sensing data, on the other hand, offer a valuable opportunity to precisely characterize the environment due to their high resolution and frequent updates. This environmental information can then be integrated with demographic data, enabling a more comprehensive and up-to-date analysis of population dynamics while considering environmental factors. However, processing these data requires suitable methods, such as deep learning, to address the spatial and temporal constraints imposed by the use of specific demographic surveys. This research addresses two levels of analysis: a city and a country. Each study is divided into two main parts. First, a domain adaptation method using deep learning for the classification of Local Climate Zones (LCZ) is developed to precisely characterize the environment of the study area. Then, this characterization is used to analyze population data while taking into account socio-economic factors, ensuring that the observed phenomena result from an environmental effect rather than from the demographic data itself. The first case study focuses on mortality in Antananarivo, the capital of Madagascar. The adaptation method uses the physical and geometric definitions of LCZ as a universal basis. We then establish links between various environmental characteristics, including LCZ, and cause-specific mortality estimated in different neighborhoods of the city. The second case study tackles malaria in Burkina Faso. The domain adaptation method for LCZ here considers seasonal variations to extract relevant information from satellite images and classify LCZ. We highlight a significant link between certain areas classified according to their LCZ characterization and the prevalence of malaria among children in households. These results highlight the importance of fine environmental characterization for analyzing population data. They suggest that this approach could enhance demographic data modeling and allow for the estimation of certain indicators.

Keywords: Deep learning, domain adaptation, remote sensing, Demography, Local Climate Zones, sub-Saharan Africa.

Chapitre 1

Introduction

Let's goo

– Hichem Boussaid

La communauté scientifique produit de plus en plus d'études utilisant des caractérisations de l'environnement notamment grâce à l'augmentation des données satellites disponibles en accès libre. La démographie est une des applications de ces données, car les populations sont directement impactées par les phénomènes environnementaux, mais agissent aussi sur cet environnement. Cependant, ces interactions sont complexes et plusieurs obstacles subsistent : Comment modéliser l'environnement adéquatement pour les études de population ? Comment intégrer cette modélisation dans une analyse démographique ? Aujourd'hui, les données satellites sont de plus en plus accessibles, grâce aux satellites d'imagerie qui permettent d'estimer des indicateurs environnementaux. Ce sont donc des sources précieuses pour les études de l'interaction entre population et environnement. Cette thèse est à l'interface entre trois domaines de recherche, mêlant informatique, télédétection et démographie. Ces travaux visent à étudier l'utilisation de ces images satellites pour des analyses de données de population, dans le contexte de l'Afrique subsaharienne. Cette région du monde connaît de fortes dynamiques de population et notamment une forte croissance démographique du fait d'une fécondité encore relativement élevée, mais aussi une forte mortalité et d'importants déplacements de population. Cependant, les données démographiques fines y sont rares ou parcellaires. La situation est la même pour les données environnementales. Les images satellites représentent une nouvelle source de donnée qui pourrait permettre d'apporter de précieuses informations environnementales qui, si appariées à des données démographiques, pourraient apporter des informations sur l'interaction entre population et environnement.

Dans ce chapitre introductif, nous donnons quelques éléments de présentation pour chaque domaine afin de mieux appréhender ce manuscrit. Nous aborderons d'abord les concepts autour de la démographie (Section 1.1), puis de télédétection (Section 1.2) et d'apprentissage automatique (Section 1.3). Ces trois sous-sections introductives permettront de mieux comprendre le contexte et les objectifs de recherche, présentés en sections 1.4 et 1.5 : établir une chaîne de traitement, depuis les images satellites jusqu'aux analyses démographiques.

1.1 Démographie

Lorsque l'on pose la question de ce qu'est la démographie, beaucoup pensent que cela se résume à « compter les gens ». Lorsque j'ai postulé à ce sujet, je pensais exactement la même chose. Bien que ce soit quand même une partie (par exemple, Léo, un collègue, tente d'atteindre le graal : estimer la population depuis la détection du bâti détecté à partir d'images satellites), la démographie est bien plus large que le simple dénombrement des individus. La démographie regroupe l'ensemble des études dont l'objectif est de caractériser des populations : leur nombre mais aussi leur composition, et leur dynamique. Les études peuvent avoir des approches variées, quantitatives, qualitatives ou mixtes, à visée purement descriptive, ou explicative. Les thèmes abordés sont nombreux, même si le coeur de la démographie porte sur les questions de fécondité, de migration et de mortalité qui sont les principaux moteurs de la dynamique de population. Ainsi, la démographie qui regroupe les études de population interagit avec de nombreuses autres disciplines comme la sociologie, l'histoire, la géographie, la santé publique... Les observations issues des analyses des données d'enquêtes démographiques peuvent servir à l'élaboration de politiques publiques, mais aussi à leur évaluation. Par exemple, l'étude de la fécondité fournit des informations cruciales pour la santé publique ou la santé de la re-

production, ainsi que la projection des besoins sociaux et économiques d'un pays à moyen et long terme. En particulier, les réflexions autour du vieillissement de la population, comme la planification du système de retraite, dépendent très fortement des projections de fécondité.

Il est compliqué de dater les débuts de la démographie, comme pour d'autres disciplines. Bien que des écrits sur la population aient été retrouvés depuis l'antiquité, comme dans la Rome Antique en Occident, les débuts de la démographie moderne sont associés aux travaux de John Graunt, avec l'aide de William Petty, qui a mis au point les premières méthodes statistiques de recensement et la première table de mortalité au 17e siècle. Ces travaux sont considérés comme le socle de la démographie moderne.

Lorsque l'on parle de l'étude de la population, cela inclut donc tous les aspects qui peuvent modifier sa dynamique. Ces dernières années, avec l'émergence des questions environnementales, l'intérêt pour l'étude des interactions entre les dynamiques de population et l'environnement ne cesse de croître. Ces interactions sont complexes la population étant dépendante de l'environnement et inversement. Outre le lien causal, d'autres facteurs sont à prendre en compte et notamment la composition de la population, afin d'assurer que les liens trouvés n'en dépendent pas. Prenons un exemple. Pourquoi les quartiers les plus bourgeois se trouvent à l'Ouest dans le bassin parisien ? Une question de vent, donc d'environnement. À partir de la révolution industrielle, les usines ont commencé à rejeter des vapeurs odorantes et dérangeantes. Or, le vent a tendance à souffler d'Ouest en Est, amenant donc les vapeurs vers l'Est. Pour éviter ce désagrément, les populations les plus riches se sont installées à l'Ouest, laissant les populations les plus modestes à l'Est. Les conditions météorologiques, ici, ont eu un impact direct sur les caractéristiques socio-économiques des populations de la région parisienne. Dans un tel cas, les différences observées entre les populations selon leur localisation et leur environnement peuvent en partie être à des différences socio-économiques qu'il est important de prendre en compte. Cet exemple nous montre aussi l'importance de l'échelle d'étude. Au niveau de la région, ou de la ville de Paris, ces disparités seraient masquées car elles sont à un niveau plus bas, celui des arrondissements. La mise en évidence de lien entre environnement et population dépend donc du niveau d'analyse.

Dans cette thèse, le mot "environnement" est utilisé pour définir une notion très large. Nous le définissons comme l'ensemble des éléments naturels ou artificiels qui entourent et peuvent agir sur une population : le climat, le type de végétation ou encore la densité d'une ville et l'exploitation des sols. La gestion des corrélations entre données de population et de cet environnement n'est pas le seul défi à relever. Si nous voulons étudier l'interaction entre environnement et population, il faut aussi une modélisation adaptée pour l'application. Il y a plusieurs façons de récupérer des informations environnementales, que l'on peut séparer selon la date d'acquisition par rapport à la collecte des données démographiques :

- Les données environnementales sont récupérées **pendant** la collecte des données de population. Ces données sont alors recueillies par l'enquêteur ou par la personne enquêtée elle-même. Si les valeurs ne sont pas récupérées par des capteurs mais par l'humain, elles peuvent être subjectives.

- Les données environnementales sont récupérées **après** la collecte des données de population. Les données doivent être rassemblées via d'autres moyens, comme à partir de bases de données nationales et internationales ou directement à partir de données brutes,

comme les images satellites. Évidemment, cela est possible sous réserve que ces données existent aux dates correspondant à l'enquête afin de ne pas avoir un biais temporel.

Les images satellites sont une source de données intéressante pour la démographie, car elles possèdent les trois caractéristiques recherchées. Elles sont disponibles à grande échelle, très fréquentes, et ont des résolutions élevées. Cela permet de caractériser l'environnement partout dans le monde, pendant toute période d'intérêt et avec une suffisamment bonne précision, même pour des études à l'échelle locale. Lorsqu'elles sont liées aux données démographiques et sanitaires, ces images peuvent fournir des informations précieuses sur les facteurs de risque environnemental [1]–[3], par exemple pour expliquer des données de santé. Par exemple, Gibb *et al.* (2023) [2] utilisent la télédétection pour étudier les interactions entre le climat et la dengue, maladie vectorielle, au Vietnam. Cette thèse se concentre sur l'Afrique subsaharienne. Ces pays subissent des contraintes climatiques, comme l'alternance entre saisons sèches et saisons des pluies, et les évènements climatiques extrêmes qui ont un impact fort sur les phénomènes de population, comme la santé [4], [5] ou les migrations [6]. Les images satellites permettent non seulement de compléter les données manquantes, mais aussi de prendre en compte ces événements climatiques, qu'ils soient soudains ou qu'ils s'inscrivent dans un temps plus long.

La partie suivante présente quelques éléments sur les données satellites qu'il est possible d'avoir, et leurs applications et apports potentiels en démographie.

1.2 Télédétection

Dans sa définition la plus générale, la télédétection est une méthode qui permet d'obtenir des informations sur un objet à l'aide d'un instrument qui n'a pas de contact direct avec cet objet. Dans la suite de ce manuscrit, le terme télédétection renvoie à la télédétection spatiale, dans laquelle l'instrument utilisé pour le capteur des informations est un engin spatial.

Les premiers satellites d'observation ont été lancés au milieu du 20^{e} siècle dans le cadre de la conquête spatiale. Dans un premier temps et dans le contexte de la Guerre Froide, ces satellites étaient réservés à un usage militaire.

Le premier satellite civil, **Landsat-1** (1972), a été lancé par la NASA pour le suivi des récoltes céréalières aux États-Unis. La France inaugure son premier satellite un peu plus tard, en 1986, avec le programme **SPOT** (**S**atellite **P**our l'**O**bservation de la **T**erre), qui est aussi le premier satellite à usage commercial. Aujourd'hui, il y a un très grand nombre de satellites d'observation de la Terre (788 en 2019[1]), qui utilisent de nombreux capteurs différents, pour des applications différentes, que ce soit la détection de bateaux, la gestion de l'utilisation des sols ou la surveillance de l'océan. Parmi tous les capteurs qui existent, nous retrouvons par exemple les satellites qui utilisent le radar, très utiles car ils ne sont pas sensibles à la couverture nuageuse, des capteurs multi-spectraux et hyper-spectraux souvent utilisés pour la caractérisation des sols, ou encore des capteurs de pollution. Cependant, les données issues des différents capteurs ne sont pas en accès libre, la modalité de partage dépendant de l'organisme ou de l'entreprise gérant les satellites. Par exemple, les entreprises Airbus Defense and Space

[1] hhttps://www.ucsusa.org/resources/satellite-database

Bande	Résolution	Longueur d'onde centrale (nm)
B01	60 m	443
B02 (visible)	10 m	490
B03 (visible)	10 m	560
B04 (visible)	10 m	665
B05	20 m	705
B06	20 m	740
B07	20 m	783
B08	10 m	842
B8A	20 m	865
B09	60 m	940
B10	60 m	1375
B11	20 m	1610
B12	20 m	2190

Tableau 1.1: Longueurs d'onde centrales de bandes spectrales du capteur d'images Sentinel-2.

et Planet produisent des données optiques à très haute résolution (inférieure à trois mètres), mais qui ne sont pas en libre accès. Depuis 2015, l'Agence Spatiale Européenne (ESA) a lancé plusieurs séries de satellites avec son programme d'observation Copernicus pour produire des données satellites disponibles, au contraire, en libre accès pour aider le développement des applications du spatial. Ces données, bien que très fréquentes, ont une résolution plus basse que certains satellites commerciaux, ce qui peut limiter leur utilisation. Les capteurs disponibles via ce programme sont nombreux : Radar (Sentinel-1), multi-spectral (Sentinel-2), ou encore de pollution (Sentinel-5p). En particulier, les satellites Sentinel-2 produisent des images avec 13 bandes spectrales dans le domaine du visible et de l'infrarouge, très adaptées pour la gestion de l'utilisation des sols et la végétation. Les caractéristiques de ces bandes sont disponibles dans le tableau 1.1. Ces images disponibles en libre accès, à très haute fréquence (cinq jours maximum) et partout dans le monde permettent de générer des indicateurs environnementaux complexes, qui peuvent être inclus dans un processus d'analyse des données de population. Dans la suite de ce manuscrit, si le contraire n'est pas précisé, les indicateurs environnementaux sont générés à partir de ces images Sentinel-2. Depuis leur lancement en 2015, les satellites ont donc produit une quantité de données astronomiques, qu'il est nécessaire de traiter avant usage. Les récentes méthodes d'apprentissage automatique, en particulier d'apprentissage profond qui est une branche de l'apprentissage utilisant les réseaux de neurones, permettent de traiter une telle quantité de données.

La section suivante donne un aperçu général du domaine de l'apprentissage automatique et de l'apprentissage profond, et décrivent leurs utilisations possibles pour caractériser l'environnement dans les pays d'Afrique subsaharienne.

1.3 Apprentissage automatique

L'apprentissage automatique (*Machine Learning*) est une discipline de l'intelligence artificielle, fondée sur des approches mathématiques pour permettre à des modèles de résoudre des tâches complexes sans avoir été explicitement programmés pour cela. Les modèles "apprennent" d'un ensemble de données qui leur est présenté, c'est-à-dire qu'ils les analysent, comprennent et retiennent quelles sont les informations intéressantes à la résolution de la tâche. L'apprentissage consiste donc en l'optimisation automatique d'un modèle pour que son erreur statistique soit la plus faible possible.

Cet entraînement de modèles est apparu avec les premiers ordinateurs. Arthur Samuel est le premier informaticien à faire usage de l'apprentissage automatique. Il programme un modèle permettant de jouer au jeu de dames, visant à chaque coup à minimiser le gain de l'adversaire et à maximiser son gain. Différentes méthodes ont été mises au point dans la seconde partie du 20^{e} siècle, suivant le gain de puissance de calcul des ordinateurs. En particulier, le développement des réseaux de neurones commence à cette époque. L'objectif est de s'inspirer du fonctionnement cognitif d'un humain pour résoudre des tâches via un ordinateur. Le premier *perceptron* a été créé en 1957 par Donald Hebb, pour la reconnaissance de forme. Celui-ci est composé d'une couche d'unités de calculs élémentaires, appelées neurones, en entrée du modèle, et une autre en sortie. Ces neurones reçoivent plusieurs entrées qu'ils combinent en fonction de l'importance relative de chacune d'entre elles, et produisent une sortie en fonction de cette combinaison et d'une fonction d'activation. Des réseaux de neurones plus complexes ont ensuite été développés à partir des années 1990. Combinés avec l'augmentation de la puissance de calcul parallèle, ils sont devenus les modèles à l'état de l'art dans de nombreux domaines, comme le traitement du langage naturel et la vision par ordinateur.

Toutes les méthodes d'apprentissage automatique, qu'elles utilisent des réseaux de neurones ou d'autres algorithmes, nécessitent généralement une phase d'apprentissage composée de deux processus principaux. Le premier processus est le processus d'entraînement du modèle. Le modèle est vu comme une fonction qui est optimisée de manière itérative. Les données d'entraînement sont présentées au modèle, et les paramètres de celui-ci sont ajustés pour résoudre une tâche précise. Après un ou plusieurs passages sur l'ensemble des données (appelés *epochs*), le modèle est suffisamment entraîné pour passer à l'étape suivante. La deuxième étape, dite de test, permet d'évaluer la qualité des prédictions du modèle. De nouvelles données, différentes des données d'entraînement, sont présentées au modèle pour effectuer des prédictions. Les résultats de cette étape permettent d'évaluer la capacité du modèle à généraliser, c'est-à-dire à transférer ses connaissances à des données qu'il n'a jamais vues auparavant. Si les résultats sont bons, cela signifie que le modèle a bien été

Comme l'apprentissage automatique est basé sur les données d'entraînement, sa capacité à avoir de bonnes performances en test est fortement dépendante de la qualité et de la quantité de ces données ainsi que du processus d'entraînement. Ce trio est indispensable pour obtenir un modèle performant en production.

L'apprentissage automatique peut naturellement être appliqué aux données satellites, comme une tâche de vision par ordinateur dans le cas des images satellites. Ces images sont

disponibles en grande quantité en particulier depuis le début de l'ère Copernicus. Avant ces débuts, l'apprentissage profond avait déjà été appliqué aux images satellites, par exemple pour la cartographie des routes [7], mais la combinaison de ces deux domaines a connu un grand essor après 2014. Plusieurs tâches, détournées des tâches usuelles d'apprentissage automatique, sont réalisées grâce à ces apprentissages. Trois des principales sont :

- **Classification d'images** : Utilisation de modèles appris pour classer les types de couverture terrestre ou détecter des objets spécifiques dans les images satellite (bâtiments, routes, véhicules).
- **Détection des Changements** : Développement de modèles pour identifier et quantifier les changements dans des séries temporelles d'images.
- **Segmentation sémantique** : Utilisation de modèles appris pour segmenter une image en régions d'intérêt, par exemple, en identifiant les différentes cultures dans une image agricole.

Toutes ces méthodes permettent de caractériser l'environnement et d'apporter des informations complémentaires aux données de population. Dans ces travaux, nous utiliserons la classification d'images, qui est la méthode la plus adaptée pour notre application. Ce choix sera justifié dans la prochaine partie.

1.4 Contexte de la thèse

Cette thèse a pour but d'explorer le lien entre population et environnement dans les pays d'Afrique subsaharienne, où les données démographiques sont souvent rares. En plus des données relatives aux ménages interrogés, certaines enquêtes démographiques fournissent leurs coordonnées géographiques. Cependant, ces coordonnées ne sont pas toujours précises : elles peuvent être seulement disponibles à une échelle supérieure ou volontairement déplacées pour préserver la confidentialité des ménages. Elles permettent tout de même de relier les ménages à leur environnement proche à travers des indices, comme pour la végétation (*Normalized Difference Vegetation Index*, NDVI), la température, les précipitations, l'aridité [8], ou l'empreinte de l'activité humaine [9], [10]. Les travaux cités ici utilisent souvent des indicateurs de base tels que la végétation ou les précipitations, et bénéficieraient d'une caractérisation environnementale plus détaillée.

Afin de soutenir l'effort de liaison des données démographiques et environnementales, les enquêtes issues du programme des enquêtes démographiues et de santé (*Demographic and Health Survey*, DHS) fournissent ce type d'indicateurs pour les ménages enquêtés. Cependant, ces données ne sont pas toujours disponibles à la date de l'enquête : par exemple, le *Global Human Settlement Layer* (GHSL) [10] n'est disponible que jusqu'à l'année 2014. Cela peut induire un biais important entre les données environnementales et la réalité terrain, notamment dans les pays où la croissance démographique et l'urbanisation sont rapides comme en Afrique subsaharienne. De plus, la résolution spatiale de ces données, généralement supérieure à 1 km,

ne permet pas toujours de percevoir la variabilité des indicateurs environnementaux qui pourrait permettre une meilleure compréhension des données démographiques. Un suivi environnemental plus précis est rendu possible par l'augmentation des données satellites disponibles et de leur résolution. En particulier, les images optiques Sentinel-2, fournies en accès libre par l'Agence Spatiale Européenne, permettent un suivi fréquent et à haute résolution. Ces images peuvent être utilisées pour la cartographie de l'environnement des ménages au moment de la collecte des données de population.

Le choix de la caractérisation ainsi que le processus de caractérisation en lui-même ne sont pas triviaux. Depuis l'augmentation de données disponibles et la spécification des besoins des domaines applicatifs, la communauté scientifique s'est activement penchée sur cette tâche. En particulier, de nombreux jeux de données globaux ont été introduits pour cartographier l'environnement à différents niveaux de précision, et servent de base pour la génération d'indicateurs en utilisant l'apprentissage automatique. Par exemple, SEN12MS [11] possède une classification à 11 classes, 10 réservées pour la végétation mais une seule pour les régions urbaines (classe *Urban and built-up areas*). BigEarthNet [12] est introduit pour résoudre un problème de classification d'image à plusieurs étiquettes, dans lequel chaque étiquette représente une classe présente dans l'image. Cependant, il ne comprend que des images de certains pays d'Europe, ce qui limite son usage à l'échelle globale. Un dernier exemple est So2Sat [13]. Il utilise le système de classification **Local Climate Zones (LCZ)** qui permet une classification basée sur des critères physiques du terrain. Ce système permet de caractériser l'environnement avec une bonne précision dans les deux contextes, urbain et rural, et est donc un avantage certain pour son utilisation dans d'autres domaines d'application comme la démographie. Il sera utilisé comme caractérisation privilégiée de l'environnement dans les travaux présentés ici.

Cette thèse résulte d'une collaboration entre l'équipe "Système Intelligent de Perception" du LIPADE (Laboratoire d'informatique Paris Descartes), spécialisée dans le traitement d'images, et l'unité de recherche Demosud (Démographie des pays du sud) de l'INED (Institut National d'Études Démographiques), qui se étudie les dynamiques démographiques en oeuvre dans les pays du sud. Cette synergie permet de combiner les compétences nécessaires en traitement d'images, analyse de l'environnement et études démographiques pour mener à bien ces recherches. Ces travaux sont financés par le "Data Intelligence Institute of Paris" (DIIP) de l'Université Paris Cité et par l'Agence Nationale de Recherche (ANR).

1.5 Objectifs de recherche

L'objectif principal de cette thèse est l'étude de l'interaction entre l'environnement et les phénomènes de population dans un contexte de pénurie de données précises comme c'est le cas en Afrique subsaharienne. Une chaîne complète, depuis le traitement des données de télédétection et de population jusqu'à la mise en valeur de l'interaction environnement/population, doit être constituée. Les objectifs de recherche suivants détaillent les grandes étapes de cette chaîne. Dans un premier temps, il s'agira de développer des méthodes d'adaptation de domaine pour la cartographie de l'environnement dans le cas complexe de l'Afrique subsaharienne. Ces méthodes doivent en plus respecter les contraintes liées à la démographie : une carte doit pouvoir être générée pour une date donnée, sur une région donnée et mises en lien avec

les données collectées à ce moment là et dans cet espace là. Afin de permettre une meilleure reproductibilité des travaux, on ajoute une contrainte supplémentaire : utiliser uniquement des données libres d'accès (gratuites) et ne pas nécessiter d'autres annotations que les jeux de données déjà existants. La seconde question de recherche est la suivante : comment lier une carte environnementale à une enquête en démographie, en prenant en compte les incertitudes sur les positions des ménages ? La solution à ce problème doit être la plus robuste possible, afin de ne pas mettre en évidence des corrélations qui pourraient être causées par une interaction mal décrite.

Enfin, la résolution de notre dernière interrogation permet de répondre à notre question principale : quelles sont les interactions entre l'environnement et la population ? L'intégration de la caractérisation de l'environnement des ménages comme variable explicative à part entière dans un modèle de démographie permet de répondre à cette question.

Le manuscrit de cette thèse est structuré autour de deux chapitres principaux (3 et 4) qui traitent des deux études réalisées. Le premier chapitre pose les bases techniques nécessaires à la compréhension des deux parties suivantes et présente l'état de l'art des méthodes d'adaptation de domaine en apprentissage profond. La première étude (chapitre 2), menée à l'échelle d'une ville, porte sur le lien entre environnement et causes de mortalité dans les quartiers d'Antananarivo, la capitale de Madagascar. La deuxième étude, menée à l'échelle nationale, porte sur le lien entre environnement et présence de paludisme au Burkina Faso. Ces deux chapitres sont organisés de la même manière. Dans un premier temps, la méthodologie employée pour la cartographie de l'environnement est présentée. L'intégration de cette caractérisation dans une analyse démographique est présentée dans un second temps. Les différentes étapes de ces processus sont illustrés en Figure 1.1, et seront détaillées dans chacun des chapitres. Les résultats et leur discussion sont présents en fin de chaque chapitre. Enfin, le dernier chapitre conclut ce manuscrit et propose des perspectives de recherche.

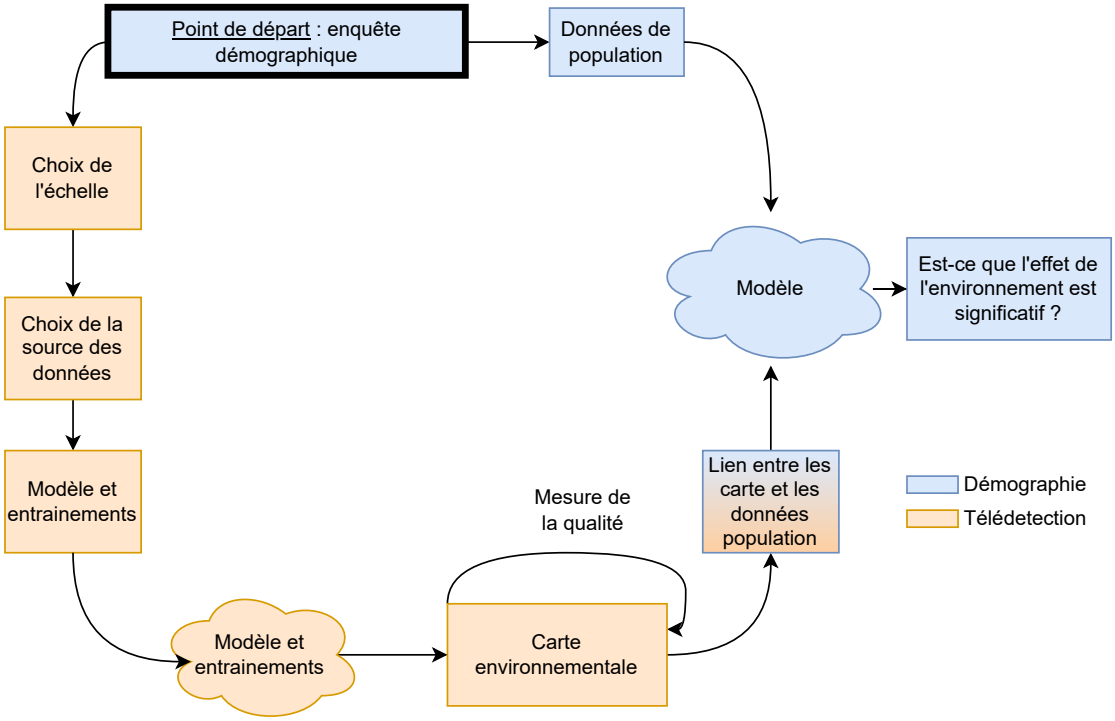


Figure 1.1: Schéma global représentant les travaux effectués pendant la thèse. Le point de départ des projets sont les données de l'enquête démographique cible. Ces données permettent d'orienter le choix des données satellites pour la caractérisation de l'environnement. Enfin, les données de population et d'environnement sont mobilisées dans des modèles statistiques pour mettre en lumière des liens entre les deux dimensions.

Chapitre 2

Description des outils

Basile, pourquoi je me suis réveillée avec les crocodiles dans la tête ?

– Lucie Vanhoutte

Pour répondre à notre problématique, les travaux menés consistent en premier lieu à étudier la génération d'indicateurs environnementaux à partir d'images satellites. Dans notre cas, il existe peu de données environnementales en Afrique subsaharienne pour pouvoir entraîner des modèles. Des méthodes d'adaptation de domaines devront donc être mobilisées.

L'utilisation de modèles d'apprentissage, comme les réseaux de neurones profonds, permettent d'analyser une image pour en extraire des caractérisations environnementales. De telles caractérisations peuvent aussi être réalisées par des experts, mais ce travail est coûteux et fastidieux en particulier lorsque les zones d'études sont grandes comme à l'échelle d'un pays. Cette étape de caractérisation peut être réalisée à l'aide de méthodes automatiques de classification. Ces méthodes, basées sur l'apprentissage automatique, sont optimisées à l'aide de jeux de données d'images et permettent de créer des cartes. Depuis leur introduction et leur application aux images satellites, les réseaux de neurones ont permis une amélioration significative de la qualité de ces cartes, ainsi que leur génération à plus large échelle. En effet, ces modèles permettent d'extraire des informations complexes, sont plus robustes aux changements et plus généralisables que les méthodes conventionnelles d'apprentissage automatique. De plus, elles permettent d'intégrer un volume de données plus important, dans notre cas les données satellites dont la quantité augmente fortement. Cette partie présente quelques méthodes d'adaptation de domaines existantes utilisant l'apprentissage profond et, dans un second temps, d'autres outils qui ont permis la production des résultats de cette thèse. Dans un premier temps, la Section 2.1 pose les bases de l'apprentissage automatique, et la spécificité de l'optimisation d'un réseau de neurones. La Section 2.2 présente l'adaptation de domaine et l'état de l'art des méthodes d'adaptation utilisant des méthodes d'apprentissage profond. Nous pouvons les séparer en deux grandes parties, présentées en Sections 2.2.2 et 2.2.3 :

- La Section 2.2.2 présente les méthodes qui cherchent à adapter les distributions des données d'entrées, pour les envoyer dans un même domaine, avant de les traiter par le modèle.
- La Section 2.2.3 présente les méthodes qui cherchent à modifier les caractéristiques extraites par le modèle, pour que celui-ci puisse extraire les des informations similaires à partir d'images de domaines différents.

Ensuite, la Section 2.3 présente plus précisément le système de classification du sol LCZ et justifie leur usage tout au long de ces travaux. Enfin, la dernière section introduit brièvement les chaînes de Markov.

2.1 Notion d'apprentissage

Comme indiqué dans l'introduction, l'apprentissage permet d'optimiser un modèle pour résoudre une tâche par laquelle ce dernier n'a pas été explicitement programmé à résoudre. Un modèle "apprend" à l'aide d'une fonction de coût.

Fonction de coût. Une fonction de coût est utilisée pour mesurer l'écart entre les sorties prédites et les étiquettes réelles. Son but est d'évaluer les performances d'un modèle d'apprentissage.

Plusieurs types d'apprentissages existent selon la disponibilité de données **étiquetées** : L'apprentissage supervisé (toutes les données sont étiquetées), l'apprentissage semi-supervisé (une partie des données sont étiquetées) et l'apprentissage non supervisé (aucune donnée n'est étiquetée). Tous ces apprentissages utilisent cependant un fonction de coût. Dans l'apprentissage supervisé, nous faisons l'hypothèse que les étiquettes de toutes les données à notre disposition sont disponibles et connues. Les données d'entraînement $X \times Y$ sont donc composées de paires $(\mathbf{x}, \mathbf{y})$ où $\mathbf{x}$ est un échantillon (dans notre cas, une image) et $\mathbf{y}$ est l'étiquette associée décrivant la classe de $\mathbf{x}$. L'objectif de l'apprentissage supervisé est d'apprendre une fonction de transfert F qui relie les entrées $\mathbf{x}$ aux sorties prédites $\mathbf{y}$, c'est-à-dire $\mathbf{y} = \mathbf{F}(\mathbf{x})$. Cet objectif est réalisé en minimisant une fonction de coût J. Une fonction de coût couramment utilisée pour optimiser une réseau de neurones pour une tâche de classification supervisée est l'entropie croisée (*Cross-Entropy*), définie comme suit :

$$CE(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=0}^{C} y_i \log(\hat{y}_i) (= J) \tag{2.1}$$

où $\mathbf{y}$ est le vecteur des vraies étiquettes, $\hat{\mathbf{y}}$ est le vecteur des prédictions du modèle, y_i est l'élément i du vecteur $\mathbf{y}$, $\hat{y}_i$ est l'élément i du vecteur $\hat{\mathbf{y}}$ et C est le nombre de classe. Cette fonction de coût prend des valeurs basses lorsque la prédiction du modèle est similaire à l'étiquette, et inversement si sa prédiction est éloignée. Le modèle est donc, idéalement, le plus performant lorsque sa valeur est faible.

Pour les apprentissages utilisant des données non étiquetées, d'autres fonctions de coût, par exemple le coût constratif qui sera détaillé dans la Section 4.3.2, sont utilisées. Le principe sous-jacent reste néanmoins le même. Cette section présente le processus d'optimisation d'un modèle d'apprentissage à l'aide de cette fonction de coût, puis la particularité de l'entraînement des réseaux de neurones. Enfin, nous discuterons des problèmes induits par ces apprentissages par rapport aux données d'entrée utilisés.

2.1.1 Descente de gradient

L'objectif de l'apprentissage est de minimiser la fonction de coût $J(\mathbf{w})$ en ajustant les paramètres $\mathbf{w}$ du modèle d'apprentissage F. Cela se fait généralement en utilisant des algorithmes d'optimisation, comme la descente de gradient qui est très couramment utilisée en apprentissage automatique :

$$\mathbf{w}_{\text{nouveau}} = \mathbf{w}_{\text{ancien}} - lr \nabla J(\mathbf{w}) \tag{2.2}$$

où lr est le taux d'apprentissage et $\nabla J(\mathbf{w})$ est le gradient de la fonction de coût, par rapport aux paramètres $\mathbf{w}$. Ici, on met à jour les paramètres du modèle itérativement en fonction de la direction donnée par la fonction de coût, en utilisant les données d'entraînement. Cette phase itérative se termine lorsqu'un critère de performance du modèle est atteint, comme une précision en validation.

Cette descente de gradient est utilisée dès lors qu'une valeur de coût est calculée, peut importe si des étiquettes sont disponibles ou non. Dans ce cas, on passe de l'apprentissage supervisé à un apprentissage semi-supervisé ou non supervisé, fréquemment utilisé pour l'adaptation de domaine. Le principe d'apprentissage, à partir de la fonction de coût, reste le même.

2.1.2 Particularité des réseaux de neurones

L'apprentissage profond, plus connu sous le terme anglais *deep learning*, est une branche de l'apprentissage automatique. Le principe d'apprentissage est donc le même, la seule différence étant que les modèles utilisés sont des réseaux de neurones.

Définition 1 *Un **réseau de neurones** est un réseau de noeuds inter-connectés dans une structure à une ou plusieurs couches.*

Le poids de chaque neurone, c'est-à-dire son importance dans le vecteur de sortie du modèle, est optimisé pendant l'apprentissage pour produire un vecteur de sortie permettant de résoudre une technique complexe, comme la vision par ordinateur ou le traitement de langage naturel. Cette sous-section présente plus précisément l'optimisation d'un réseau.

L'entraînement des réseaux de neurones utilise cette descente de gradient, mais adaptée pour leur structure séquentielle. Dans ce cas, la descente de gradient s'effectue par rétropropagation (ou *backpropagation*). Son objectif est de calculer le gradient pour chaque neurone, afin de lui donner un direction pour son optimisation. Dans la rétropropagation, l'erreur de prédiction totale est d'abord calculée (via une fonction de coût) et cette erreur est propagée dans le sens inverse du modèle en utilisant les dérivées locales, dans chaque couche, via la règle de la chaîne. Ce principe peut être formulé de la manière suivante :

$$\frac{\partial J}{\partial w_{ij}} = \frac{\partial J}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_{ij}} \tag{2.3}$$

Dans cette équation, w_{ij} est le poids de la connexion entre le neurone i de la couche précédente et le neurone j de la couche actuelle, z_j est l'entrée pondérée du neurone j et J une fonction de coût. Ainsi, il est possible de calculer des gradients relatifs à chaque neurone pour optimiser le modèle entier. Cette rétropropagation peut être utilisée dès lors qu'une valeur de coût est calculée en sortie du modèle, si des étiquettes sont disponibles, ou non. C'est par ce principe que tous les réseaux de neurones sont entraînés. Bien que ces réseaux aient une transférabilité accrues par rapport aux méthodes conventionnelles, leurs performances baissent significativement lorsqu'il sont confrontés à des données inconnues. Ces performances baissent d'autant plus que les données cibles sont éloignées des données d'entraînement, c'est-à-dire différentes de ce que le modèle connaît. Ce sujet est central pour beaucoup d'applications et en particulier dans notre cas, car nous allons chercher à cartographier l'Afrique subsaharienne. Peu de données étiquetées y sont disponibles, et nous voulons développer des méthodes ne nécessitant pas d'annotations humaines, fastidieuses et sujettes aux erreurs d'interprétations.

2.1.3 Le problème des données

Les deux sous-sections précédentes montrent que les modèles d'apprentissage dépendent fortement des données utilisées pour l'entraînement. La quantité et la qualité de ces données sont essentielles pour obtenir un modèle performant. Si les données d'entraînement ne respectent pas certains critères, le modèle appris ne pourra pas offrir de bonnes performances sur les données de test. Par exemple, ces données doivent être suffisantes en nombre et en diversité, pour ne pas que le modèle apprenne des schémas trop spécifiques et ne puisse pas se généraliser. Lorsque des images satellites sont utilisées, ce problème est illustré par les différences régionales entre les images, ou des différences de capteur. Dans notre cas, nous ne considérons que les différences géographiques. Un modèle qui apprend des caractéristiques sur l'Europe aura du mal à prédire ces mêmes caractéristiques sur l'Afrique, car la morphologie du paysage et la végétation sont très différentes. Dans l'idéal, il faudrait avoir des images étiquetées de l'Afrique subsaharienne afin de résoudre notre tâche, ce qui n'est pas toujours possible étant donné car peut de référence terrain sont disponibles pour cette région (par exemple, So2Sat [13] ne possède que des références pour Le Cap et Nairobi). Cet écart entre les échantillons d'entraînement et les échantillons de test est appelé écart de domaine. Les méthodes d'adaptation de domaine cherchent à réduire cet écart pour que le modèle appris soit aussi performant sur le domaine de test que sur le domaine d'entraînement. Ces méthodes sont présentées dans la section suivante.

2.2 Méthodes d'adaptation de domaine

L'adaptation de domaine est un champ de recherche de l'apprentissage automatique et de l'apprentissage par transfert. L'objectif est de transférer l'apprentissage d'un modèle depuis un domaine source vers un domaine cible. Lorsque l'on utilise des images satellites, l'adaptation peut être faite entre des différents capteurs [14], différentes régions [15] mais aussi différents périodes temporelles [16]. Dans ce manuscrit, nous nous concentrons sur l'adaptation de domaine spatiale. Cette section présente une vue d'ensemble des méthodes développées pour résoudre cette tâche à l'aide d'apprentissage profond. La partie 2.2.1 pose les bases formelles de l'adaptation de domaine. Deux principales familles de méthodes sont explorées pour cette adaptation, regroupées selon l'étape dans le traitement de l'image qu'elles modifient. La section 2.2.2 présente les méthodes qui cherchent à rapprocher les données en amont du modèle. Ensuite, la section 2.2.3 présente les méthodes qui cherchent à modifier les représentations des données dans l'espace latent du modèle. Enfin, la section 2.2.4 discute des familles de méthodes qui peuvent être utilisées pour notre application.

2.2.1 Formalisation

L'adaptation de domaine a pour but de transférer l'apprentissage d'un modèle depuis un domaine source vers un domaine cible, qui a une distribution de probabilités différente que le domaine source.

Soit X l'espace des données d'entrées et Y l'espace des étiquettes de X, et D le jeu de don-

nées regroupant X et Y. L'objectif de l'apprentissage automatique classique est de retrouver la fonction h, permettant de relier un échantillon de X à son étiquette correspondante dans Y. Cette fonction est estimée à partir d'un sous ensemble $S = (x_i, y_i)_{i\in[1,m]} \in (X \times Y)^{(m\times m)}$. La tâche d'adaptation de domaine est similaire mais inclus deux jeux de données différents. Soit $D_S = X_S, Y_S$ et $D_T = X_T, Y_T$ les ensembles de données et étiquettes pour les données sources (S) et cible (T, *Target*). Il s'agit d'optimiser la fonction h, en utilisant des données, annotées ou non, des deux jeux de données D_S et D_T. Plusieurs types d'entraînements sont possibles selon la disponibilité de D_S et D_T :

- Adaptation supervisée : Toutes les données (sources et cibles) sont connues et utilisables pour l'entraînement.
- Adaptation semi-supervisée : Toutes les données sources et leurs étiquettes sont disponibles, mais seulement une partie des données cibles l'est.
- Adaptation non supervisée : Les données sources sont connues, mais les étiquettes des données cibles sont inconnues. L'utilisation d'une information relative à l'image mais qui n'est l'étiquette, comme la date à laquelle l'image a été prise, est suffisamment forte pour considérer l'adaptation comme semi-supervisée.

Les deux derniers types sont les plus adaptés au traitement d'images satellites, car l'étiquetage d'une grande quantité d'images est très coûteuse en temps et en argent. L'adaptation de domaine supervisée sera donc mise de côté pour ce manuscrit. En télédétection, deux principales problématiques se distinguent, selon la provenance du changement de domaine entre les données source et de test : la différence de capteur utilisé, ou la différence de région géographique. Lorsque l'on applique un réseau de neurones à des zones non explorées ou à des images acquises, les performances de cartographie sont, en général, médiocres. Cette difficulté découle du postulat selon lequel les données cibles suivent la même distribution que les données d'entraînement, une hypothèse peu réaliste dans le contexte des images de télédétection, caractérisées par une grande diversité de morphologies urbaines et paysagères à travers le monde, avec des variations saisonnières importantes, par exemple dans les pays d'Afrique subsaharienne. En conceptualisant ce problème en tant que problème d'adaptation de domaine, il devient impératif de considérer ces diversités dans l'apprentissage.

La communauté scientifique a proposé différentes techniques pour résoudre ces problèmes d'adaptation. Il y a principalement deux familles de modèles, selon l'approche choisie pour rapprocher les représentations des données sources et cibles dans le modèle. La première famille cherche à rapprocher les données cibles des données sources, c'est-à-dire les méthodes qui cherchent à modifier les données en amont du modèle. Ensuite, certaines méthodes agissent dans l'espace latent du modèle, en cherchant à ne sélectionner seulement les caractéristiques robustes, ou en minimisant la divergence des représentations sources et cibles.

Les parties de cet état de l'art suivent les deux principales familles de l'adaptation de domaine : la Section 2.2.2 présente les méthodes qui modifient les distributions des données avant le modèle, et la Section 2.2.3 présente les modèles qui modifient les représentations des données après traitement passage dans le modèle.

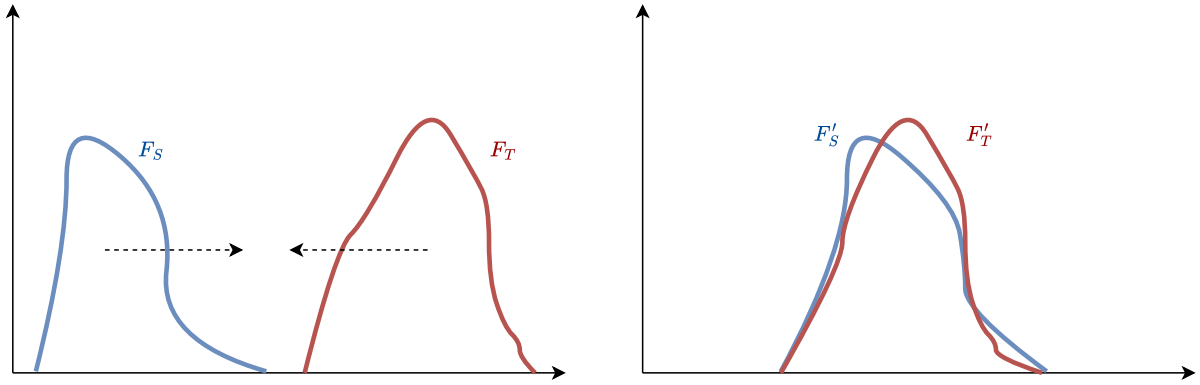


Figure 2.1: Principe d'alignement des distributions source (F_S) et cible (F_T) en adaptation de domaine. A gauche sont les distributions avant adaptation, et à droite après adaptation.

2.2.2 Adapter les distributions des données

"Adapter la distribution" fait référence à des méthodes visant à réduire l'écart entre les domaines source et cible **avant** de les traiter avec le réseau. Elles cherchent donc à modifier les données d'entrées brutes, ce qui peut être réalisé en transformant les images cibles dans le domaine source, en transformant les images sources dans le domaine cible, ou en trouvant une représentation intermédiaire D_I.

Cela peut s'apparenter à des technique de changement de style. Si on considère des peintures, le changement de style serait la modification de la peinture utilisée, de la texture et de l'aspect général de la peinture. Cette sous-section présente quelques travaux tentant de changer les images dans le domaine opposé. Ces techniques peuvent générer de fausses images d'entraînement pour produire de nouvelles données fictives afin d'intégrer les données cibles dans l'entraînement au même titre que les données sources. Des méthodes conventionnelles et l'apprentissage profond peuvent tous les deux résoudre cette tâche.

Yang *et al.* (2020) [17] utilisent une analyse de Fourier pour remplacer les basses fréquences d'une image cible par les fréquences d'une image source. En conséquence, les couleurs et les textures sont modifiées pour correspondre à l'imagerie source, tout en représentant sémantiquement l'imagerie cible. Cependant, cette technique suppose que les images source et cible sont sémantiquement proches, ce qui n'est pas le cas pour les images satellites.

Bien que certaines méthodes utilisent des techniques conventionnelles de traitement d'images pour effectuer ce changement de style, les méthodes récentes utilisent des méthodes d'apprentissage profond pour effectuer ce changement. La plupart utilisent des modèles génératifs, comme les modèles génératifs adversariaux (*Generative Adversarial Networks*, GAN). Les GAN ont révolutionné le domaine de l'apprentissage automatique depuis leur introduction par Goodfellow *et al.* (2014) [18] en permettant la génération de données réalistes. Un GAN est un modèle de deep learning composé de deux réseaux neuronaux qui sont entraînés avec des objectifs opposés. Le premier réseau, appelé générateur, crée des données synthétiques, tandis que le second modèle, appelé discriminateur, essaie de distinguer les données synthétiques des données réelles. Pendant la phase d'entraînement, le générateur essaie donc de générer des images toujours plus réalistes, et on espère tromper le discriminateur.

Avec l'entraînement itératif, les deux réseaux s'améliorent jusqu'à ce que le générateur puisse produire des données indiscernables de celles de l'ensemble de données d'origine. Dans le changement de style, le générateur ne crée pas une image de zéro, mais change le style de l'image en espérant tromper le discriminateur, qui a la même tâche : déterminer si l'image est vraie ou générée. Par exemple, les GAN conditionnels [19] permettent de faire ce transfert. De nombreuses méthodes ont été développées pour la compréhension de scène de rue en milieu urbain. Cette tâche est particulièrement intéressante car liées à de nombreuses application comme le développement d'algorithmes pour voitures autonomes. Il est nécessaire de créer des modèles robustes aux changements, car la voiture doit pouvoir comprendre la scène même dans des situations ou villes jamais vues précédemment. Cependant, il est impossible d'avoir des images ou vidéos de toutes les situations existantes, donc des jeux de données simulées [20], [21] ont été créés pour compléter le jeu de données réelles, Cityscapes [22]. La tâche d'adaptation est alors assimilable à un changement de capteur : *Comment passer d'images issues d'un capteur optique à des images issues d'un capteur simulé ?* Cette conversion de style n'est pas triviale, car les images ne sont pas disponibles dans les deux versions. Lors d'un passage d'un style à un autre, il faut donc pouvoir contrôler la cohérence sémantique entre l'image d'origine et l'image générée [23], mais aussi la cohérence du nouveau style [24]. Ce principe est également utilisé pour des images satellites. Une vue générale est proposée en Figure 2.2. L'idée est ici de conserver les informations sémantiques d'une images, comme le bâti ou les forêts, et d'en modifier leur aspect (couleur et texture), toujours en utilisant un GAN. ColorMapGAN [25] génère de fausses images d'entraînement de cette façon. ColorMapGAN est alimenté avec des images cibles et change sa carte de couleurs en fonction d'une image de l'ensemble de données source. Autrement dit, on récupère les couleurs d'une image source pour les mettre dans une image cible. Le modèle développé par Letheule *et al.* (2023) [26] permet la simulation d'images SAR à partir d'images Sentinel-2, à l'aide d'un GAN conditionnel. Le classificateur est affiné en utilisant ces nouvelles données pour effectuer l'adaptation.

2.2.3 Rapprocher les caractéristiques extraites

Les méthodes présentées dans cette partie agissent plus tard dans le processus d'entraînement. Alors que les méthodes de la partie précédente modifiaient les données en entrées, celles-ci cherchent à modifier les représentations des images dans l'espace latent du modèle, donc après que les images aient été traitées. Ces méthodes peuvent se diviser en trois sous familles. D'une part, il y a les méthodes basées sur le calcul de divergence, et qui essaient d'aligner les représentations des images sources et cibles dans l'espace latent. D'autre part, les méthodes adversariales cherchent aussi à aligner les représentations en utilisant des mécanismes adversariaux sur les caractéristiques extraites. Enfin, les méthodes de la dernière sous-famille extraient les caractéristiques invariantes aux domaines sources et cibles.

Méthodes basées sur la divergence

Ces méthodes visent à réduire la disparité statistique entre les deux domaines en minimisant une mesure de distances. Un schéma global de ce type de méthode est présenté en Figure 2.3. Ces notions de distances ont émergé dans les années 90, avec la démonstration de la perti-

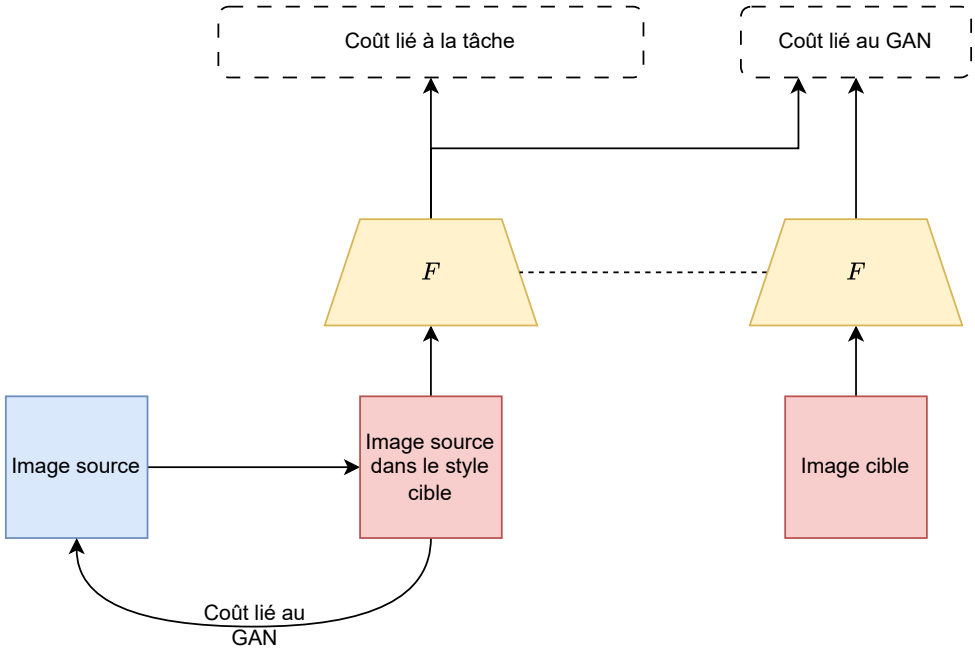


Figure 2.2: Vue d'ensemble d'un modèle utilisant le changement de style pour effectuer l'adaptation de domaine.

nence de la distance de Kullback-Leibler (KL) pour l'adaptation de domaine [27], puis par la formalisation de la *Maximum Mean Discrepancy* (MMD) [28] qui mesure le distance entre deux distributions. L'objectif principal est donc d'aligner les distributions des domaines sources ou cibles, que ce soit les distributions marginales ou conditionnelles. Les distributions marginales sont les distributions spécifiques au domaine que l'on souhaite aligner afin de permettre le transfert d'un modèle entraîné sur D_S vers D_T.

Les distributions conditionnelles sont elles les distributions relatives aux classes à travers les domaine. En plus de vouloir aligner les domaines ensembles, on va vouloir aligner les distributions des classes dans chaque domaine, toujours afin de permettre une meilleure classification. Le calcul de cette divergence peut être intégré à des méthodes de réduction de dimensions classiques, comme l'Analyse de Composantes Principales pour construire une représentation robuste des images [29]. Avec l'avènement de l'apprentissage profond, ces méthodes d'adaptation de domaine basées sur la divergence ont pu être intégrées dans des architectures de réseaux de neurones, plutôt que dans des méthodes conventionnelles. En pratique, lorsque des réseaux de neurones sont utilisés comme extracteur de caractéristiques, l'idée est d'aligner les distributions des représentations latentes cibles et sources à l'aide d'une couche d'adaptation. Long *et al.* (2015) [30] sont les premiers à prouver l'efficacité de alignement des distributions marginales pour l'adaptation de domaine, avec des couches utilisant la MMD. Ces couches ont ensuite été adaptées pour aussi aligner les distributions conditionnelles [31]. Sun & Saenko (2016) [32] adaptent la méthode CORAL [33] pour l'apprentissage profond avec DeepCORAL, qui cherche à minimiser la distance entre les covariances des caractéristiques sources et cibles extraites par le modèle. Zellinger *et al.* (2018) [34] étend la MMD pour l'adapter à des moments d'ordres supérieurs et améliorer l'alignement des caractéristiques. Dans ce processus d'apprentissage, la divergence est calculée après l'extraction

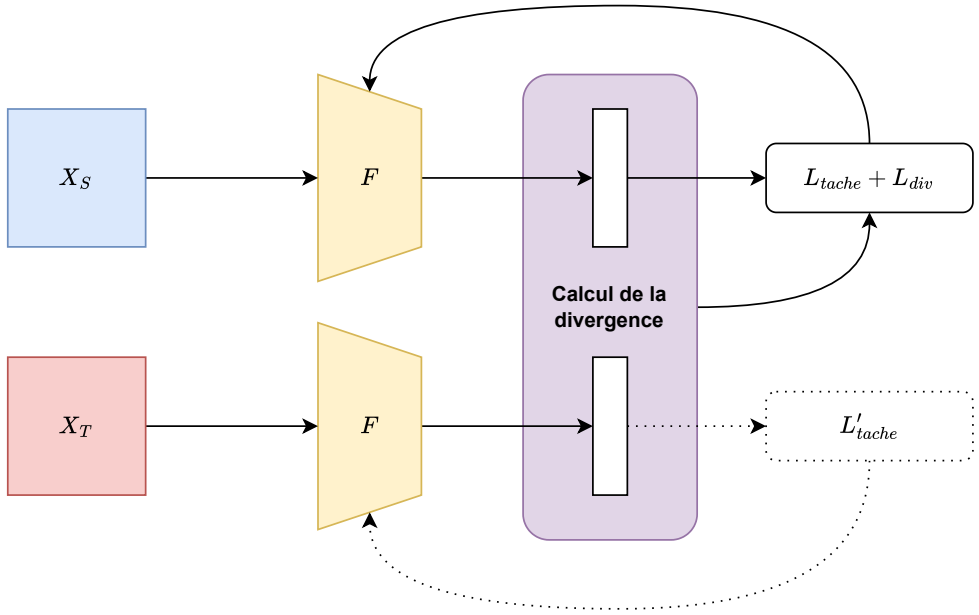


Figure 2.3: Vue d'ensemble d'un modèle utilisant un calcul de divergence. Les parties en pointillé sont les parties qui existent dans certaines méthodes, mais qui ne sont pas obligatoires.

complète des caractéristiques. Zhu *et al.* (2021) [35] ont aussi adapté la MMD pour la rendre locale et l'utiliser à chaque couche d'un réseau de neurones, afin d'aligner chaque sous-domaine et pas uniquement le domaine des caractéristiques extraites. Ce terme de divergence, une fois ajouté à la fonction de coût de classification, agit comme un coefficient de régularisation pour que le modèle ne sur-apprenne pas par rapport au jeu de données d'entraînement (source). Zhu *et al.* (2021) [36] calculent une MMD conditionnelle dans un module d'extraction de caractéristiques multi-niveaux, afin d'aligner chaque domaine à chaque afin d'aligner les distributions conditionnelles à plusieurs niveaux. Les efforts de la communauté scientifique dans ces méthodes utilisant la divergence tendent à chercher des techniques astucieuses pour aligner les distributions dans des domaines de plus en plus précis, là où les représentations des données permettent une meilleures séparations des classes et un meilleur alignement des domaines. Ces efforts permettent de modéliser des relations plus complexes, ce qui était limité par l'utilisation d'une MMD simple. Les méthodes adversariales, présentées dans la partie suivante, permettent ces modélisations plus complexes.

Méthodes adversariales

Une autre manière d'aligner les représentations dans l'espace latent consiste en l'utilisation de réseaux adversariaux. Ces méthodes ne sont pas des GAN, comme dans la section 2.2.2, car n'utilise pas de génération d'images. Le principe de discrimination est cependant conservé. Les vecteurs en entrée du discriminateur en sont pas des images, mais les caractéristiques extraites par le modèle, pour chaque domaine. Les objectifs adversariaux sont donc les suivants: le discriminateur doit différencier les représentations sources des représentations cibles, et le modèle doit produire des caractéristiques similaires pour les deux domaines afin de tromper le

discriminateur. L'utilisation de l'adversarial permet de modéliser des relations plus complexes que les méthodes basées sur la divergence, car ils permettent de modéliser des caractéristiques d'ordres supérieurs. Cette fonction de coût adversariale a ensuite le même rôle que la divergence décrite dans la sous-section précédente : intervenir comme coefficient de régularisation et éviter le sur apprentissage sur les données sources d'entraînement, tout en permettant des modélisations plus complexes des données. Ganin *et al.* (2016) [37] introduisent cet apprentissage pour la première fois en utilisant un extracteur de caractéristiques et deux autres modèles de prédiction. Le premier, appelé le "prédicteur d'étiquette", doit classer l'image. Le deuxième, appelé le "prédicteur de domaine", doit déterminer le domaine d'appartenance de l'image. L'intégration de ce classifieur de domaine dans l'entraînement s'effectue à l'air d'une couche d'inversion du gradient. Pour effectuer l'alignement des domaines pendant l'entraînement, on souhaite que l'extracteur de caractéristiques trompe le classifieur de domaine. Intuitivement, on voudrait que l'entraînement de l'extracteur aille dans le sens opposé du discriminateur. Ainsi pendant la rétro-propagation, le gradient est multiplié par une constante négative pour changer son sens et ainsi change le sens de l'entraînement: lorsque le discriminateur réussit mieux, l'extracteur de caractéristiques est pénalisé. Ce principe a ensuite été développé par la communauté. L'alignement des distributions globales ne permet pas toujours de différencier les classes, notamment les classes similaires, après adaptation. Pour résoudre ce problème, Chen *et al.* (2017) [38] ont appliqué cette partie adversarial pour aligner deux types de distribution. Le premier entraînement adversarial sert, comme pour [37], à aligner les domaines sources et cibles. En plus de l'alignement des domaines globaux, un autre mécanisme adversarial est utilisé pour aligner les distributions des domaines par classes. Cela permet de limiter la confusion entre les classes. Saito *et al.* (2018) [39] utilisent des classifieurs spécifiques à la tâche à résoudre pour détecter les échantillons cibles éloignés du domaine source. Un réseau est ensuite entraîné pour produire des caractéristiques cibles proches du domaine source afin de tromper ces classificateurs, afin d'empêcher la génération de cartes de caractéristiques aberrantes ou proches des frontières des classes. Chen *et al.* (2022) [40] combinent l'alignement en entrée du modèle et dans l'espace des caractéristiques pour l'adaptation de domaine en segmentation sémantique. Le premier module, en amont du modèle, rapproche les données cibles du domaine source en conservant la sémantique de l'image. Le second module, dans l'espace des caractéristiques, regroupe des caractéristiques issues de niveau différent pour éviter que l'alignement ne se fassent à qu'à des hauts niveaux. Les différents niveaux sont traités par un mécanisme d'attention, et la carte de prédiction résultante est utilisée pour le mécanisme adversarial.

Les méthodes adversariales, comme les méthodes basées sur la divergence, cherchent à aligner les distributions. Cela suppose que cet alignement est possible, et que toutes les caractéristiques des domaines peuvent êtres "traduites" dans l'autre domaine. La partie suivante présente des méthodes qui ne reposent pas sur cet hypothèse, mais cherche au contraire à extraire des données les caractéristiques invariantes.

Sélection de caractéristiques robustes avec des prototypes

La famille de modèles cherchant à sélectionner les caractéristiques robustes à la différence entre les deux domaines agissent aussi sur les représentations des images dans l'espace latent. Au lieu d'ajuster les poids du modèle pour que deux images des deux domaines ayant la même

information sémantique aient une représentation similaire dans l'espace latent, ces méthodes visent à identifier et extraire les caractéristiques les plus pertinentes et discriminantes des données. L'idée est d'obtenir, à la fin de l'entraînement, un modèle robuste aux variations des domaines et qui se concentrent sur leurs caractéristiques communes. L'obtention de ces caractéristiques communes peut se faire via l'usage de *prototypes*. Un schéma simple de ce types de méthode est proposé en Figure 2.4 En apprentissage automatique, un prototype est une instance qui représente un groupe de données, appelé catégorie. En adaptation de domaine, ces prototypes sont des représentations de catégories invariantes selon le domaine. L'intuition général est que les données peuvent être représentées grâce à leur exemples encodés dans l'espace latent d'un modèle [41]. Ils sont optimisés pendant l'apprentissage et servent ensuite d'a priori lors du traitement de l'image. Saito *et al.* (2019) [42] introduisent ces prototypes directement dans la couche de classification du modèle, avec un vecteur de "poids". Ces vecteurs de poids doivent orienter la décision du modèle vers la classe la plus probable, et ainsi sont des représentations particulières de ces données. Zhang *et al.* (2021) utilisent du *pseudo-labelling*, l'entraînement du modèle en considérant que la prédiction du modèle est suffisamment bonne pour être l'étiquette. Ce type d'entraînement est très sensibles aux valeurs aberrantes, compliquées à prédire pour un modèle car hors des distributions connues. Ces mauvaises prédictions influent négativement sur un apprentissage par *pseudo-labelling*. Pour résoudre ce problème, ce *pseudo-labelling* est contrôlé par la distance entre les prototypes et l'exemple. Si cette distance est trop grande, l'impact de cet exemple sur l'apprentissage est diminué pour conserver de la stabilité dans l'entraînement. Ce concept de prototype à été utilisé conjointement avec d'autres techniques d'apprentissage automatique, comme l'apprentissage auto-supervisé [43] ou l'apprentissage contrastif [44].

Ces prototypes ont aussi été utilisé pour l'adaptation de domaine en télédétection. Pour améliorer leur représentation, Gao *et al.* (2023) proposent des "multi-prototypes" à plusieurs dimensions pour modéliser des données plus complexes. Zhu *et al.* (2023) [45] introduisent un modèle à deux chemins, un adversarial et l'autre à base de prototypes pour aligner à la fois les distributions et obtenir des représentations robustes aux domaines. Le chemin adversarial est classique, comme décrit dans la sous-section précédente. L'autre chemin utilise les prototypes comme a priori à comparer avec les caractéristiques extraites du modèle à l'aide de mécanismes d'attention. Ces prototypes sont ensuite optimisés à l'aide de *pseudo-labelling*.

2.2.4 Discussion

Cette partie avait pour objectif de présenter quelques méthodes permettant de faire de l'adaptation de domaine pour des tâches de vision par ordinateur. Lorsque l'on considère des images satellites, ce problème se complexifie. Les images satellites ont une seconde particularité : la variance intra-classes est plus grande que pour les images naturelles, notamment à causes des environnements différents. Cela a plusieurs implications concernant les méthodes d'adaptation de domaines qui pourront être utilisées dans la thèse. Les méthodes adaptées à l'adaptation de domaine d'images satellites se concentrent sur des problèmes avec une adaptation faible, comme entre villes ou régions. Tasar *et al.* (2020) [25] ont réalisé une adaptation de domaine en utilisant quatre villes européennes, tandis que Zhu *et al.* (2023) [45] ont utilisé les ensembles de données ISPRS Vaihingen et Potsdam. Ces stratégies ne sont pas adaptées à

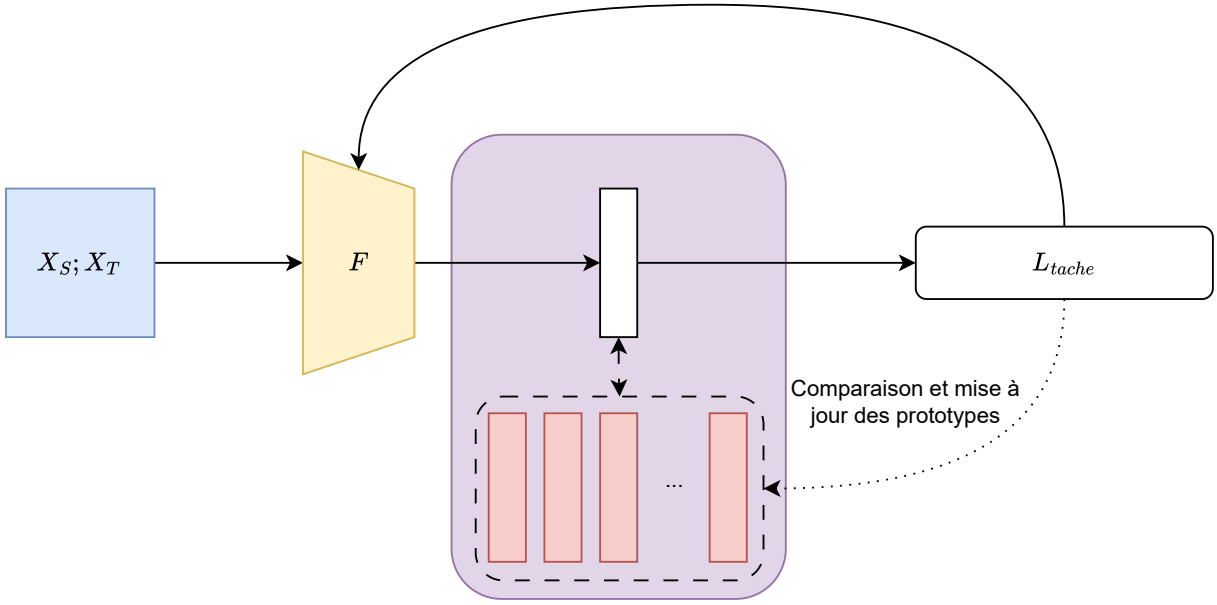


Figure 2.4: Vue d'ensemble d'un modèle utilisant les prototypes. Les caractéristiques extraites sont comparées à des vecteurs qui sont optimisés pour représenter les différentes classes.

la cartographie de la couverture terrestre à l'échelle d'un pays, car elles ne prennent pas en compte les complexités environnementales d'un pays entier.

Cette discussion doit donc nous permettre de répondre à la question suivante : quelle famille de méthode est la plus adaptée pour l'adaptation des données environnementales depuis les bases de données existantes (i.e. principalement sur l'Europe, l'Asie et l'Amérique) vers le contexte africain ?

Bien que les GAN aient montré des résultats prometteurs à l'aide de méthodes de changement de style, leur utilisation n'est pas adaptée lors de grand décalage d'environnement. En suivant ces méthodes, il faudrait pouvoir changer le style d'une ville africaine vers un style par exemple européen, ce qui est un écart très important. Non seulement les textures et les couleurs sont très différentes, mais aussi la structure et l'organisation des villes. Pour cette raison, les GAN ne seront pas utilisés dans ces travaux.

Comme il est complexe de modifier le style des images dans notre cas, il semble plus adéquat de modifier l'espace latent des modèles pour avoir de meilleures représentations. Les LCZ ont été définies par rapport à des critères physiques, définis en section 2.3.1, construites pour être indépendantes du contexte du pays. Ainsi, ces descripteurs ne doivent pas dépendre des caractéristiques de style des images, et doivent être une représentation plus globale. La définition physique des LCZ peut être vu comme des prototypes de chacune des classes, c'est-à-dire comme des vecteurs de représentation des données et qui sont invariants au domaine.

2.3 Local Climate Zones

Les LCZ sont une classification des zones urbaines basée sur les caractéristiques climatiques et physiques locales, afin de la rendre culturellement indépendant. Elles ont été introduites par Stewart & Oke [46] d'abord pour l'étude des îlots de chaleurs, mais ont rapidement été utilisées pour d'autres applications, comme la gestion de l'énergie [47], du climat [48] ou encore les géosciences [49]. Cette section présente d'abord la construction de ces LCZ et justifie leur usage dans cette thèse. Enfin, nous retracerons un historique de la cartographie des LCZ depuis leur introduction.

2.3.1 Construction des Local Climate Zones

Les LCZ sont définies comme une région délimitée avec des caractéristiques uniformes telles que la densité du bâti, la présence de végétation, la rugosité du terrain ou l'albédo, qui influencent directement le climat local. Ces caractéristiques ne sont pas déterminées par des humains lors de la caractérisation mais par des critères physiques pré-déterminés. Ces valeurs sont données dans le Tableau 2.1. Dans ce Tableau, les critères physiques sont:

- **Sky View Factor** - **Facteur de vue du ciel** : Proportion de ciel visible depuis le sol, influençant le rayonnement thermique.
- **Aspect Ratio** - **Rapport d'aspect (H/L)** : Rapport entre la hauteur moyenne des bâtiments (H) et la largeur des rues (L), influençant l'ombre et la ventilation.
- **Building Surface Fraction** - **Fraction de surface bâtie** : Pourcentage de la surface occupée par les bâtiments, influençant l'imperméabilité et l'absorption de chaleur.
- **Impervious Surface Fraction** - **Fraction de surface imperméable** : Pourcentage de la surface recouverte de matériaux imperméables, comme l'asphalte ou le béton.
- **Pervious Surface Fraction** - **Fraction de surface perméable** : Pourcentage de la surface recouverte de matériaux perméables, comme les sols naturels ou la végétation.
- **Height of Roughness Elements** - **Hauteur des éléments rugueux** : Hauteur moyenne des objets sur une surface (bâtiments, arbres) qui influence la rugosité et l'écoulement de l'air.
- **Terrain Roughness Class** - **Classe de rugosité du terrain** : Classification de la rugosité du terrain, influençant la résistance au flux d'air et la turbulence selon Davenport *et al.* (2000) [50]
- **Surface Admittance** - **Admittance thermique de surface** : Mesure de la capacité d'une surface à absorber et relâcher la chaleur, influençant la variation de température au cours de la journée.
- **Surface Albedo** - **Albédo de surface** : Fraction de la lumière solaire réfléchie par la surface, influençant la température locale.

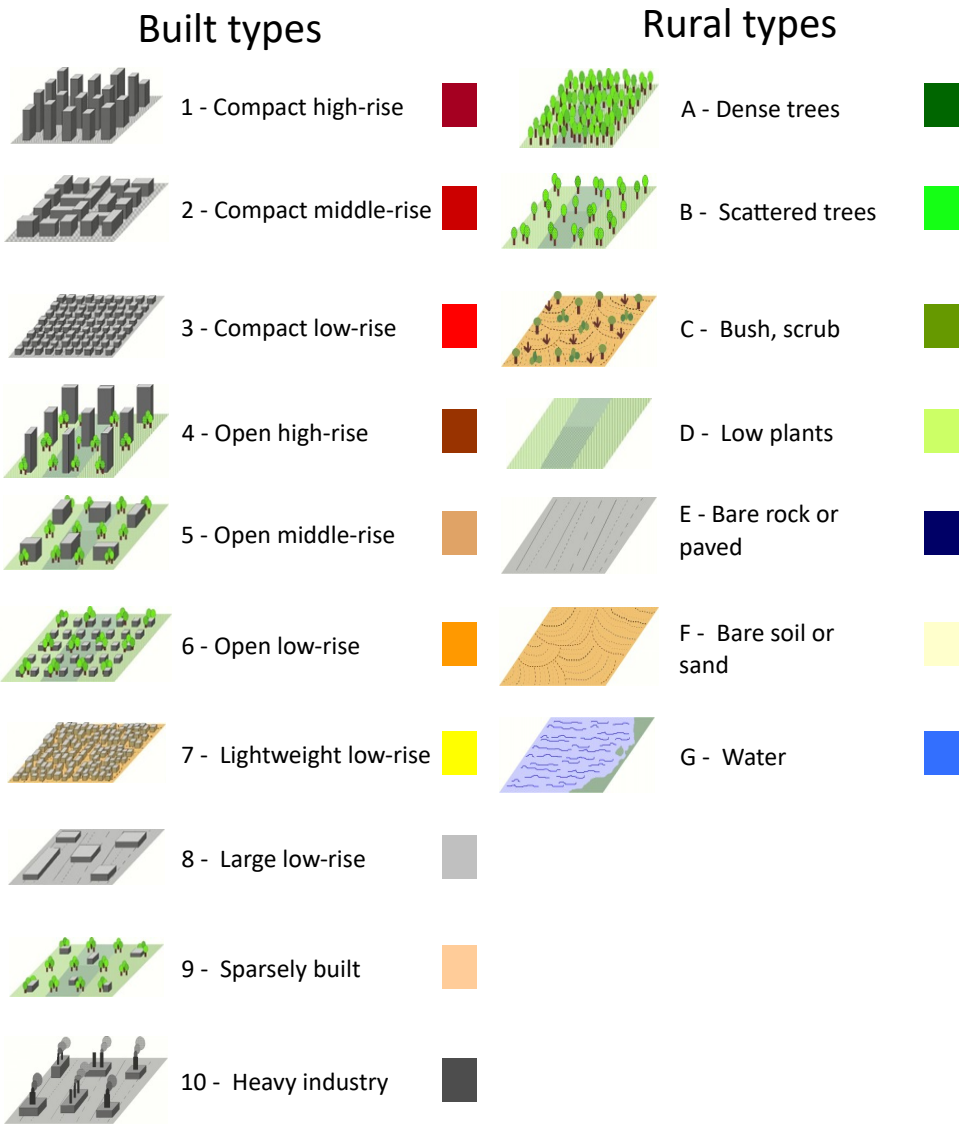


Figure 2.5: Visualisation et code couleur des classes LCZ. Les illustrations sont issues de Stewart & Oke [46].

- **Anthropogenic Heat Output - Émission de chaleur anthropique** : Quantité de chaleur produite par les activités humaines, comme le chauffage, la climatisation et les transports, influençant les températures locales.

Une visualisation des classes est donnée en Figure 2.5.

2.3.2 Justification de leur usage

L'utilisation des LCZ pour caractériser l'environnement pour l'analyse de données issues d'enquêtes démographiques se justifie par plusieurs raisons. Dans un premier temps, les LCZ fournissent une représentation détaillée et fine de l'environnement tant dans les zones urbaines que dans les zones rurales, ce qui leur donne un avantage conséquent par rapport à

LCZ	Sky view fact.	Aspect ratio	Building fract.	Imperious fract.	Pervious fract.	Height roughness	Terrain roughness	Surface admittance	Surface albedo	Anthropogenic heat ouput
1	0.2-0.4	>2	40-60	40-60	<10	>25	8 1,500–1,800	0.10–0.20	50–300	
2	0.3-0.6	0.75-2	40-70	30-50	<20	10-25	6-7	1,500–2,200	0.10–0.20	<75
3	0.2-0.6	0.75-1.5	40-70	20-50	<30	3-10	6	1,200–2,800	0.10–0.20	<75
4	0.5-0.7	0.75-1.25	20-40	30-40	30-40	>25	7-8	1,400–1,800	0.12–0.25	<50
5	0.5-08	0.3-0.75	20-40	30-50	20-40	10-25	5-6	1,400–2,000	0.12–0.25	<25
6	0.6-0.9	0.3-0.75	20-40	20-50	30-60	3-10	5-6	1,200–1,800	0.12–0.25	<25
7	0.2-0.5	1-2	60-90	<20	<30	2-4	4-5	800–1,500	0.15–0.35	<35
8	>0.7	0.1-0.3	30-50	40-50	<20	3-10	5	1,200–1,800	0.15–0.25	<50
9	>0.8	0.1-0.25	10-20	<20	60-80	3-10	5-6	1,000–1,800	0.12–0.25	<10
10	0.6-0.9	0.2-05	20-30	20-40	40-50	5-15	5-6	1,000–2,500	0.12–0.20	>300
A	<0.4	>1	<10	<10	>90	3-30	8	inconnu	0.10–0.20	0
B	0.5-0.8	0.25-0.75	<10	<10	>90	3-15	5-6	1,000–1,800	0.15–0.25	0
C	0.7-0.9	0.25-1.0	<10	<10	>90	<2	4-5	700–1,500	0.15–0.30	0
D	>0.9	<0.1	<10	<10	>90	<1	3-4	1,200–1,600	0.15–0.25	0
E	>0.9	<0.1	<10	>90	<10	<0.25	1-2	1,200–2,500	0.15–0.30	0
F	>0.9	<0.1	<10	<10	90	<0.25	1-2	600–1,400	0.20–0.35	0
G	>0.9	<0.1	<10	<10	>90		1	1,500	0.02–0.10	0

Tableau 2.1: Intervalles des descripteurs pour chacune des classes LCZ.

d'autres systèmes n'utilisant que le terme "zones urbaines". L'utilisation de ce second type de caractérisation entraînerait des omissions de différences importantes entre les ménages. Cette précision, en particulier dans les milieux urbains, est cruciale dans un contexte où des variations subtiles de l'environnement peuvent avoir un impact significatif sur la santé, le bien-être et les comportements des individus. Elle intègre d'ailleurs plus de précisions sur la qualité des milieux de vie des ménages sondés dans les enquêtes.

De plus, les LCZ offrent une approche standardisée, reproductible et indépendant des contextes/cultures des pays pour la classification des zones urbaines, ce qui facilite la comparaison entre différentes études et régions géographiques. Ce système de classification est donc objectif car ne dépend pas d'une décision humaine. Il permet d'avoir la représentation de l'environnement la plus globale et non biaisée possible.

Pour ces raisons, les LCZ nous semble les plus appropriées pour décrire l'environnement dans le contexte d'une analyse de données de population.

2.3.3 Cartographie des Local Climate Zones

Cette section propose un historique des travaux sur la cartographie des LCZ. Depuis leur introduction en 2012, de nombreuses équipes de recherche se sont penchées sur cette tâche. Deux types d'études peuvent être différenciées : les études à petite échelle (dans le cas d'une ville par exemple) et les études à plus grande échelle (au niveau de plusieurs villes ou d'un pays). Différentes méthodes de cartographies ont été utilisées pour ces deux types d'études.

Pour les études "sur site", deux méthodes sont possibles. La première est issue de l'article fondateur des LCZ [46] et consiste en 3 étapes : récupérer les méta-données de l'image en question, définir la source thermale, et classifier les zones en fonction des caractéristiques urbaines et naturelles. La deuxième méthode, consiste en une évaluation visuelle de la zone par des experts pour définir la classe LCZ correspondante. Cette méthode requiert de l'expertise en télédétection et une très bonne connaissance du terrain afin de produire des cartes avec une précision acceptable pour l'application. En outre, des problèmes de reproductibilité des résultats se posent. Bien que ces méthodes aient beaucoup été utilisées pour la cartographie des LCZ [51]–[53], elles ne permettent que des productions à petite échelle, pour les études sur site. Des méthodes utilisant la télédétection ont été développées pour effectuer une cartographie à plus large échelle et de manière systématique.

Les méthodes utilisant les images satellites se sont particulièrement développées après l'introduction du projet WUDAPT [54], cumulé avec l'explosion du nombre de données satellites disponibles. Ce projet a pour objectif de créer un processus de cartographie unique et reproductible pour la cartographie des LCZ. Ce processus comprend plusieurs étapes. D'abord, la sélection de zones d'entraînement LCZ à partir de Google Earth sur la base de connaissances d'experts de la télédétection. Ces zones sont ensuite associées à des images Landsat pour être traité par une forêt aléatoire, jusqu'à l'obtention d'une carte d'une précision acceptable. Ce processus permettrait d'obtenir des cartes avec une précision supérieur à 50% [55]. Une plateforme utilisant Google Earth Engine a été développée pour faciliter la génération de ces cartes [56]. L'utilisation de toutes les zones d'entraînements définies dans cette plateforme et de forêtes aléatoires entraînées à partir de caractéristiques dérivées d'images satellites ont permis la génération de carte LCZ à l'échelle des Etats-Unis d'Amérique [57], de l'Europe [58] et de toute la surface terrestre [59].

En parallèle de ces travaux, Rosentreter *et al.* (2018) [60] sont les premiers à utiliser les réseaux de neurones à convolutions pour cartographier des villes allemandes. Qiu *et al.* (2018) [61] utilisent aussi un réseau de neurone pour montrer l'intérêt des images Sentinel-2 et Landsat pour la cartographie des LCZ. L'intérêt de l'utilisation de données temporelles, notamment saisonnières sur une année, est aussi démontré via l'utilisation de modèles à convolution et récurrent [62]. Le jeu de données So2Sat [13] a été introduit pour encourager cet effort nouveau sur l'exploitation de réseaux de neurones pour la cartographie des LCZ. Il est composé d'images Sentinel-1 et Sentinel-2 de 42 zones urbaines autour du globe. Dix villes, provenant de 10 zones culturelles ont ensuite été rajoutées pour une meilleure évaluation des modèles. Le jeu de données est disponible en plusieurs versions avec différents fractionnement :

- "*Random*" : Les fractionnements d'entraînement et de tests sont construits en tirant les images aléatoirement.
- "*Block*" : Les villes sont séparées en deux parties (pour l'entraînement et la validation) en fonction de leurs positions spatiales. On peut parler ici d'une légère adaptation de domaine.
- "*Cultural_10*" : Les 42 deux villes d'entraînement de So2Sat sont conservées pour l'entraînement, et 10 autres villes provenant de 10 zones culturelles différentes sont ajoutées pour le test. Cette version propose une adaptation de domaine plus complexe.

En particulier, So2Sat a été utilisé pour générer des cartes LCZ de 1692 villes autour du monde [63]. Bien que l'apprentissage ait été fait sur ce jeu de données global, la précision des cartes n'est pas toujours visuellement correcte. En particulier, les villes qui ne sont pas proches des régions d'entraînements ne sont pas correctement cartographiées. Du point de vue d'un problème d'adaptation de domaine, ces villes cibles appartiennent au domaine cible D_T qui est très différent des villes d'entraînement du domaine source D_S. Pour combler cet écart de domaine, un modèle ensembliste a notamment été utilisé par Zhao *et al.* (2023) [15]. Il repose sur cinq ResNet [64] entraînés en parallèle sur les mêmes données d'entraînements issues des données cibles et des données sources. A chaque itération de l'entraînement, chaque réseau reçoit un *batch* **différent** d'images sources et le **même** *batch* cible. Une fonction de coût classique est calculée sur les images sources. Pour les images cibles, une pseudo-étiquette est

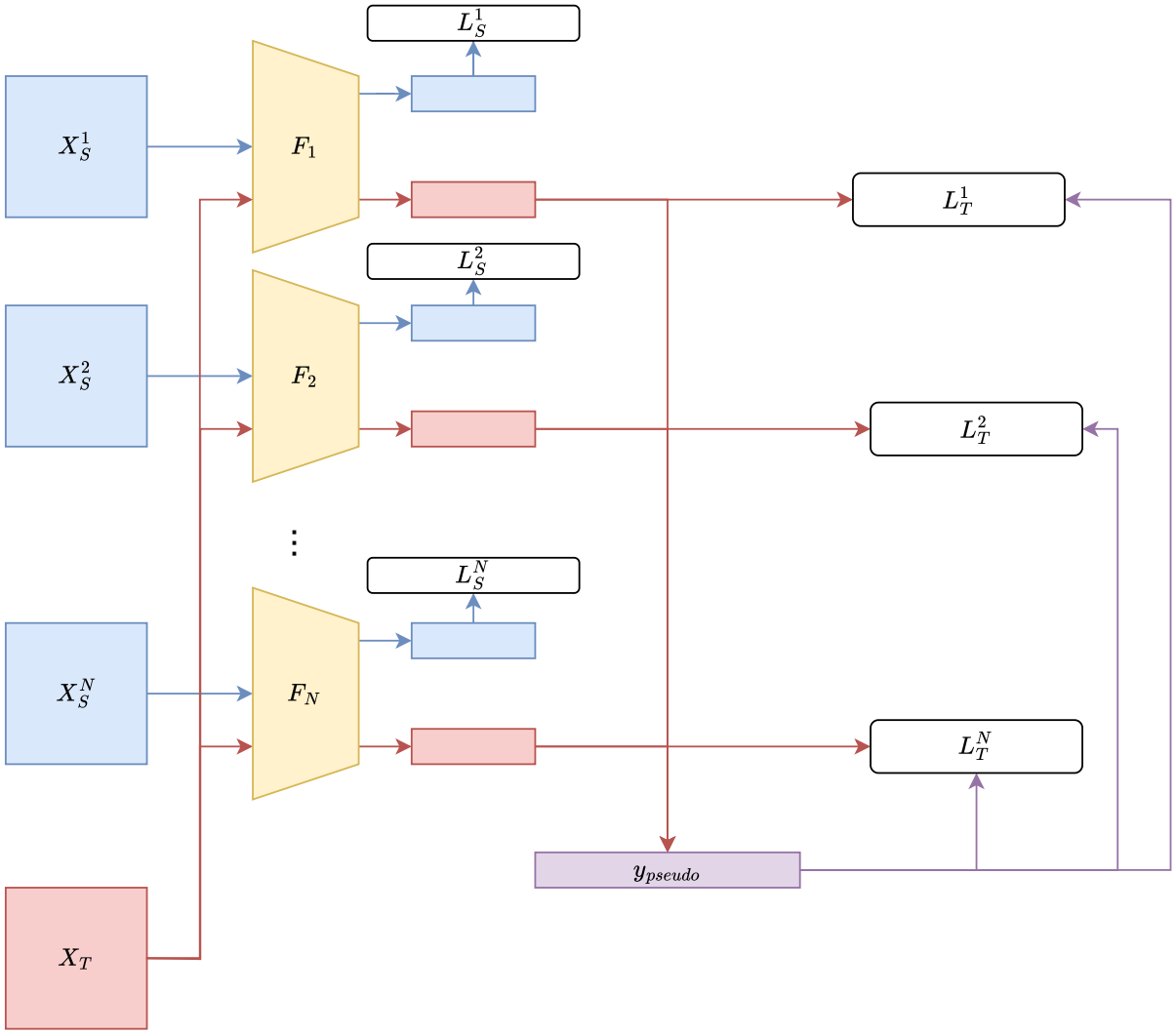


Figure 2.6: Vue d'ensemble du modèle ensembliste introduit par Zhao *et al.* (2023) [15]. N représente le nombre de modèles. y_{pseudo} est la pseudo étiquette pour l'image cible X_T calculée à partir de la moyenne des prédictions des N modèles. Les coûts sont ensuite combinés dans chaque chemin pour la rétropropagation de chaque réseau.

calculée à partir de la moyenne des résultats des prédictions des cinq modèles. L'hypothèse implicite est que la prédiction des cinq modèles est suffisamment robuste pour permettre une inclusion correcte des caractéristiques cibles dans l'entraînement. Une visualisation du modèle est donnée en Figure 2.6. Les modèles pour l'adaptation de domaine seront évalués sur le version "Cultural 10".

2.4 Courte introduction aux chaines de Markov

Les chaînes de Markov sont un concept fondamental en théorie des probabilités et en statistiques. Elles sont largement utilisées dans de nombreux domaines depuis leur introduction par Andreï Markov en 1906, comme en informatique ou en économie. Une chaîne de Markov

est un modèle stochastique décrivant une séquence de transitions d'un état à un autre, suivant la propriété faible de Markov. Cette propriété est aussi appelée propriété de mémoire à court terme. Cette formalisation de la transition d'état leur permet d'être intégré dans des systèmes permettant une pris een compte du temps. Les chaînes de Markov seront utilisées dans le chapitre 4, pour effectuer une régularisation temporelle de cartes environnementales générées et améliorer la qualité de la carte finale. Une chaîne peut être formulée de la manière suivante :

Propriété 1 *Soit $(M_i)_{i\in\mathbb{N}}$ une séquence de variables aléatoires représentant les états successifs d'un processus stochastique. Cette séquence satisfait la propriété de Markov si, pour tout $n \geq 0$ et tout état x_i,*

$$P(M_{n+1} = M_{n+1}|M_0 = m_0, M_1 = m_1, \ldots, M_n = m_n) = P(M_{n+1} = m_{n+1}|M_n = m_n)$$

Cela signifie que la probabilité de passer à l'état M_{n+1} au temps $n + 1$, est uniquement conditionnée par l'état actuel m_n et ne dépend pas de l'historique complet du processus.

Les probabilités de passage d'un état à un autre dans un processus de Markov peuvent être écrit sous forme de transition. Dans cette matrice, le coefficient à la $j - ieme$ ligne et à la $c - ieme$ colonne représente la probabilité de passer de l'état j à l'état c.

Propriété 2 *La matrice de transition d'une chaîne de Markov est stochastique. Soient $E = [1, N], N \in \mathbb{N}$ et $P = [p_{j,c}]_{l,c\in E^2}$ une matrice de transition d'une chaîne de Markov à N états. Alors P est stochastique :*

$\sum_{c=1}^{N} p_{jc} = 1$

Lorsque l'on étudie une chaine de Markov, la matrice de transition est en général définie et fixée. Pour résumer, les éléments clés d'une chaîne de Markov sont :

1. **Les états** : Ce sont les différentes conditions ou situations possibles du système que vous modélisez. Par exemple, dans un modèle météorologique, les états peuvent être "ensoleillé" (A), "nuageux" (B), "pluvieux" (C), "orageux" (D). Un graphe illustre cette chaîne de Markov en Figure 2.7.

2. **La matrice de transition** : Cette matrice décrit les probabilités de transition d'un état à un autre. Chaque élément de la matrice représente la probabilité de passer d'un état initial à un état final. Dans notre exemple, elle eut être définie de la manière suivante:

   $$P = \begin{pmatrix} 0.0 & 0.7 & 0.3 & 0.0 \\ 0.0 & 0.9 & 0.0 & 0.1 \\ 0.5 & \mathbf{0.5} & 0.0 & 0.0 \\ 0.25 & 0.1 & 0.15 & 0.5 \end{pmatrix}$$

   Dans ce modèle, la probabilité de passer d'un état pluvieux (état C) à un état nuageux (état B) est de **0.5**.

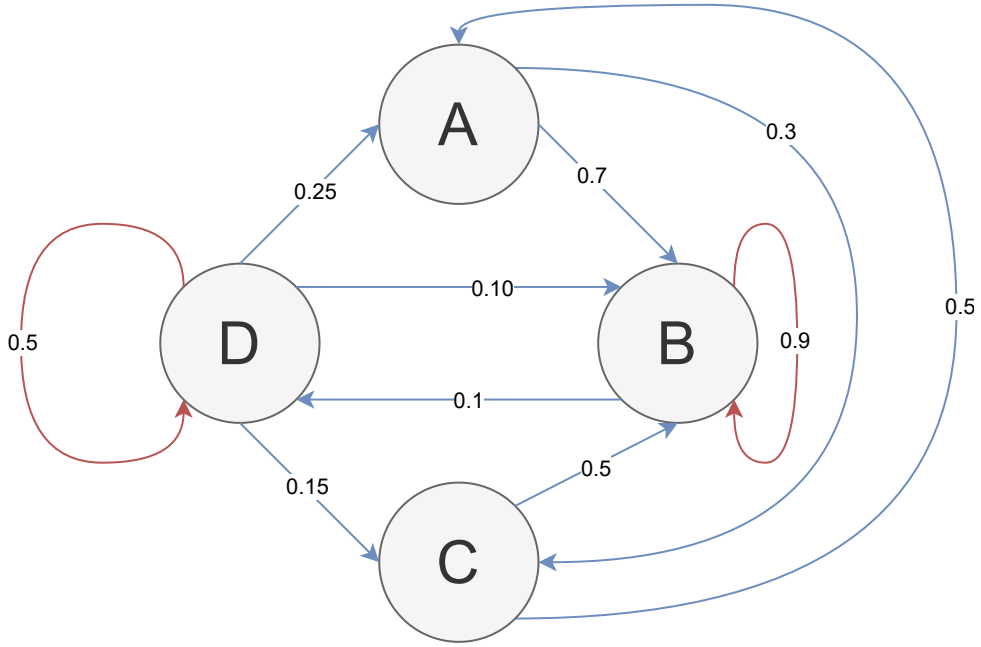


Figure 2.7: Schéma d'une chaîne de Markov à quatre états. Les liens rouges sont les liens qui bouclent sur un état, et les liens bleus permettent le changement d'état.

3. Les transitions d'état sont **discrètes**. À chaque étape, le système se déplace d'un état à un autre en suivant les probabilités de transition définies dans la matrice de transition P.

2.5 Environnement et démographie

Cette section aborde le volet démographique de cette thèse. L'augmentation des sources de données environnementales, notamment les données satellites, permet à la communauté des démographes d'explorer les relations entre les caractéristiques climatiques et environnementales et certains phénomènes de population. Ces liens sont établis par la corrélation entre les données environnementales et les données démographiques, en tenant compte des caractéristiques socio-économiques des individus étudiés afin d'assurer l'indépendance des résultats.

Les indicateurs sont généralement utilisés comme des intermédiaires pour intégrer d'autres phénomènes, directement liés aux populations, dans l'analyse démographique. Thiede *et al.* (2020) [65], Nicholas *et al.* (2021) [66] et Grace *et al.* [5] étudient la croissance des enfants via l'indicateur de la taille en fonction de l'âge (HAZ, *Height-for-Age Z-score*), en relation avec certaines caractéristiques environnementales. Thiede *et al.* (2020) [65] analysent les effets des températures et des précipitations sur la santé des enfants en Indonésie, révélant que des retards dans le début de la mousson sont systématiquement associés à une détérioration de la santé des enfants. En particulier, les retards dans la mousson pendant la période prénatale sont corrélés à une réduction du HAZ des enfants âgés de 2 à 4 ans, tandis que le poids des enfants de moins de 2 ans est négativement affecté par des retards dans la mousson la plus récente, en particulier à Java. Nicholas *et al.* (2021) [66] montrent que des excès de précipi-

tations durant la période prénatale sont liés à une réduction du HAZ chez les enfants ruraux dans leurs premières années de vie. Ces observations ont été faites à partir des données de la base de données *Climatic Research Unit-time series*, qui indique les températures et précipitations moyennes par mois. Grace *et al.* [5] examinent la relation entre les caractéristiques des paysages, telles que la profondeur des points d'eau, et les résultats de santé des enfants dans les populations pastorales et agro-pastorales du Sahel. Leur étude, utilisant des données satellites et des données d'enquête sanitaire, révèle que la profondeur des points d'eau proches a un impact significatif sur le score HAZ des enfants. Ce lien est modulé par les pratiques de subsistance et la source d'eau potable des ménages, illustrant la vulnérabilité des enfants dans ces zones face aux changements climatiques.

L'environnement et le climat influencent également la fécondité. Macfarlan *et al.* (2021) [67] analysent la saisonnalité des naissances sur la péninsule de Basse-Californie, un désert aride marqué par des fluctuations climatiques saisonnières liées à la mousson nord-américaine. Les auteurs constatent que la disponibilité énergétique locale, observée via le NDVI, joue un rôle central dans la saisonnalité des conceptions, les naissances étant planifiées pour coïncider avec la période de "verdissement" saisonnier. Cela suggère une connaissance des populations de leur environnement visant à améliorer la santé et le bien-être des nouveau-nés.

En démographie, les indicateurs de santé des enfants sont souvent utilisés car ces populations sont très sensibles aux conditions de vie, telles que les conditions socio-économiques ou environnementales. Cependant, les données environnementales peuvent également soutenir l'analyse démographique pour d'autres populations. Par exemple, Sikarwar *et al.* (2023) [68] montrent qu'une plus grande exposition à la verdure, mesurée par le NDVI, est associée à une réduction significative du risque de décès liés à la COVID-19 au niveau des districts en Inde. Plus précisément, les districts avec des niveaux de verdure plus élevés présentent une réduction du risque de décès de 32% à 47% par rapport aux districts les moins verts. Cette association reste robuste même après ajustement pour des facteurs tels que la pollution de l'air, la densité de population et les conditions socio-économiques. Les données liées aux maladies vectorielles peuvent aussi être analysées avec des données environnementales. Buczak *et al.* (2012) [69] développent un modèle prédictif de la dengue en combinant des données satellites et météorologiques à haute résolution. L'incidence antérieure de la dengue, les indices de végétation (NDVI, *Enhanced Vegetation Index*, EVI), les précipitations, et d'autres variables climatiques sont essentiels pour prédire les épidémies. En ajustant ces variables à une échelle spatiotemporelle précise, leur méthode permet de faire des prévisions efficaces des flambées de dengue, mettant en évidence l'importance des données satellites pour la surveillance et la prévention des maladies.

Cette section a exploré l'impact des données environnementales, en particulier les données satellites, sur l'analyse démographique des phénomènes liés à la santé et à la fécondité. Les études examinées montrent que les caractéristiques climatiques et environnementales jouent un rôle significatif dans les indicateurs de santé et dans les tendances de fécondité, et illustrent comment ces outils peuvent améliorer les prévisions et la gestion des épidémies de maladies vectorielles. L'intégration des données environnementales, en particulier celles fournies par les satellites, enrichit l'analyse démographique en permettant une évaluation plus précise des impacts environnementaux sur la santé et les tendances démographiques. Ces approches reposent sur trois piliers : des données démographiques géolocalisées, des données satellites et

un contexte socio-économique pour contrôler des effets d'interaction socio-économiques dans le lien entre population et environnement. Les travaux présentés dans ce manuscrit doivent avoir ces trois éléments avec des résultats significatifs.

Chapitre 3

A l'échelle locale

Oh mince, pourquoi ils ont tué le chien ? :(

– Julie Maestri

3.1 Introduction

Ce second chapitre se concentre sur une application à l'échelle locale, afin d'évaluer non seulement l'apport de la télédétection pour la démographie mais aussi discerner les premières limites de son usage. Les données démographiques recueillies à l'échelle locale se déroulent sur des régions restreintes, comme par exemple une ville ou une agglomération. La caractérisation de l'environnement doit être fine, car elle doit permettre de percevoir les différences entre zones voisines. Le capteur et le système de classification utilisés doivent être choisis pour respecter cette contrainte spatiale et mener à bien l'analyse.

Dans un premier temps, il faut sélectionner des données provenant d'un capteur suffisamment bien résolu. Par exemple, le capteur MODIS[1], avec une résolution d'un kilomètre, ne permet pas une cartographie diversifiée à l'échelle d'une ville. A l'inverse, les images Sentinel-2 sont suffisamment résolues pour ce type d'études (10 - 60 mètres). Dans certains cas où la diversité environnementale des villes est forte, des images très résolues, comme les données PlanetScope, doivent être employées. Cependant, ces images ne sont pas disponibles en accès libre, donc nous utiliserons les images Sentinel-2, disponibles en accès libre et à grande échelle.

Dans un second temps, Il faut choisir un système de classification du sol. Tous les systèmes de classification ne sont pas adaptés à un usage local, car ne décrivent pas l'environnement précisément. Cette imprécision se traduit souvent par un nombre de classes trop faible pour caractériser des zones qui ont des types d'environnement proches. Prenons l'exemple des bases de données internationales et des cartes globales générées à partir d'images satellites. Le projet Worldcover de l'ESA [70], [71] a pour objectif de cartographier la surface du globe avec une résolution de 10 mètres, à partir d'images Sentinel-2. Le système de classification du sol est composé de 11 classes, dont 10 sont rurales et une seule urbaine. Cette dernière, dénommée "zone construite" (en anglais *Built-up*), n'indique que les zones où il y a des infrastructures ou habitations construites, sans donner d'information sur leur nature (industrielle, habitations, etc..), densité ou morphologie. Bien que suffisante pour des études à grande échelle, de cette classification à l'échelle locale est limitée. De la même manière, le système utilisé dans la base de données SEN12MS [11] n'utilise que la classe *Urban and Built-Up Lands*, laissant penser que toutes les zones construites sont considérées comme urbaines. Le Global Human Settlement Layer (GHSL) [10] quant à lui indique la présence humaine notamment par une estimation de la densité de population et de bâtis. Bien que cette densité de bâtis permet d'avoir une caractérisation plus précise des zones urbaines, elle ne fournit pas toujours assez d'information critique, en particulier lorsque l'on traite de données de santé. En effet, la hauteur des bâtiments, leurs matériaux et la morphologie de la zone méritent d'être pris en compte [72]–[74]. Les LCZ sont donc adaptées pour une étude dans les milieux urbains, car proposent un cadre plus global que les indicateurs présentés ci-dessus. En particulier, les 10 classes urbaines incluent les caractéristiques de matériaux et de morphologie. Cela permet d'avoir une cartographie de l'environnement des villes plus précise, et a fortiori une distinction plus robuste des environnements des quartiers. D'autres indicateurs plus classiques, comme le NDVI, les modèles d'altitude ou une détection du bâti, peuvent aussi être utilisés pour compléter ce système.

Ce chapitre porte sur l'utilisation de la télédétection pour générer une cartographie fine

[1] https://modis.gsfc.nasa.gov/

de l'environnement, et estimer son impact sur la mortalité - toutes causes confondues et pour certaines causes de décès spécifiques - à Antananarivo, la capitale de Madagascar. La section suivante de ce chapitre présente la ville d'Antananarivo, son contexte environnemental, ainsi que les données de mortalité.

La section 3.3 présente la méthode de cartographie des LCZ, à l'échelle locale. Cette méthode est fondée sur les définitions physiques des LCZ ainsi que sur la base de données So2Sat [13] pour adapter le modèle au contexte de la ville d'Antananarivo. La génération de la carte de la ville est présentée en section 3.4. La dernière section présente les résultats de l'interaction entre environnement et mortalité au niveau des quartiers de la ville, selon les différents indicateurs utilisés et en contrôlant les caractéristiques socio-économiques des quartiers.

3.2 Données de mortalité à Antananarivo

Les données de mortalité de la communauté urbaine d'Antananarivo ont été récupérées pendant une mission à l'Institut Pasteur de Madagascar en octobre 2023. Pendant cette mission, nous avons visité les différents quartiers de la ville pour nous rendre compte de la diversité d'environnements et les problématiques locales. La sous-section 3.2.1 présente la ville d'Antananarivo et les observations que nous avons pu faire lors de notre visite. Ensuite, la sous-section 3.2.2 présente les données de mortalité, et la dernière sous-section présente les indicateurs environnementaux utilisés.

3.2.1 Antananarivo

Antananarivo est la capitale économique et politique de Madagascar. La ville, la plus grande du pays, est localisée au cœur des hautes terres centrales dans la région Analamanga. Elle est située à une altitude d'environ 1 280 m et est entourée de collines, ce qui crée une topologie inégale. Les régions de l'est de la ville ont une altitude plus élevée que les bas quartiers à l'ouest de la ville qui sont sujets à de nombreuses inondations. Les zones à l'est de la ville Le climat d'Antananarivo est classé comme "tropical d'altitude" selon la classification de des climats de Köppen. Dans ce climat, il y a deux saisons principales avec des températures moyennes relativement proches. La ville subit des saisons sèches modérées de mai à octobre (avec des températures entre 10°C et 20°C) et des saisons des pluies de novembre à avril (avec des températures entre 20°C et 30°C). Grâce à son altitude relativement élevée, la présence de moustiques est faible, ce qui réduit le risque de maladie vectorielle comme le paludisme sans toutefois préserver la population d'épisodes épidémiques [75].

Antananarivo est composé de 6 arrondissements et de 192 quartiers, appelés *fokontany*. La structuration des quartiers de la villes reflète les différentes divisions de la population qui ont pu exister au cours du temps (selon le niveau socio-économique notamment) . Sur les hauteurs de la ville, dite "la haute ville", se trouve le Palais de la Reine, entouré de zones résidentielles pour les populations les plus aisées. Depuis l'époque pré-coloniale, les populations les plus pauvres et les migrants ruraux occupent les bas-quartiers de l'ouest de la ville, non loin des quartiers administratifs de la ville, plus aisés, qui sont aussi en bas de la colline. Ce sont aussi les quartiers les plus denses, notamment la cité des 67 hectares et Isotry. Tout l'espace est utilisé

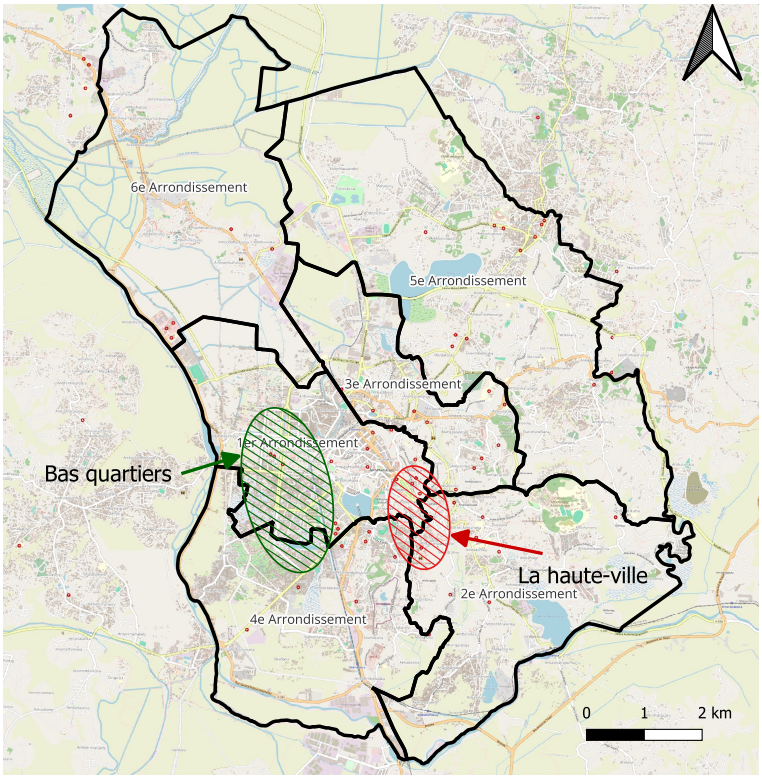


Figure 3.1: Carte d'Antananarivo indiquant les arrondissements, la haute-ville et les bas quartiers de l'ouest de la ville. Le fond de carte provient d'OpenStreetMap. Le sixième arrondissement, au nord-ouest, n'est pas inclus dans la base de données de mortalité et n'est pas considéré dans cette étude.

pour le bâti alors que peu est réservé à des espaces verts. Les bas-quartiers sont constitués d'un mélange de bâti formel (par exemple, des maisons construites en briques) et de zones de bidonvilles (construction en matériaux légers, comme la tôle), ce qui les fragilise encore plus lors des fréquents épisodes d'inondations en saison des pluies. Ces zones informelles ne sont pas bien délimitées et sont intriquées dans les zones formelles : les populations s'installent là où il y a de la place, au coeur de la ville. Dans ces quartiers, l'accès à l'eau potable ainsi que l'évacuation des eaux usées sont limités. Les déchets sont en général jetés dans les différents canaux qui traversent la ville du nord au sud. Ces canaux se retrouvent souvent bouchés, ce qui complique encore plus l'évacuation des déchets. Les fokontany à l'est de la ville sont situés sur des collines, plus basses que celle de la haute ville. Cela leur permet d'être moins sujets aux inondations. Ces quartiers sont pour les uns résidentiels, avec des populations plus aisées, et pour les autres ruraux avec des rizières. Les différents arrondissements de la ville et les points d'intérêts évoqués dans cette partie sont présentés Figure 3.1.

Bien qu'il existe des transports en commun, l'usage de la voiture est très répandu. L'organisation de la ville ne permet qu'une circulation compliquée : les embouteillages sont très importants à toute heure de la journée, notamment dans le centre de la ville, autour de la haute ville. Le parc de véhicules étant ancien, la pollution au gaz d'échappement est très forte, un peu partout dans la ville.

(a) Cultures urbaines. (b) Parc Ambohijatovo. (c) Canal Nord/Sud

Figure 3.2: Photos prises pendant la mission à Antananarivo d'octobre 2023. (a) Vue des cultures du 2e arrondissement depuis le Rova (Palais de la Reine). (b) Vue du parc Ambohijatovo, dans la Haute Ville. (c) Vue du canal traversant la ville du nord au sud en passant par les bas quartiers. Nous pouvons voir les déchets jetés dans le canal qui empêchent l'évacuation des eaux.

L'environnement direct des habitants d'Antananarivo est très varié, non seulement par la topologie du sol qui est très différente selon les quartiers mais aussi par l'évolution de l'urbanisme de la ville. Au sein d'un même fokontany, des ménages peuvent avoir des conditions très différentes, en particulier lorsque l'on regarde la nature des habitations. Une caractérisation fine de l'environnement est donc nécessaire pour pouvoir capturer la complexité de ces environnements, qui ne dépend pas uniquement de la réalité topographique mais aussi de la réalité socioéconomique des ménages. Ces caractéristiques des ménages, à la fois environnementales et socioéconomiques, devraient avoir des conséquences sur les causes de mortalité des habitants. Les données de mortalité, présentées dans la section suivante, permettent d'évaluer l'impact des caractéristiques environnementales sur la santé.

3.2.2 Les décès enregistrés et leurs causes

Alors que la proportion de décès non déclarés est encore importante à Madagascar[2], un enregistrement systématique des décès existe à Antananarivo. En effet, de longue date, la Communauté Urbaine d'Antananarivo (CUA) ne fournit de permis d'inhumer que sous condition d'un certificat de décès délivré par un médecin qui note, outre l'identité du défunt, son adresse, son âge et son sexe, ainsi que la cause probable du décès.

Ce permis peut-être obtenu gratuitement auprès du Bureau Municipal d'Hygiène (BMH). Le BMH Isotry, dans le centre de la ville, s'occupe de délivrer les certificats pour les cinq premiers arrondissements de la ville.

La mise en place de ce système a permis la récupération de données de décès, ce qui permet d'effectuer un suivi historique des causes de décès de la ville.

Historiquement, l'INED a soutenu la saisie des informations collectées jusqu'en 2015. Depuis 2016, c'est l'Institut Pasteur de Madagascar (IPM) qui assiste le Bureau Municipal d'Hygiène (BMH) et gère la base de données des décès.

[2]http://unstats.un.org/unsd/demographic/CRVS/CR_coverage.htm

Si le décès a lieu à l'hôpital, la famille du défunt se voit remettre un certificat de décès, rempli par les médecins de l'hôpital, à présenter au BMH qui complète une fiche spécifique. Si le décès a lieu à domicile, un médecin du BMH recueille des informations sur les circonstances pour établir dans la mesure du possible une cause de décès, il peut le cas échéant se déplacer sur le lieu du décès afin de le constater ou enquêter en cas de doute sur la cause pour remplir la fiche. Ces trois documents, fiches de constatation et déclaration de décès, sont présentés dans la Figure 3.3c.

Toutes les informations sont ensuite saisies dans une même base de données. A la suite d'un accord entre l'IPM et l'INED, ont pu être récupérées pour les décès de 2016 à 2023 les informations suivantes : date de naissance, sexe, date du décès, fokontany de résidence du défunt et code cim10[3] de la cause du décès. Grâce au quartier de résidence, il est possible de relier la mortalité et l'environnement au niveau local. Dans une partie ultérieure la mortalité sera estimée pour la période 2016-2020 en mobilisant, outre les décès, les effectifs de population disponibles à partir du recensement national de la population de 2018.

3.2.3 Choix des indicateurs environnementaux

Les données de mortalité sont ainsi accessibles au niveau des fokontany, ce qui constitue le degré de détail le plus précis pour notre analyse de l'environnement. Dans cette sous-section, nous allons traiter les interrogations suivantes :

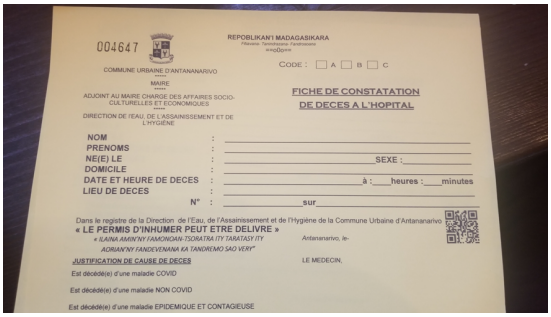
1. Quels aspects de l'environnement voulons-nous modéliser ?
2. Quels indicateurs allons-nous utiliser pour ces modélisations ?

Premièrement, nous voulons modéliser l'environnement de la manière la plus complète possible. En effet, comme indiqué en sous-section 3.2.1, l'environnement au niveau des fokontany est complexe. Il est donc nécessaire de modéliser non seulement la présence de bâti, mais aussi d'avoir une caractérisation plus fine de celui-ci, prenant en compte la densité, la morphologie et les matériaux utilisés. La présence de végétation et l'altitude, étant données les différences fortes entre arrondissements mais aussi entre les différents fokontany d'un même arrondissement, doivent aussi être modélisées.

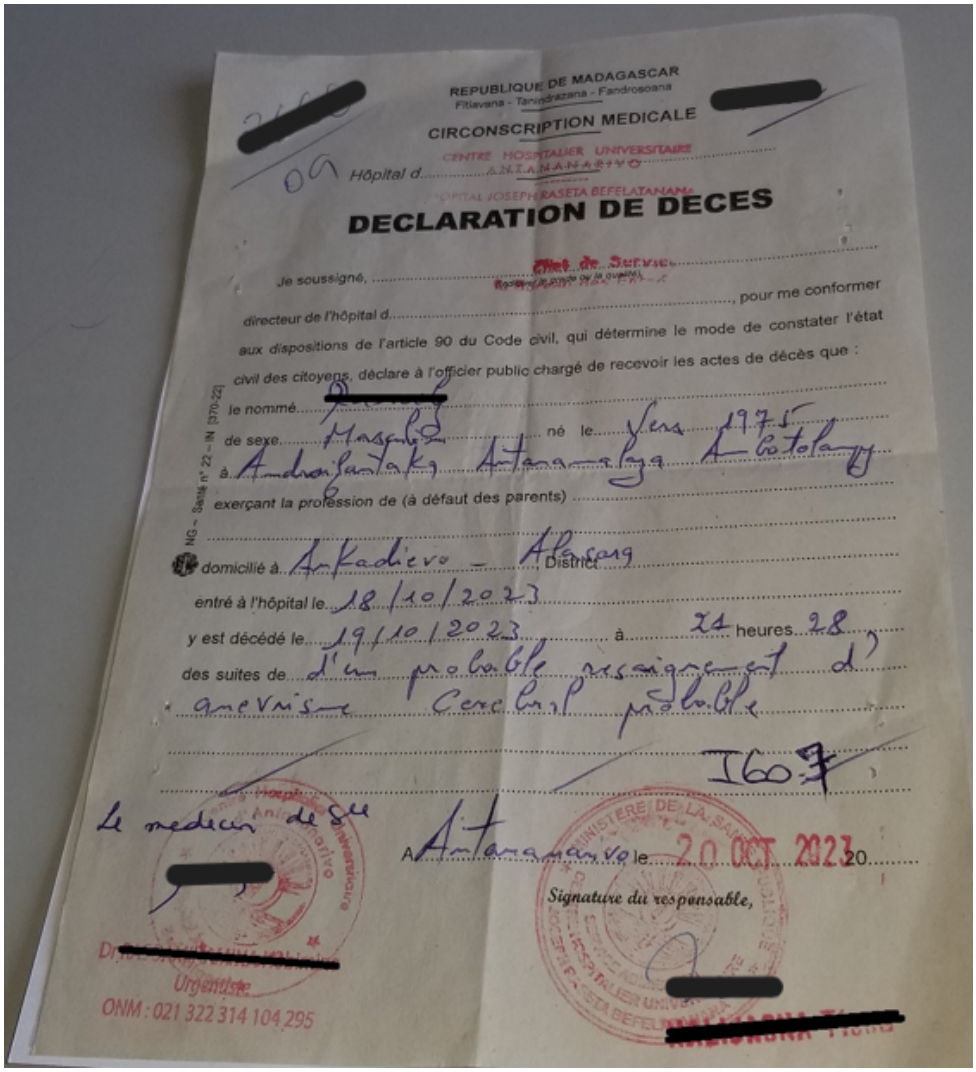
Morphologie de la ville. La caractérisation des structures construites par l'humain mobilise plusieurs aspects. Le système de classification LCZ est suffisamment complet pour apporter des informations sur la densité, la morphologie et les matériaux utilisés pour les bâtiments. En particulier, il est possible de faire la différence entre les quartiers très denses (LCZ 3 *Compact low-rise*) et les quartiers avec des constructions en matériaux légers, plus proches des bidonvilles (LCZ 7 *Lightweight low-rise*). Cette différenciation ne peut pas se faire avec une estimation de la densité de bâtis seule. Les données LCZ disponibles pour la ville sont limitées. Certaines cartes sont librement rendues librement accessible par la communauté scientifique, mais ne sont pas forcément disponible à la date désirée. Une carte peut aussi être générée à partir d'un modèle et de données, comme So2Sat [13], mais ces cartes souffrent de problème d'adaptation de domaine? Une méthode d'entraînement d'un réseau de

[3] https://icd.who.int/browse10/2008/fr

(a) Constatation de décès à domicile.



(b) Constatation de décès à l'hôpital.



(c) Déclaration de décès.

Figure 3.3: Formulaires de constatation de décès à domicile (a) et à l'hôpital (b), ainsi que la déclaration de décès (c).

neurones sera mis en place, et décrite en section 3.3, spécifiquement pour la génération d'une carte LCZ d'Antananarivo. De plus, l'entreprise Google met à disposition des cartes de détection de bâti, notamment pour Madagascar. Ces cartes ont été obtenues à partir de méthodes

d'apprentissage profond et d'images satellites. Pour chaque bâtiment détecté, la confiance du modèle dans sa détection est indiquée. Comme les prédictions du modèle se basent sur un entraînement préalable, elles sont d'autant meilleures que les données d'entrée sont similaires aux données d'apprentissage. A l'inverse, des données d'entrée très différentes des données d'entraînement aboutiront à des prédictions avec des scores de confiance plus faibles. Nous pouvons donc penser que, si le modèle effectue une détection avec une prédiction faible, l'image en entrée montre un bâtiment peu usuel, et à de grandes chances d'être du type des constructions légères qui caractérisent l'habitat informel, comme dans les bidonvilles. Ainsi, nous faisons l'hypothèse que le score de confiance du modèle apporte une information sur la caractérisation du bâti.

Caractérisation de la végétation. L'indicateur le plus facile d'accès pour calculer la présence de végétation est le *Normalized Difference Vegetation Index (NDVI)*. Cet indicateur est calculé directement à partir des images satellites, sans besoin d'apprentissage. En particulier, pour les images Sentinel-2, il est une combinaison des bandes spectrales rouges et proches infra-rouges :

$$NDVI = \frac{B08 - B04}{B08 + B04} \tag{3.1}$$

où $B08$ est la bande proche infrarouge et $B04$ est la bande rouge (les longueurs d'ondes associées sont disponibles dans le Tableau 1.1. Les NDVI peuvent donc être calculés pixel par pixel. Une carte des NDVI a été élaborée pour l'année 2019 avec une résolution de 10 mètres, correspondant aux données Sentinel-2.

Caractérisation de l'altitude. Un modèle d'altitude (DEM) a été utilisé pour cartographier l'altitude des quartiers d'Antananarivo. Ce modèle d'altitude a été téléchargé à partir du portail de données de la NASA[4].

Les indicateurs de végétation et d'altitude sont issus directement de capteurs ou après des calculs élémentaires. En revanche, une carte LCZ est générée avec des méthodes de traitement d'images plus complexes. La partie suivante présente une méthode d'apprentissage profond pour la production de telles cartes, partout dans le monde et en particulier à Antananarivo. Cette méthode est apprise sur le jeu de données So2Sat et sur de l'adaptation de domaine pour étendre les connaissances d'un modèle appris à toutes les régions du monde.

3.3 Cartographie du sol par critères physiques

La définition des LCZ est basée sur des critères physiques (Tableau 2.1), qui ne dépendent pas d'aspects culturels mais seulement des formes et matériaux présents dans les zones d'études. Pour appartenir à une classe, une zone (pour nous, la zone recouverte par l'image satellite) doit donc posséder certaines caractéristiques respectant des intervalles de valeurs, pour chaque critère physique. Cette représentation par intervalle peut être associée à la notion de prototype : une manière de représenter les données, censée ne pas dépendre des domaines. Cette section porte sur l'utilisation de ces prototypes pour la classification des LCZ, dans un con-

[4]https://www.earthdata.nasa.gov/

texte d'apprentissage supervisé classique et dans le contexte de l'adaptation de domaine. La sous-section 3.3.1 présente la classification d'images supervisée conventionnelle, dont le but est de prédire ces descripteurs et non la classe LCZ en elle-même. La classe est déduite des valeurs de descripteurs. Ensuite, la sous-section 3.3.2 présente un modèle exploratoire qui intègre ces prototypes comme vecteurs de références dans l'adaptation de domaine. Ce sont donc des prototypes pré-définis, qui ne sont pas appris pendant la phase d'entraînement du modèle.

3.3.1 Apprentissage des Local Climate Zones par descripteurs

Dans cette sous-section est présentée la méthode d'apprentissage supervisé pour la prédiction de descripteurs. Prédire les descripteurs, plutôt que les classes, permettrait de nous apporter des informations supplémentaires sur les performances en classification. En effet, les prédictions du modèle seront plus interprétables, car il sera possible d'identifier les descripteurs mal estimés. Selon les données utilisées, deux conclusions pourront être tirées de ces classifications insatisfaisantes. Si les données de test sont similaires aux données d'entraînement (par exemple dans le cas d'une séparation aléatoire dans un jeu de données), nous pourrons observer quels sont les descripteurs, donc les informations les plus difficiles à extraire dans l'image. Si les données de test sont différentes des données d'entraînement (donc dans le cas d'une différence de distribution entre les données d'entraînement et de test), nous pourrons observer quels descripteurs sont difficiles à transférer d'un domaine vers un autre.

L'apprentissage de ces indicateurs n'est pas trivial, car les étiquettes ne sont pas des valeurs exactes mais des intervalles. Cette section se penche sur l'étude de fonctions de coûts adaptées pour un tel apprentissage.

Apprentissage par intervalles

La tâche à résoudre est un peu différente d'une tâche de régression classique. Dans la régression conventionnelle, le but est de prédire une valeur, et le modèle est optimisé selon l'écart entre la valeur réelle et la valeur prédite. Dans notre cas, nous n'avons pas de valeur prédite déterminée mais un intervalle de possibilités. Intuitivement, on voudrait que le coût soit nul dans cet intervalle, et de plus en plus fort lorsque l'on s'éloigne des bornes de cet intervalle. En entraînant sur beaucoup de données, on espère que le modèle finira par estimer les valeurs réelles des descripteurs de manière satisfaisante.

Soit $F(.)$ un réseau de neurones qui accepte pour entrée une image $\mathbf{x}$. La couche de sortie de $F(.)$ est composée de deux couches linéaires. La première couche produit un vecteur $d = [d_1, d_2, ..., d_N]$ dont chaque coefficient d_i est la valeur estimée pour le descripteur i, et N est le nombre de descripteurs décrits par des intervalles. La deuxième couche est une couche de classification, pour prédire la valeur d_{rugo} du descripteur de rugosité, décrit par des classes (Tableau 2.1). La classe LCZ déduite de ce jeu de descripteurs est la classe dont le plus d'intervalles, et la classification, sont satisfaits. Si L_i^{desc} est la fonction de coût de régression pour le descripteur i décrivant l'écart entre la prédiction d_i et l'intervalle de valeurs acceptables, L_{rugo}^{desc} la fonction de coût de classification du descripteur de rugosité, et $\alpha \in [0, 1]$,

la fonction de coût totale est la suivante :

$$L = \alpha \times L_{rugo}^{desc} + \frac{1}{N} \times \sum_{i=1}^{N} (1-\alpha) \times L_i^{desc} \tag{3.2}$$

Nous utilisons les trois versions de So2Sat [13], présentées en section 2.3.3, pour entraîner et évaluer notre modèle : *Random* (pas d'adaptation), *Block* (adaptation faible) et *Cultural_10* (adaptation forte). Ces trois versions nous permettront d'évaluer l'impact de chaque descripteur sur la classification, en fonction du contexte. Plusieurs fonctions de coût ont été implémentées pour cet apprentissage par intervalle. Leur objectif est de laisser suffisamment de liberté au modèle dans l'intervalle des descripteurs mais en le contraignant fortement en dehors. Nous définissons donc les cinq fonctions de coût ci-dessous:

- L_{int}**, la fonction de coût par intervalle.** Soit la fonction g définie par morceaux sur $\mathbb{R}$, étant la fonction nulle dans un intervalle $I = [a, b], (a, b) \in \mathbb{R}^2$ et la norme L_1 en dehors :

$$L_{int} = \begin{cases} 0 & \text{si } d \in I \\ |d| - a & \text{si } x < a \\ |d| - b & \text{si } x > b \end{cases} \tag{3.3}$$

  Cette fonction est la fonction la plus simple permettant de résoudre notre tâche. Cependant, ses discontinuités en a et b la rendent peu adaptée à l'entraînement de réseaux de neurones. De plus, un coût nul donnerait une valeur de gradient aussi nulle : le modèle n'apprend pas.

- L_{phuber}**, la fonction de coût "pseudo" Huber.** Cette fonction est une adaptation de la fonction de Huber qui prend des valeurs quadratiques dans un intervalle, et des valeurs linéaires hors de cette intervalle. La fonction de pseudo-Huber est une approximation lisse de la fonction de Huber, qui apporte un coefficient à la fonction linéaire hors de l'intervalle, avec δ un paramètre permettant de contrôler la pente de la partie linéaire et l'intervalle de la partie quadratique :

$$L_{phuber} = \begin{cases} \frac{1}{2}d^2 & \text{si } |d| \leq \delta \\ \delta(|d| - \frac{1}{2}\delta) & \text{si } |d| > \delta \end{cases} \tag{3.4}$$

- L_{rugo}^{desc} est la fonction d'entropie croisée :

$$L_{rugo}^{desc} = -\sum_{i=1}^{C} d_{rugo}^{i} \log(\hat{d}_{rugo}^{i}) \tag{3.5}$$

  où :

  - C est le nombre de classes,
  - d_{rugo}^{i} est la vraie étiquette,

- $\hat{d}^i_{rugo}$ est la probabilité prédite pour la classe i.

- L_2 est la fonction de coût quadratique :

$$L_2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{d}_i)^2 \tag{3.6}$$

où :

- n est le nombre total d'exemples dans l'ensemble de données,
- y_i est la valeur réelle de la cible pour l'exemple i,
- $\hat{y}_i$ est la valeur prédite pour l'exemple i.

- L_{bas}**, la fonction de coût basse.** Nous définissons une fonction de coût pour éviter les discontinuités aux bornes de l'intervalle, et éviter d'avoir des valeurs de coût nulles. L'idée est d'avoir une fonction qui prend des valeurs très basses dans un intervalle, et qui prend les valeurs d'une L_1 en dehors. Cette fonction, inspirée de fonctions polynomiales, est définie par morceaux.

  Le raisonnement sous-jacent est le suivant :

  1. Dans l'intervalle, les valeurs doivent être proches de 0. La compression de la fonction est d'autant plus grande que l'intervalle est grand. Soit $|I| = |b - a|$ la taille de l'intervalle. Nous cherchons un réel η tel qu'entre $-\eta$ et η la fonction est de la forme $L_{bas} = \eta \times t(|I|) \times x^p$, où t est une fonction définie sur $\mathbb{R}_+^*$, et p un entier naturel pair. Une fonction de ce type permet de contrôler la pente de la fonction en fonction de la taille de l'intervalle $|I|$.
  2. En dehors de l'intervalle, la fonction doit prendre les valeurs de L_1 : $L_{bas} = |d| \pm \beta$

  Nous allons exprimer les valeurs de η et β en fonction de $t(i)$ pour assurer la continuité de la fonction. Ce système peut être résolu grâce aux égalités des morceaux de la fonction et des égalités de leur dérivées, en η.

$$\begin{cases} t(|I|) \times \eta^p = \eta + \beta \\ t(|I|) \times p \times \eta^{p-1} = 1 \end{cases} \tag{3.7}$$

$$\begin{cases} \beta = \eta \times (t(|I|) \times \eta^{p-1} - 1) \\ \eta = \sqrt[p-1]{\frac{1}{t(|I|)\times p}} \end{cases} \tag{3.8}$$

$$\begin{cases} \beta = \sqrt[p-1]{\frac{1}{t(|I|)\times p}} \times (\frac{1}{p} - 1) \\ \eta = \sqrt[p-1]{\frac{1}{t(|I|)\times p}} \end{cases} \tag{3.9}$$

La Figure 3.4 présente une représentation graphique de toutes ces fonctions.

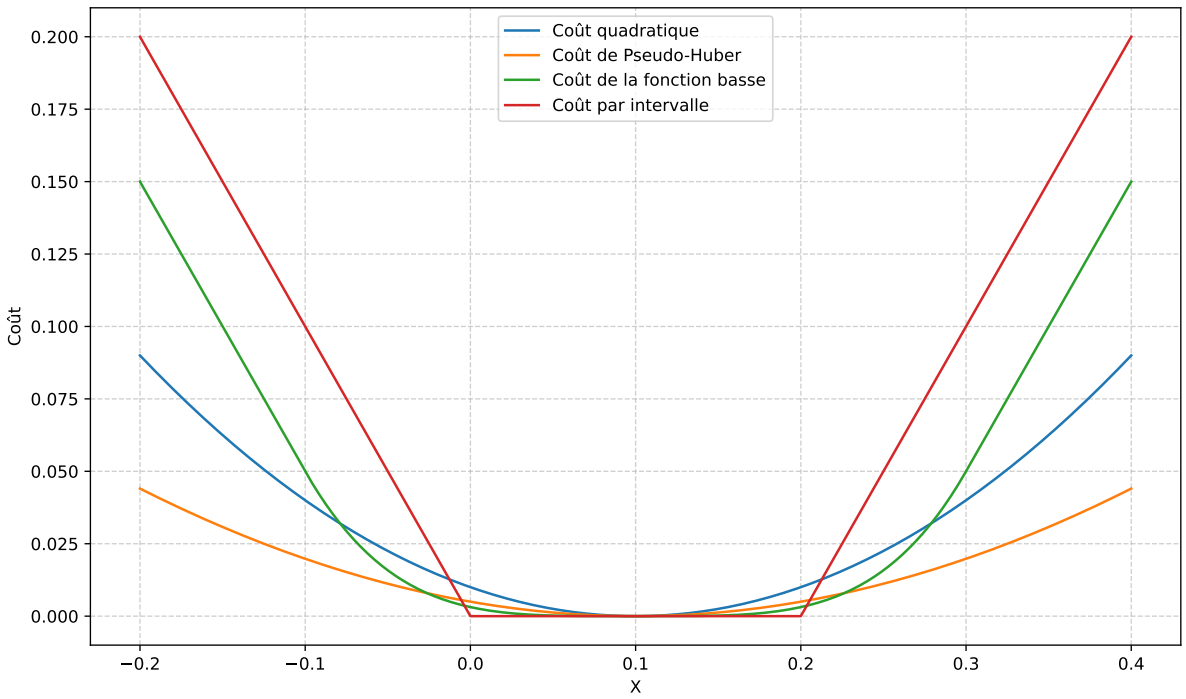


Figure 3.4: Représentation graphique des fonctions de coûts pour un intervalle de valeur acceptables de $[0, 0.2]$

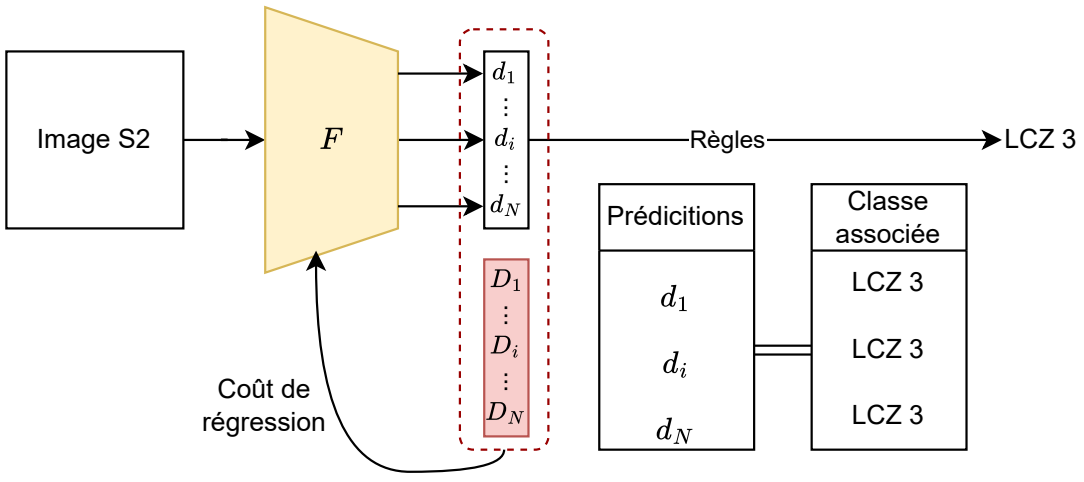


Figure 3.5: Vue d'ensemble du modèle d'apprentissage par intervalle. Un réseau de neurones F prend en entrée une image Sentinel-2 et produit en sortie un vecteur de N valeurs prédites pour chacun des N descripteurs. La classe LCZ finale est déduite de ces prédictions.

Paramètres d'entraînement

Le modèle $F(.)$ est un ResNeXt [76] et sera entraîné à partir de zéro. Ce modèle permet d'extraire une plus grande diversité de caractéristiques en regroupant les convolutions, ce qui conduit à des représentations plus riches. Pour toutes les fonctions de coût (L_{int} , L_2, L_{bas} et L_{phuber}), nous utilisons un taux d'apprentissage de 0,001, qui a empiriquement donné les meilleurs résultats. La taille du *batch* pour l'entraînement est fixée à 128. Nous appliquons également des techniques classiques d'augmentation de données à nos données d'entraînement : retournement horizontal, retournement vertical, et standardisation. Cette standardisation consiste en retirant aux pixels des images la moyenne et en divisant par l'écart-type des pixels du dataset d'entraînement tout entier. Cette méthode permet de mettre les caractéristiques des images sur une échelle similaire pour améliorer l'apprentissage. Les phases d'entraînement, de validation et de test ont été réalisées à l'aide de cartes graphiques NVIDIA V100 de 16/32 Go, disponibles sur le supercalculateur HPE SGI 8600 Jean-Zay.

Résultats

Le Tableau 3.1 présente les résultats des modèles sur les trois versions de la base de données So2Sat, avec différentes fonctions de coûts. Dans tous les cas, les résultats de classification après régression sont inférieurs aux résultats utilisant une couche de classification directe. Entre les modèles de régression, les résultats dépendent de la tâche à résoudre.

Pour la classification la plus simple, sans adaptation de domaine, la fonction de coût permettant d'obtenir les meilleurs résultats est la L_2, qui exerce une contrainte plus forte sur le modèle lorsque la prédiction est éloignée du centre de l'intervalle, mais reste 3 points en dessous du modèle conventionnel. Les modèles ayant été entraînés avec L_{bas}, L_{int} et L_{phuber} sont respectivement 9, 10 et 14 points en dessous du modèle classique.

Pour la tâche de classification intermédiaire, avec une légère adaptation de domaine au niveau des villes, la fonction L_2 permet une nouvelle fois d'obtenir les meilleurs résultats pour les modèles de régression. L'écart avec les autres modèles de régression est cependant plus faible, avec trois, quatre et cinq points pour respectivement L_{bas}, L_{int} et L_{phuber}.

Pour la tâche de classification la plus compliquée, impliquant une adaptation de domaine au niveau de zones culturelles, la tendance semble s'inverser entre les modèles de régression. Le modèle le plus performant a été entraîné avec L_{phuber} qui possède une métrique supérieure de 1,23 point par rapport au modèle entraîné avec L_{bas}, 1,36 par rapport au modèle entraîné avec L_{int} et 1,41 par rapport au modèle entraîné avec L_2.

Discussion

L'utilisation de plusieurs régressions permet de souligner les forces et faiblesses du modèle entraîné. Chacun des descripteurs correspond à une caractéristique de l'image précise, que le modèle doit extraire pour effectuer les régressions. Les résultats de classification sur les versions *random* et *block*, avec peu ou pas d'adaptation, permettent de mettre en évidence

Coût utilisé	***Random***	***Block***	***Cultural_10***
Entropie croisée (classification)	97.74 %	84.75 %	59.91%
L_{int}	87.38 %	77.91 %	55.97 %
L_2	**93.90** %	**81.31** %	55.92 %
L_{bas}	88.29 %	77.49 %	56.10 %
L_{phuber}	83.35 %	75.98 %	**57.33** %

Tableau 3.1: Résultats de classification sur les jeux de test des trois version de So2Sat.

les descripteurs les plus difficiles à classifier. Les résultats sur la version *Cultural_10* permettent d'évaluer la transférabilité d'un modèle relative à chacun des descripteurs. La Figure 3.6 présente les histogrammes d'erreurs de chaque descripteur dans le jeu de test des trois versions de So2Sat, pour le modèle utilisant le coût L_2. Le modèle commet une erreur lorsque la valeur de régression prédite pour un descripteur est hors de l'intervalle de prédiction de ce descripteur pour une classe donnée. D'après les histogrammes de la Figure 3.6 indique que deux descripteurs sont mal classés sur les trois versions : la production de chaleur anthropique et la classification de la rugosité. La chaleur anthropique est la chaleur produite par les activités humaines, comme les transports, le chauffage ou encore la climatisation. Les valeurs définies pour les classes rurales (A-G) sont nulles, car il n'y a peu d'activité humaine. Dans le modèle de régression, l'intervalle correspondant pour l'apprentissage a été fixé à $[0, 0.01]$. Les régressions qui ne sont pas dans cet intervalle, même proches de zéro, sont comptées comme incorrectes et augmentent artificiellement le nombre d'erreurs pour ce descripteur.

A part pour la chaleur anthropique, le modèle commet le plus d'erreurs pour les descripteurs *Aspect Ratio* et *Pervious surface fraction*. L'*aspect ratio* est le rapport hauteur des éléments de l'image et la largeur des espaces ouverts. Par exemple pour les zones urbaines, *aspect ratio* est calculé être la hauteurs des bâtiments et la largeur des rues parcs et autres. Pour les zones rurales, ce calcul est fait avec la hauteur des arbres ou de la canopée et la largeur des prairies ou des clairières. Lorsque les espacements sont suffisamment étroits, en particulier inférieurs à la résolution des images Sentinel-2, cette caractéristique de l'image est plus difficile à extraire pour le modèle. Les erreurs de régression sont ainsi plus fréquentes avec ce descripteur. La prédiction de la valeur de *Pervious surface fraction*, ou la fraction de surface perméable, est aussi complexifiée par la résolution de l'image, en particulier pour les classes représentant des zones bâties. Dans ces cas, les surfaces sont à la fois recouvertes de bâtiments et de zones perméables et imperméables, de tailles inférieures à la résolution de Sentinel-2.

Un descripteur est plus transférable lorsque le nombre d'erreurs est faible : le modèle n'a pas de mal à extraire cette information de l'image. La perte de précision du modèle lorsque l'adaptation est plus forte (*random* $\Rightarrow$ *cultural_10*), indiquée dans le Tableau 3.1, n'est pas causée par la mauvaise classification d'un descripteur en particulier. Une hausse des erreurs générale pour tous les descripteurs est visible.

D'autre part, plusieurs fonctions de coût ont été utilisées pour l'apprentissage des LCZ, laissant plus ou moins de liberté au modèle de prédire des valeurs de régressions dans l'intervalle. La L_{int} et la L_{bas} prennent des valeurs basses ou nulles dans l'apprentissage, ce qui permet au modèle de prédire des valeurs variées sans que les répercussions dans l'apprentissage soient

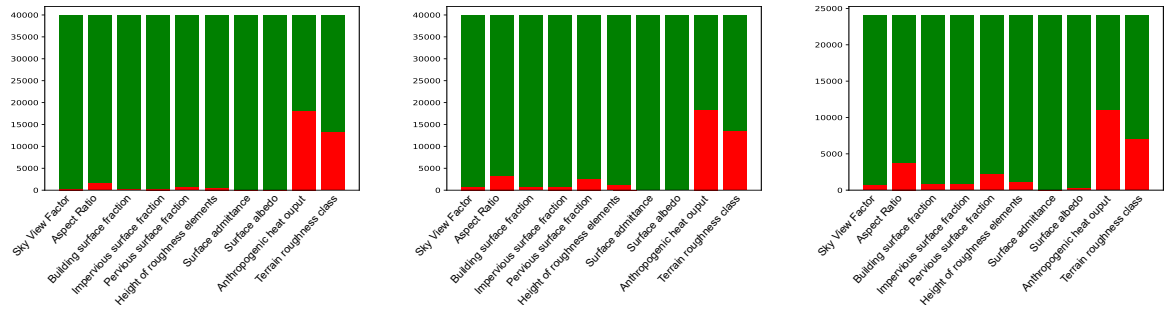


Figure 3.6: Histogrammes des erreurs par descripteurs pour les jeux de test des versions *random*, *block* et *cultural_10* pour un modèle entraîné avec la L_2. En rouge sont indiquées les erreurs, c'est-à-dire lorsque les valeurs prédites ne sont pas dans l'intervalle de description.

importantes. Il faut tout de même noter qu'avec la fonction L_{bas} et sa forme polynomiale dans l'intervalle, le modèle sera incité à prédire cette valeur centrale. A l'inverse, les fonctions L_2 et L_{phuber} mettent beaucoup plus de contrainte sur le modèle pendant l'entraînement, car les valeurs de coût vont croître fortement lorsque l'on s'éloigne de la valeur centrale de l'intervalle. Le modèle est donc incité à prédire cette valeur et non à prédire une valeur qui pourrait correspondre à la valeur réelle du descripteur pour cette classe. Les Figures 3.7 et 3.8 présentent les histogrammes des régressions pour la classe *Large low-rise*, entraînés avec la L_2 et la L_{int}, qui posent respectivement le plus et le moins de contrainte au modèle pendant l'entraînement. Comme prévu, le spectre des valeurs prédites est bien plus large pour la fonction de coût L_{int}, alors qu'il est très étroit et proche de la valeur centrale pour la L_2. Cet intervalle étroit ne représente pas les valeurs réelles, car ne permettent pas de représenter des images avec une grande diversité comme celles du jeu de données So2Sat, qui proviennent, de zones culturelles différentes. Le modèle prédit des valeurs similaires pour toutes les images. A l'inverse, le modèle entraîné avec la fonction L_{int} produit des histogrammes bien plus larges, et qui ne sont pas centrés sur les valeurs centrales des intervalles possibles. L'évaluation quantitative de ces prédictions nécessite l'étiquetage de chaque descripteurs des images de test, ce qui n'a pas été fait dans le cadre de ces travaux. Il est cependant possible de discuter ces résultats de manière qualitative, en observant la tendance des histogrammes.

Exemple 1 *Sky view Factor : Ce descripteur désigne le rapport entre la partie du ciel visible depuis le sol et la partie du ciel visible lorsque celui-ci n'est pas obstrué. Un Sky view Factor proche de 1 signifie que l'observateur, depuis le sol, voit une grande partie du ciel, donc que les bâtiments qui l'entourent sont très bas. A l'inverse, un Sky view Factor très faible indique que les bâtiments qui l'entourent sont très hauts et très denses. Dans la classe Large low-rise, les bâtiments sont larges, peu densément construits, et bas car sinon la zone associée pourrait être classée comme Open high-rise. La distribution des erreurs pour cette classe et ce descripteur, affichée en Figure 3.8, indique que la plupart des images ont un Sky view Factor autour de 0.8, ce qui correspond à des bâtiments de quelques étages, et que le nombre d'images ayant un Sky view Factor supérieur à 0.8 décroît rapidement. Cette observation est plausible, car des bâtiments larges de ce type se limitent à quelques étages, mais n'en ont rarement qu'un seul.*

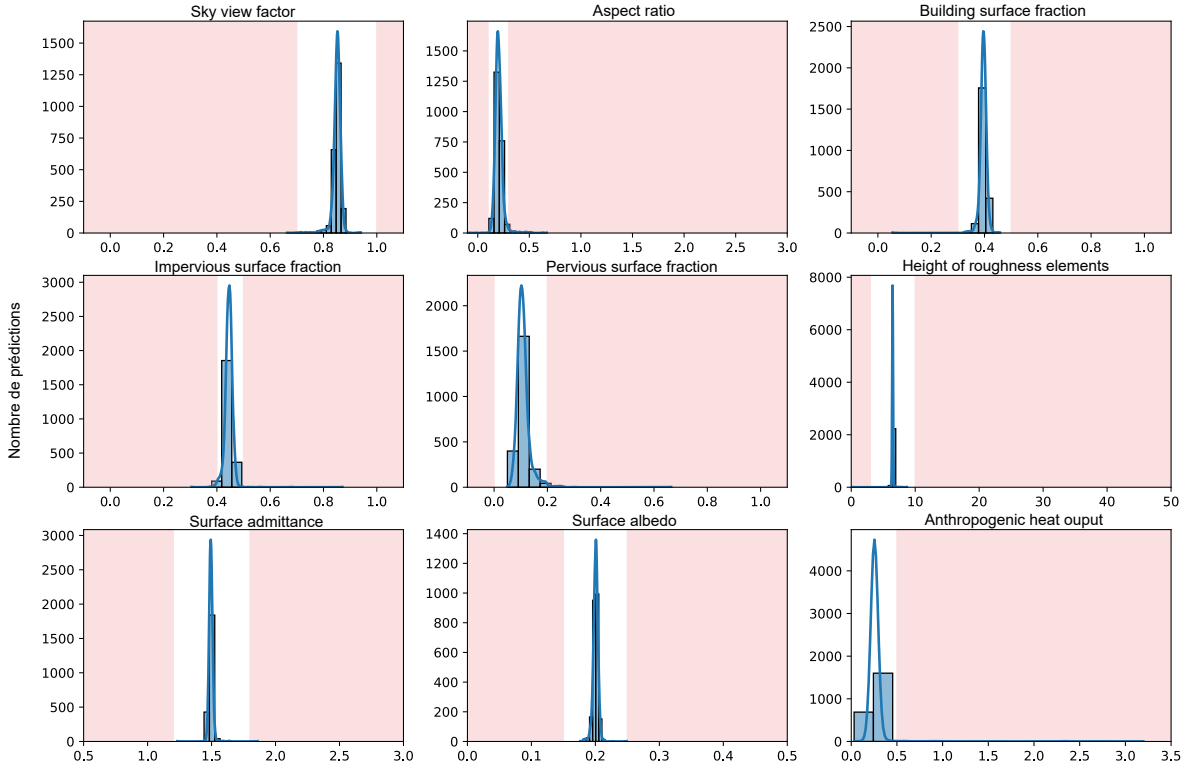


Figure 3.7: Histogramme des valeurs prédites pour la classe *Large low-rise* par un modèle entraîné avec la fonction de coût L_2. Les intervalles blancs sont les intervalles de valeurs possibles pour les descripteurs. Les intervalles rouges sont hors de l'intervalle étiquette.

Exemple 2 *Height : Ce descripteur désigne la hauteur moyenne des bâtiments de la zone, en mètres. L'histogramme de la Figure 3.8 indique que le nombre d'images, donc de zones, est croissant lorsque la hauteur moyenne grandit. Cette observation est cohérente avec l'exemple précédent : dans le jeu de données d'images de la classe Large low-rise, il y a plus de zones de bâtiments hauts que de bâtiments bas. Cette distribution est aussi plausible mais doit être validée quantitativement avec un jeu de données étiquetées.*

Les résultats de classification à l'aide des descripteurs sont encourageants, et peuvent être interprétés plus facilement, tant sur la classification des LCZ que sur la transférabilité de ces modèles de classification. La section suivante utilise ces descripteurs comme une représentation particulière des LCZ, universelle, pour permettre une adaptation de domaine. Cette représentation est, d'une part, un point d'ancrage pour les domaines sources et cibles et est, d'autre part, une représentation universelle pour l'oeil humain.

3.3.2 Adaptation de domaine basée sur des descripteurs physiques

Bien que la définition des LCZ ait été définie pour être universelle, la traduction de leur classification en tant que tâche de traitement d'image ne les exempt pas des problèmes de vision.

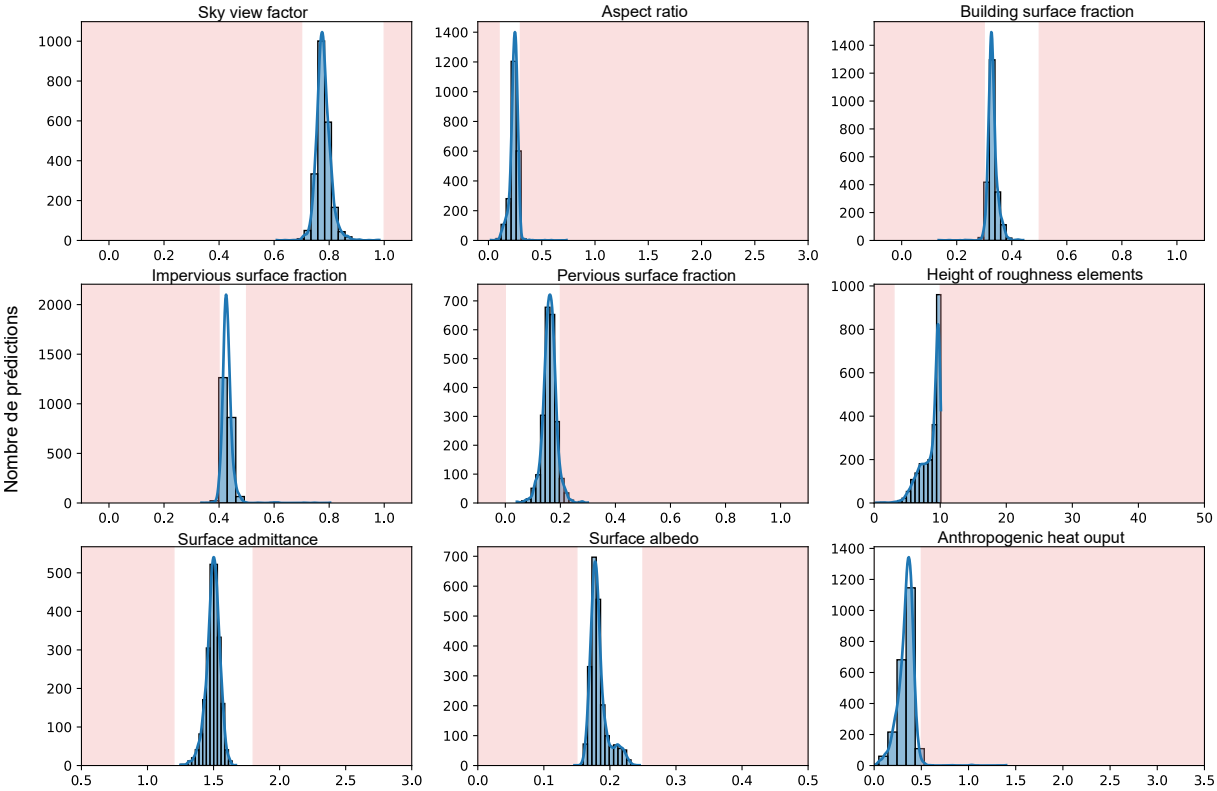


Figure 3.8: Histogramme des valeurs prédites pour la classe *Large low-rise* par un modèle entraîne avec la fonction de coût L_{int}. Les intervalles blancs sont les intervalles de valeurs possibles pour les descripteurs. Les intervalles rouges sont hors de l'intervalle étiquette.

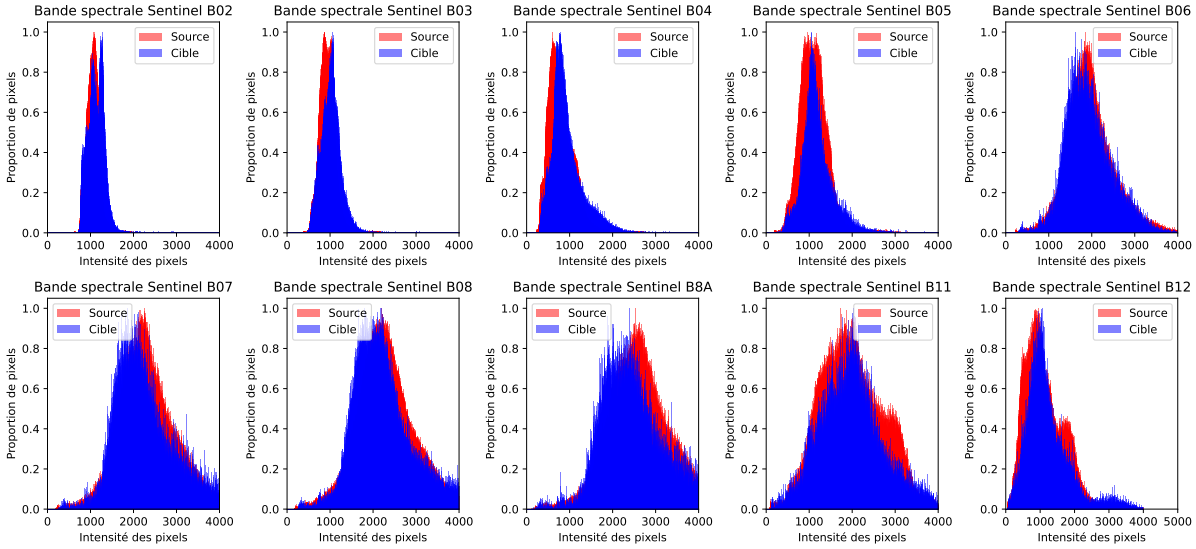


Figure 3.9: Histogrammes par bande spectrale des images sources et cibles pour la classe *Scattered trees*. Dans ce cas, il y a peu de différences entre les domaines.

Le transfert d'un apprentissage d'une région à une autre pose des questions d'adaptation de domaine, même si les caractéristiques physiques sont censées être universelles. Par exemple dans le cas des LCZ, la différence de matériau, de forme des habitations ou encore des différences de type de végétation change l'aspect général de l'image, ce qui explique le besoin d'une méthode d'adaptation. Cette sous-section présente la méthode d'adaptation de domaine développée pour la classification des LCZ. Cette méthode se base sur la définition des LCZ. Dans un premier temps, nous justifions le besoin d'une méthode d'adaptation en section 3.3.2. Ensuite, la méthode d'inclusion de la définition physique des LCZ est présentée en section 3.3.2.

Justification du besoin d'adaptation

Dans ces travaux, notre domaine source est le jeu d'entraînement de la version *Cultural_10* de So2Sat, et le domaine cible est le jeu de test, construit à partir de villes de régions culturelles différentes du jeu d'entraînement. L'écart entre domaines peut être observé en regardant les histogrammes des images des données sources et cibles. Les Figures 3.9 et 3.10 présentent les histogrammes pour chaque bande spectrale Sentinel-2 et pour les classes LCZ *Scattered trees* et *Bush, scrub* . La Figure 3.9 montre une adaptation qui semble être légère, car la moyenne et la variance des distributions semblent être similaires. Dans ce cas, on s'attend à ce qu'un modèle d'apprentissage profond arrive à transférer ces connaissances d'un domaine à un autre. A l'inverse, la Figure 3.10 montre une adaptation plus complexe. Comme les distributions sont très différentes, le modèle aura du mal à transférer ses connaissances. L'observation de ces différences suggère le développement de méthodes d'adaptation pour la cartographie des LCZ.

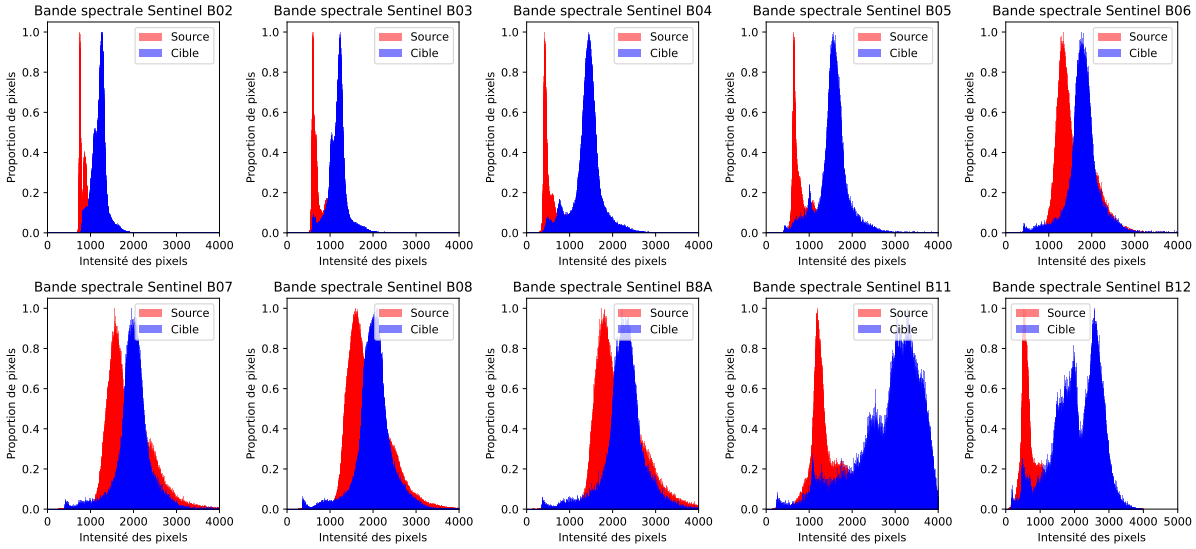


Figure 3.10: Histogrammes par bande spectrale des images sources et cibles pour la classe *Bush, scrub*. L'écart entre les domaines est important et les formes des distributions sont différentes.

Adaptation de domaine directe

Les exemples indiqués dans la partie précédente montrent que, pour les LCZ, l'écart de domaine n'est pas une différence de moyenne et de variance entre les différentes distributions. La méthode d'adaptation développée doit pouvoir créer des représentations complexes, d'ordre plus haut que la moyenne et la variance. Ces représentations plus complexes peuvent être modélisées avec un apprentissage adversarial, plutôt que les méthodes basées sur la divergence, pour aligner les distributions source et cibles dans l'espace latent. Soit $F(.)$ un réseau de neurones prenant en entrée une image x et produisant un vecteur de sortie $s = [s_1, s_2, ..., s_{17}]^T$ de score de prédiction pour chaque classe LCZ. La classe prédite est la classe qui obtient le plus haut score de prédiction. Soit $D_S = (x_i^T, y_i^T)_{i \in [\![1, n_S]\!]}$ l'ensemble de données sources, ici la base d'entraînement de *Cultural_10* et $D_T = (x_i^T, y_i^T)_{i \in [\![1, n_T]\!]}$ l'ensemble de données cible tirées de la base de test de *Cultural_10*. L'objectif est de maximiser la précision sur D_T. Les prototypes en adaptation de domaine sont utilisés comme une représentation des classes dans l'espace latent, communes aux deux domaines. Lors de l'apprentissage, des informations invariantes sont extraites des données et stockées pour servir de référence lors de la phase de test. Nous proposons ici de prendre le problème dans le sens inverse. Nous avons déjà une représentation supposée universelle des LCZ via les intervalles de descripteurs. Soit M une matrice de taille $[N, 2]$ dont les coefficients par ligne sont les bornes inférieures et supérieures des intervalles de descripteurs :

$$M = \begin{pmatrix} D_1^{inf} & D_1^{sup} \\ D_2^{inf} & D_2^{sup} \\ \dots & \dots \\ D_N^{inf} & D_N^{sup} \end{pmatrix}$$

Cette matrice est intégrée dans le processus d'apprentissage, composé de deux chemins : le premier pour aligner les distributions dans l'espace latent, et le second pour forcer une

Modèle	***OA***
Zhao *et al.* (2023) [15]	**61,03** %
LCZmemoryNet (notre modèle)	57,28 %

Tableau 3.2: Résultats de classification le jeu de test de la version *Cultural_10* de So2Sat [13].

représentation de l'espace des données selon des prototypes fixes qui représentent M en utilisant un mécanisme d'attention. Ce processus est représenté sur la Figure 3.11.

Le premier chemin utilise l'apprentissage adversarial, dont le but est de rapprocher les représentations sources et cibles dans l'espace latent. Une image x (source ou cible) est présentée au modèle, qui produit une représentation F_T de x dans l'espace latent. F_T est ensuite traitée par un réseau discriminateur $Disc(.)$ qui prédit son ensemble de provenance : D_S ou D_T. Un coût adversarial $L_{adversarial}$ est calculé et utilisé pour entraîner à la fois $Disc(.)$ et $F(.)$. Par ailleurs, F_T est traité par un classifieur C_1 qui prédit la classe LCZ, avec un coût supervisé $L^1_{classification}$. La fonction de coût totale pour ce chemin, et pour optimiser $F(.)$ est :

$$L_{chemin1} = L^1_{classification} + L_{adversarial} \tag{3.10}$$

En plus de la classification, le second objectif du modèle $F(.)$ est donc de produire des représentations similaires pour les données sources et cibles.

Le second chemin intègre la matrice M dans l'apprentissage via un mécanisme d'attention. Seules les images sources, étiquetées, sont utilisées dans ce chemin. Dans un premier temps, la matrice M et la carte des représentations F_T sont projetées dans un espace de même dimension, produisant les vecteurs m et f'_T. m et f'_T sont utilisés comme entrées pour un mécanisme d'attention, qui mesure la similarité entre les deux représentations. f'_T est ensuite projeté en vecteur *query* et m en vecteurs *key* et *value* pour le mécanisme d'attention, ce qui produit une sortie O. Un classifieur C_2 prédit la classe LCZ à partir de cette sortie, produisant le coût $L^2_{classification}$. La fonction de coût utilisée pour l'entraînement est la suivante :

$$L = L^1_{classification} + L_{adversarial} + L^2_{classification} \tag{3.11}$$

Résultats

Cette partie présente les résultats de la méthode présentée en Section 3.3.2. Celle-ci est comparée à la méthode présentée par [15], actuellement à l'état de l'art de l'adaptation de domaine des LCZ, avec les données du jeu de test de la version *Cultural_10*. La métrique utilisée pour comparer ces deux méthodes est l'*accuracy* globale (*Overall Accuracy, OA*) Les résultats de classification sont présentés dans le Tableau 3.2. La méthode proposée dans ces travaux ne permet pas d'obtenir des résultats supérieurs à la baseline supervisée, indiquée dans le Tableau 3.2 et du modèle à l'état de l'art, avec un écart de quatre points.

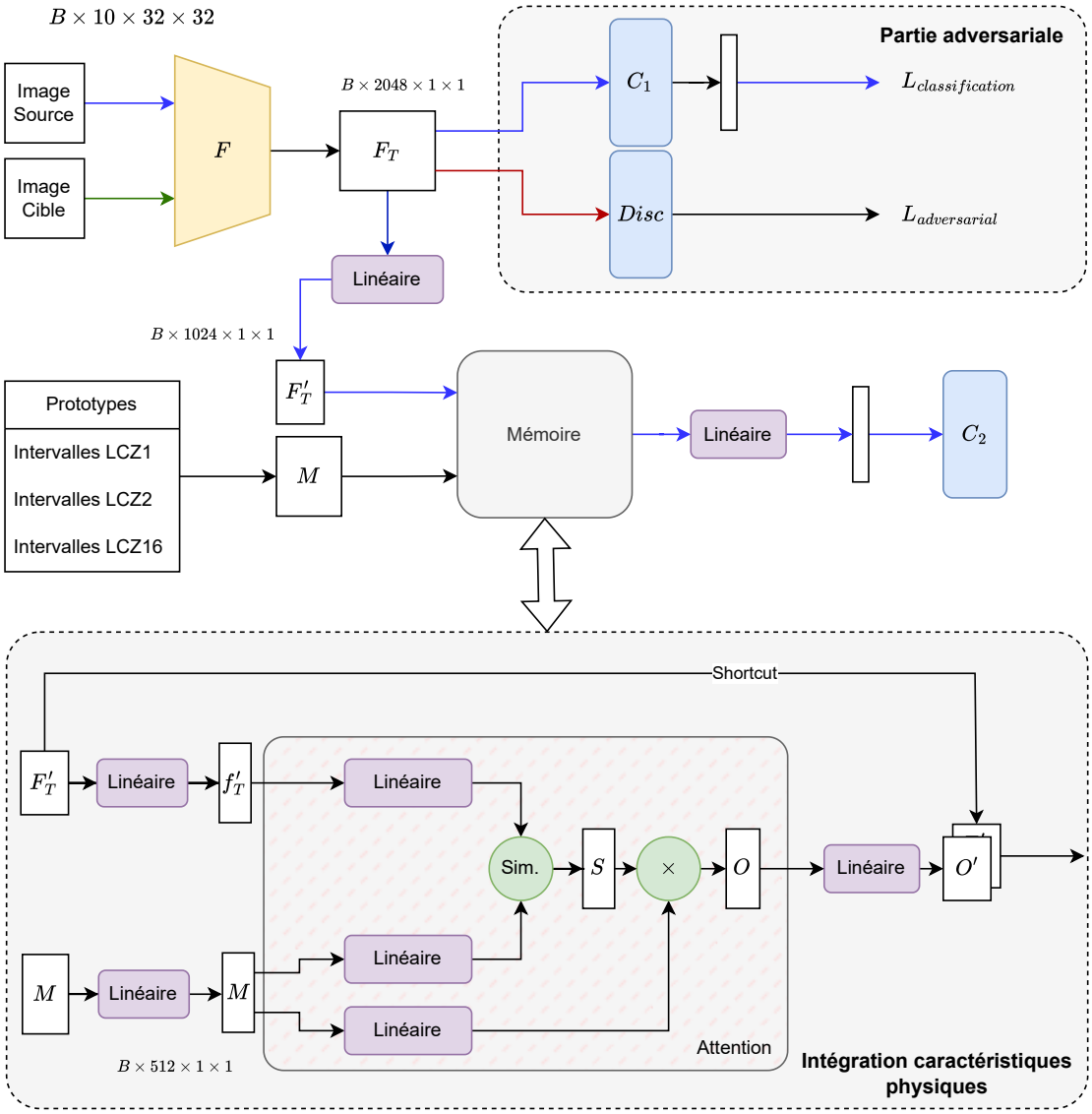


Figure 3.11: Vue d'ensemble de la méthode d'adaptation de domaine proposée. Cette méthode possède deux chemins: le premier utilise l'apprentissage adversarial pour aligner les distributions dans l'espace latent, et le deuxième introduit une représentation supposée universelle via les descripteurs physiques

Discussion

Dans cette sous-section, les valeurs de descripteurs physiques de LCZs ont été utilisés comme une représentation connue. Cette représentation devait permettre d'avoir une représentation optimale des classes, communes aux domaines source et cible. Comme supposée universelle, cette représentation n'est pas optimisée pendant l'apprentissage : seulement sa transformation dans l'espace latent l'est. Deux hypothèses ont donc été faites. La première hypothèse est qu'il est possible d'apprendre par rapport à des prototypes fixes, et non appris par le modèle. Les méthodes d'adaptation de domaine de l'état de l'art proposent de laisser ces prototypes libres afin que le modèle puisse les définir à partir de son extraction d'information. Dans notre cas, le modèle est contraint par les prototypes, supposés universels. La seconde hypothèse est que ces prototypes fixes peuvent être intégrés dans l'apprentissage avec un mécanisme d'attention. En calculant l'attention entre les caractéristiques extraites par le modèles et une transformation des prototypes, nous espérons guider l'extraction d'information pour qu'elle sépare les différentes classes dans l'espace latent selon notre représentation fixe.

Son intégration dans l'entraînement n'a pas donné les résultats escomptés. Deux options sont possibles pour expliquer ce déficit de résultat. Premièrement, l'intégration de cette représentation, la mémoire, ne permet pas de produire des prototypes références dans l'espace latent, communs aux représentations des deux domaines. Les descripteurs physiques sont d'abord envoyés vers un nouvel espace, dans lequel ils peuvent être comparés aux caractéristiques extraites par le modèle. Cette comparaison s'effectue via un mécanisme d'attention entre les deux vecteurs résultants, qui doit permettre d'aligner les caractéristiques extraites avec les caractéristiques correspondantes des descripteurs. La nouvelle représentation des descripteurs est calculée à l'aide d'une couche linéaire, qui peut ne pas offrir assez de liberté pour obtenir des représentations suffisamment différentes pour chaque classe. Ces représentations sont alors similaires et ne permettent pas de modifier l'espace latent du modèle pour séparer les classes.

Deuxièmement, la représentation par descripteur peut être universelle pour les humains, mais pas pour le modèle qui extrait des informations depuis l'image. Cependant, nous avons vu dans la Section 3.3.1 que le modèle arrivait à prédire ces descripteurs, donc arrive à extraire les caractéristiques essentielles à ces prédictions. Nous écartons donc l'hypothèse selon laquelle le modèle n'arrive pas à extraire cette représentation universelle pour les humains. La fixation stricte des prototypes peut empêcher le modèle d'apprendre les bonnes représentations. Cependant, les résultats en classification se rapprochent de ceux obtenus avec la méthode supervisée et le modèle de pointe, ce qui laisse penser que ces travaux mériteraient une analyse plus approfondie, qui n'a pu être réalisée ici en raison de contraintes temporelles.

3.4 Cartographie de la ville d'Antananarivo

Afin de pouvoir utiliser les LCZ comme un indicateur environnemental pour l'étude de la mortalité à Antananarivo, une carte de la ville doit être générée. Cette section présente l'entraînement du modèle pour la cartographie de la ville d'Antananarivo spécifiquement. Premièrement, nous verrons comment transférer l'apprentissage du modèle au contexte de la ville, puis nous verrons la génération de la carte.

3.4.1 Ajout d'Antananarivo dans la base d'entraînement

Pour rappel, le jeu d'entraînement *Cultural_10* que nous utilisons est composé de deux parties : une étiquetée qui représente notre domaine source et une autre, non étiquetée, qui représente notre domaine cible. Pour espérer améliorer les performances du modèle sur notre ville cible, Antananarivo, il est nécessaire d'ajouter de ces images de la ville dans le jeu d'entraînement. Pour cela, une tuile Sentinel-2 a été téléchargée depuis la plateforme du projet Copernicus[5], pendant la saison sèche de l'année 2019 pour correspondre à notre année d'intérêt et à l'année des autres images du jeu de données So2Sat. Les bandes avec une résolution de 60 m (B01, B09 et B10) ont été supprimées pour correspondre au modèle de So2Sat. Cette tuile a ensuite été découpée en images de taille $32\times32\times10$, soit une résolution au sol de 320m. Les bandes dont la résolution est de 20m sont re-échantillonnées à 10m avec une interpolation bicubique. Au total, 2493 images d'Antananarivo ont été ajoutées au jeu d'entraînement cible. Comme elles ne sont pas annotées, ces images ont été ajoutées au jeu d'entraînement **cible**, et seront aussi utilisées pour la cartographie, présentée dans la partie suivante.

3.4.2 Cartes de la ville

Dans cette sous-section, nous allons produire des cartes LCZ d'Antananarivo, d'une résolution de 320m à l'aide de plusieurs modèles entraînés. Pour cela, toutes les images Sentinel-2 de la ville, dont l'extraction a été détaillée en sous-section 3.4.1 sont classées par les modèles puis assemblées pour reconstituer les cartes. Nous utilisons les modèles suivants pour la production des cartes :

- LCZmemoryNet entraîné **sans** les images d'Antananarivo,
- LCZmemoryNet entraîné **avec** les images d'Antananarivo,
- modèle ensembliste par [15] entraîné **sans** les images d'Antananarivo,
- modèle ensembliste par [15] entraîné **avec** les images d'Antananarivo.

La Figure 3.12 présente les cartes ainsi produites.

Une évaluation visuelle des classes majoritaires pour chaque carte permet d'avoir une première idée de leur qualité. La carte générée par LCZmemoryNet sans avoir ajouté d'images d'Antanarivo dans l'entraînement prédit principalement les classes *Lightweight low-rise* et *Open low-rise* pour les régions urbaines, et la classe *Bare rock or paved* pour les zones de végétations. La carte générée par LCZmemoryNet avec ajout d'images indique principalement la classe *Lightweight low-rise* pour les zones urbaines et la classe *Bush, scrub* pour les zones rurales. Dans les deux cas, la représentation des zones urbaines n'est pas cohérente avec la réalité. Par exemple, les zones de bâtis légers sont principalement présents dans les premier et quatrième arrondissements (voir les bas quartiers indiqués en Figure 3.1), et non dispatchés

[5] https://scihub.copernicus.eu/

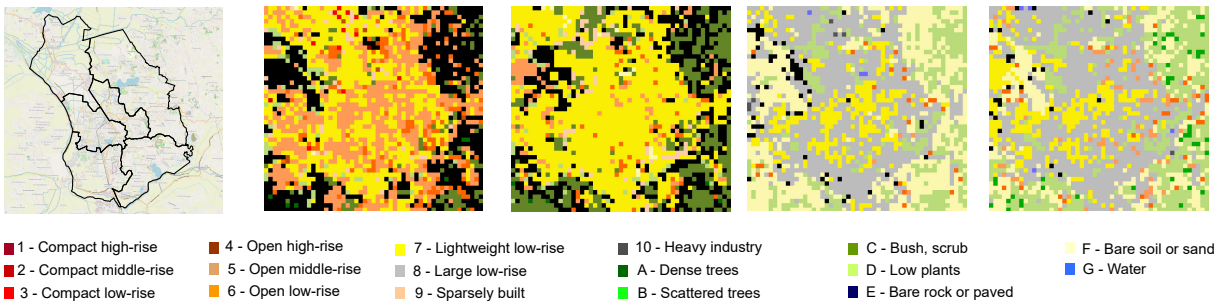


Figure 3.12: Cartes LCZ d'Antananarivo générées. De gauche à droite : OpenStreetMap, LCZmemoryNet sans les images supplémentaires, LCZmemoryNet avec les images supplémentaires, [15] sans les images supplémentaires, [15] avec les images supplémentaires.

dans toute la ville. Le même raisonnement peut être tenu sur les images générées avec le modèles introduit par Zhao *et al.* (2023) [15]. Les zones urbaines sont représentées par la classe *Large low-rise*, ce qui ne correspond pas à une ville comme Antananarivo qui est dense.

Par conséquent, les cartes ainsi générées ne peuvent pas être utilisées en lien avec la démographie, car elles apporteraient de fausses corrélations et pourraient par conséquent mener à des interprétations erronées. Pour la suite de cette partie, nous utiliserons la carte proposée par Demuzere *et al.* [59], qui couvre toute la surface terrestre. Cette carte est issue de forêts aléatoires entraînées sur des zones étiquetées par des experts ou dans des plateformes de crowdsourcing [57]. Une carte d'Antananarivo a pu être extraite de cette carte globale.

La section suivante présente l'étude de données de population de ces premiers travaux. L'idée est d'étudier l'effet de l'environnement, en utilisant plusieurs caractérisations, sur la mortalité et les causes de décès à Antananarivo. Pour cela, nous utilisons le registre des décès de la ville (présenté en section 3.2.2) combiné aux données du recensement national de population de 2018.

3.5 Étude du lien entre causes de mortalité et environnement

Cette section lie les indicateurs environnementaux récupérés ou générés aux données sur les causes de mortalité, tout en contrôlant les caractéristiques socio-économiques des quartiers à partir des informations disponibles au niveau des ménages et des individus dans le recensement de population. La section 3.5.1 présente le traitement qu'il a fallu effectuer sur les données brutes de mortalité pour pouvoir localiser le maximum de décès dans les quartiers de la ville. Ensuite, les sections 3.5.2 et 3.5.3 présentent les caractéristiques socio-économiques et environnementales d'Antananarivo, au niveau des fokontany. Enfin, les résultats et leur discussion sont présentés en sections 3.5.6 et 3.5.7.

3.5.1 Traitement des données de population

La première étape pour lier les données de population aux données environnementales au niveau des quartiers est de nettoyer les données démographiques, en l'occurrence celles portant sur le quartier de résidence des personnes décédées, pour les rendre exploitables. En effet, les données sont construites à partir des déclarations des proches, inscrites sur les formulaires papier dédiés, puis transcrites en format numérique. Or, des erreurs ou incohérences peuvent subvenir à chaque étape. Ainsi, lefokontany indiqué par les déclarants ou noté dans la base ne correspond pas toujours à la liste des fokontany que nous possédons (que ce soit celle fournie par le BMH ou celle qui vient de OCHA - United Nations Office for the Coordination of Humanitarian Affairs). Les noms des quartiers peuvent correspondre à des lieux dits, ou à des localités et de nombreuses fautes d'orthographe ou imprécisions subsistent. Au total, plus de 60 000 valeurs doivent être triées, dont environ 37 000 sur notre période d'intérêt entre 2016 et 2020. Cette section présente les étapes mises en places pour le formatage des données.

Correspondance	Adresse	Fokontany indiqué	Fokontany final
Exacte	III C 53	AMBANIN'AMPAMARINANA	Ambanin'ampamarinana
Groupée	/	67ha Sud - 67ha Ouest - 67ha	67ha
Orthographe	/	andoh**o**tapenaka	Andoh**a**tapenaka
Localité	/	FORT DUCHESNE	Andraisoro
SDF	/	/	/
Inconnue	/	/	/

Tableau 3.3: Exemple d'association de fonkontany suivant plusieurs situations. "Exacte" signifie que la correspondance est exacte entre le fokontany indiqué dans le questionnaire et le fokontany officiel. "Groupée" signifie que le fokontany indiqué appartient à un regroupement de fokontany officiels. "Orthographe" signifie qu'il y a une faute d'orthographe d'écart. "Localité" signifie que le fokontany indiqué est une localité, c'est-à-dire un lieu dit, et non un fokontany officiel.

Les décès de personnes identifiées comme sans domicile ("SDF" "Sans domicile" "Sans domicile fixe" dans la base) n'ont pas été pris en compte dans l'analyse. L'étude portant sur la population générale, cette population très spécifique pourrait biaiser les interactions entre la mortalité et l'environnement à l'échelle fine des quartiers.

Regroupement de certains fokontany

Comme indiqué dans l'introduction de cette partie, les noms de fokontany indiqués ne sont pas toujours complets. Pour pouvoir attribuer le maximum de décès et pour regrouper les fokontany comptant peu de décès, certains fokontany sont fusionnés. Un tableau de correspondance est fourni en annexe 6.1

Exemple 3 ***67ha :*** *La cité des 67 ha est une zone résidentielle située dans le premier arrondissement d'Antananarivo. Administrativement, il est séparé en quatre fokontany : 67 ha*

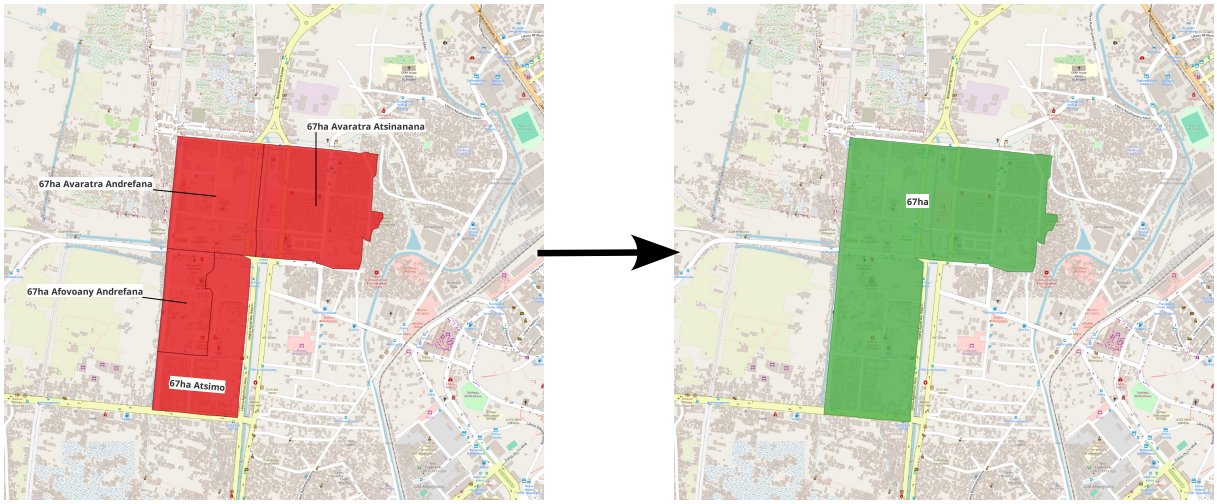


Figure 3.13: Exemple de regroupement des fokontany de la cité des 67 ha en un seul groupe. A gauche sont affichés les quatre fokontany officiels, et à droite leur regroupement.

Atsimo[6]*, 67 ha Avaratra-Andrefana*[7]*, 67 ha Avaratra-Atsinanana*[8]*, 67 ha Afovoany-Andrefana*[9]*. Les références à ces quartiers dans la base décès sont diverses : 67ha, 67ha n-o, 67ha avaratra et autres dénominations ne permettant pas toujours de conclure sur le fokontany précis. Ainsi, ces quatre quartiers ont été fusionnés en un fokontany dit "groupé" : 67ha. Ce regroupement est illustré en Figure 3.13.*

Traduction et correction de l'orthographe

Quelques corrections manuelles, lorsqu'elles sont évidentes, ont été effectuées au début du traitement. Ces corrections ont été complétées au fur et à mesure des associations, donc au fil de l'eau pendant toute la période d'analyse. Elles incluent des traductions de mots du français vers le malgache, ainsi que des corrections basiques d'orthographe et de fautes de frappe.

Exemple 4 ***Traduction :*** *"Nord" vers "Avaratra", "ne" vers "Avaratra Atsinanana".*
Frappe et orthographe : *"manrintsoa" vers "Manarintsoa", "anocibe" vers "Anosibe".*

Recherche des mots communs

Certaines dénominations indiquées sont incomplètes, mais l'orthographe est correcte. Cela arrive, en général, lorsque les noms officiels sont composés de plusieurs mots, mais un seul mot est indiqué dans la base. Nous faisons l'hypothèse que les arrondissements de résidence (numérotés de 1 à 5) sont indiqués correctement. Pour chaque décès, les noms indiqués sont

[6]traduction française: Sud
[7]traduction française: Nord-Ouest
[8]traduction française: Nord-Est
[9]traduction française: Centre-Ouest

comparés avec les noms officiels des fokontany de l'arrondissement concerné, afin de faire un premier filtre. Lorsqu'un nom indiqué ne comporte qu'un mot, qui est inclus dans un nom officiel, le fokontany officiel est stocké. Les fokontany sont directement associés lorsqu'il n'y a qu'un choix, et filtrés manuellement lorsqu'il y a plusieurs choix. Si aucune association n'est possible, nous effectuons le même processus dans les autres arrondissements et filtrons manuellement.

Recherche automatique sur OpenStreetMap

Après ces premières étapes, les lieux indiqués toujours non attribués sont :

- Ceux dont les fautes de frappe n'ont pas été prises en compte.
- Ceux qui ne possèdent pas de mots en commun avec les fokontany officiels. Ce sont les localités, ou les "lieux dits" : ils existent vraiment mais ne sont pas recensés dans les fokontany officiels.

L'API d'OpenStreetMap a été utilisé pour filtrer ces localisations. Pour chaque valeur restante, une recherche a été faite dans ces données. Si cette recherche aboutissait à un résultat, sa localisation a été comparée aux zones recouvertes par les fokontany officiels, et lorsqu'il y a recouvrement, le fokontany correspondant a été sélectionné. Ce filtrage est nécessaire car certains quartiers indiqués ont les mêmes noms que d'autres villes de Madagascar, qui sont indiquées en priorité par la recherche.

Recherche manuelle en ligne

Pour le reste, des recherches manuelles ont été faites, car les données OpenStreetMap sont incomplètes. De plus, ces recherches dépendent aussi de la qualité des fokontany indiqués. Plusieurs sites sont été utilisés : Google Maps, OpenStreepMap, Mapcarta ainsi que des recherches simples Google pour trouver des points d'intérêts. Si la localité est directement trouvée, le décès est associé au fokontany officiel la recouvrant. Si des points d'intérêts sont trouvés (comme des sites religieux ou des écoles), et que ces points sont spatialement proches, nous faisons l'hypothèse que la localisation correspondante a intentionnellement été indiquée dans le questionnaire. Le fokontany correspondant est associé dans la base de mortalité.

Au final, seuls 757 lieux indiqués sur les 37 000 décès de la période 2016 - 2020 n'ont pas pu être attribués, ce qui équivaut à un taux de non classification de 2%.

3.5.2 Données socio-économiques

La base des décès ne permet pas de collecter les informations relatives aux caractéristiques socio-économiques des quartiers. En revanche, ces données peuvent être récupérées via d'autres enquêtes, si celles-ci ont été effectuées à des dates proches de la période d'étude.

Le recensement de la population malgache, effectué en 2018, collecte des données au niveau des individus et des ménages sur les conditions d'habitat et les conditions de vie des membres du ménage, notamment du chef de ménage. Disponibles au niveau individuel grâce à un accord cadre entre l'INSTAT et l'INED, ces données permettent de produire des indicateurs socioéconomiques au niveau des fokontany. Nous estimons pour chaque fokontany :

- La part d'individus habitant dans des logements dont les murs sont en briques ou en parpaings (Part mur brique).
- La part d'individus ayant un accès privé à un robinet. Cet accès à l'eau peut être individuel (au sein du ménage) ou collectif (dans la cour) mais privé (Part robinet privé).
- La part d'individus ayant accès à des toilettes avec plateforme en béton lisse, en porcelaine ou en fibre de verre (Part toilettes améliorées).
- La part d'individus habitant dans un logement avec un système d'évacuation des eaux usées (fosse dans le logement ou égouts) (Part évacuation des eaux).
- La part d'individus ayant accès à l'électricité (Part électricité).
- La part d'individus ayant un véhicule motorisé (voiture, moto ou scooter, Part véhicules motorisés).
- La part d'individus dont le logement est équipé d'un appareil de cuisson électrique ou au gaz (Part cuisson améliorée).
- La part d'individus habitant dans un logement la part d'individus disposant d'un droit d'occupation de leur logement (titre foncier, enregistrement au cadastre, certificat de régularisation, en cours de régularisation, terre ancestrale sans titre, part droit terrain)
- La part d'individus vivant dans des ménages dont le chef a au moins un niveau d'éducation secondaire (part niveau secondaire).
- Le nombre d'individus dans le fokontany.

Les Figure 3.15 et 3.14 affichent les cartes pour chaque indicateur, au niveau des fokontany. Cette visualisation spatiale atteste des disparités entre les fokontany pour certains indicateurs socioéconomiques. Nous pouvons observer des différences selon deux axes. Premièrement, il y a des inégalités Est-Ouest, en particulier entre les quartiers résidentiels à l'Est et les bas-quartiers de la ville à l'Ouest, connus pour être les plus pauvres. Deuxièmement, nous observons des différences entre le centre de la ville, situé au niveau des collines et de la haute ville, et les périphéries. Ces différences proviennent de l'installation des populations les plus riches près du Palais de la Reine, sur le point culminant. En plus des indicateurs socioéconomiques, il faut pouvoir définir des indicateurs environnementaux à ce même niveau d'étude pour pouvoir tester l'impact de l'environnement sur la mortalité, en contrôlant ces variables socioéconomiques.

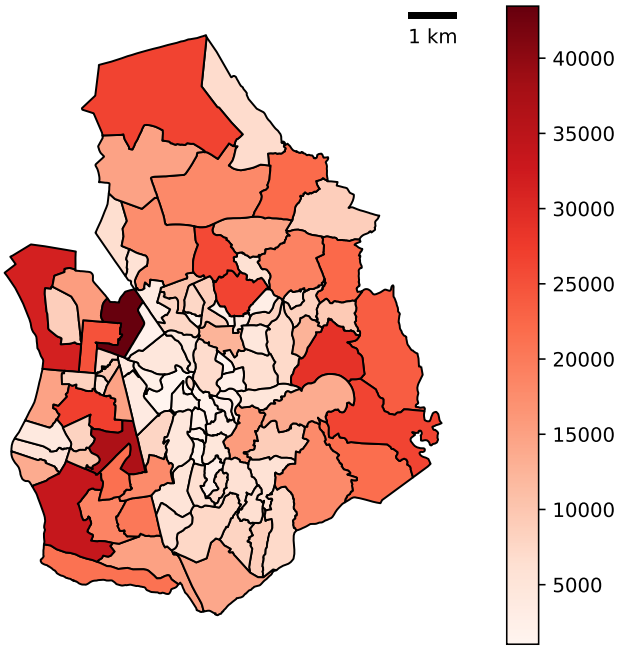


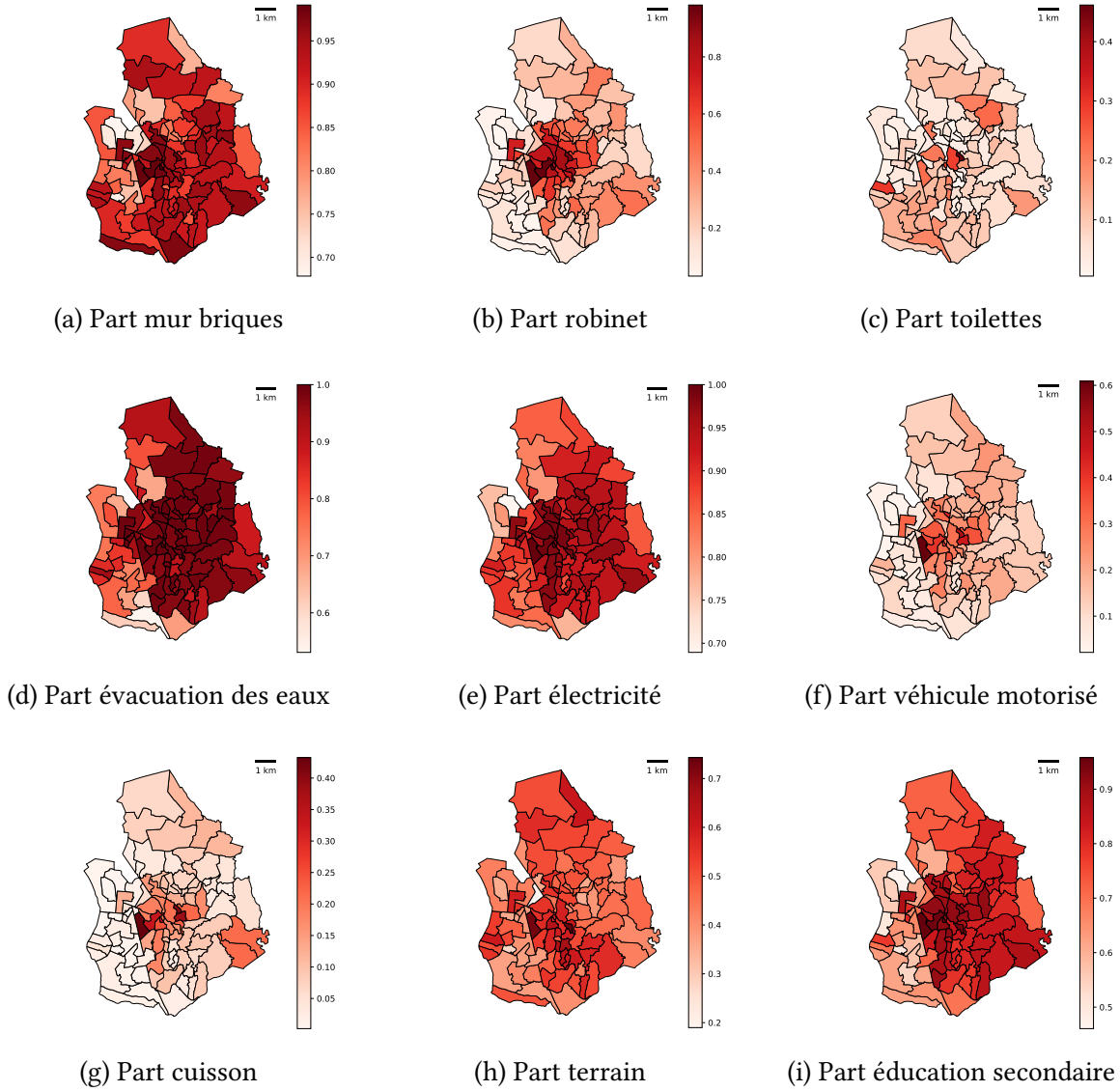
Figure 3.14: Nombre d'individus par fokontany à Antananarivo

3.5.3 Données environnementales

Cette sous-section présente le traitement des différents indicateurs environnementaux utilisés. Les indicateurs sont regroupés par fokontany pour correspondre à notre niveau d'étude.

Végétation

Le taux de végétation est calculé via le NDVI. Plus celui-ci est fort, plus la présence de végétation dans la zone est forte. Des cartes de valeurs de NDVI sont possibles à partir de bandes spectrales Sentinel-2, d'une résolution de 10 m. Il est donc possible de capturer la diversité intra-fokontany. La valeur moyenne permet d'indiquer la quantité moyenne de végétation par fokontany, et la variance nous indique l'éventuelle présence de "zones vertes" dans un quartier très urbain. L'utilisation de la variance permet de différencier certains quartiers ayant des valeurs moyennes identiques, mais qui correspondent à des environnements différents. Par exemple, à NDVI moyen égal, un quartier ayant une variance forte sera un quartier avec peu de végétation, mais caractérisé par la présence d'un parc, alors qu'une variance de NDVI faible est le signe que le degré de verdure est uniformément répartie sur toute la surface. La Figure 3.16 affiche les valeurs moyennes et de variance de NDVI par fokontany au niveau de la ville. Les bas quartiers de la ville, autour de la cité des 67ha, ont des taux moyens de végétation et une variance faible. Il y a donc peu de végétation dans ces quartiers, ni de parcs ou de zones vertes.



(a) Part mur briques (b) Part robinet (c) Part toilettes

(d) Part évacuation des eaux (e) Part électricité (f) Part véhicule motorisé

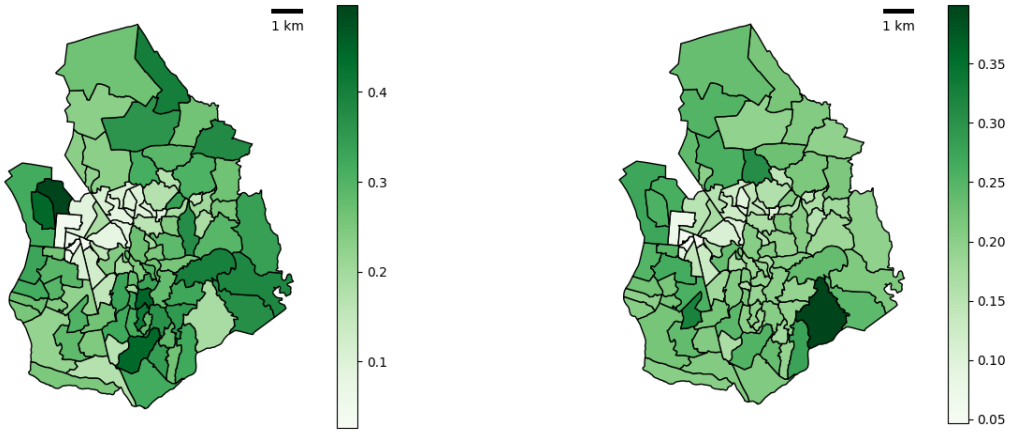
(g) Part cuisson (h) Part terrain (i) Part éducation secondaire

Figure 3.15: Répartition spatiale des indicateurs socio-économiques par fokontany.

Altitude

Antananarivo est située sur trois collines, donc les fokontany ont des altitudes moyennes très variées. Cela a un impact sur les conditions de vies des ménages, notamment par rapport aux risques d'inondations pendant les périodes de pluie. Ces conditions de vies plus difficiles en basse altitude ont d'ailleurs façonné les différences de richesse entre les fokontany, les plus bas étant souvent ceux habités par les populations les plus pauvres. Ces valeurs de l'altitude sont récupérables via des modèles numériques d'élévation, disponibles en libre accès[10]. De la même manière que pour les valeurs de NDVI, et étant donnée la diversité de l'altitude dans la ville, les valeurs moyennes et les variances sont calculées pour chaque fokontany. Les valeurs de la variance indiquent si un fokontany se situe en pente, et le différencient d'autres fokontany qui pourraient être situés à la même altitude moyenne, mais sur un plateau, ou différencient les

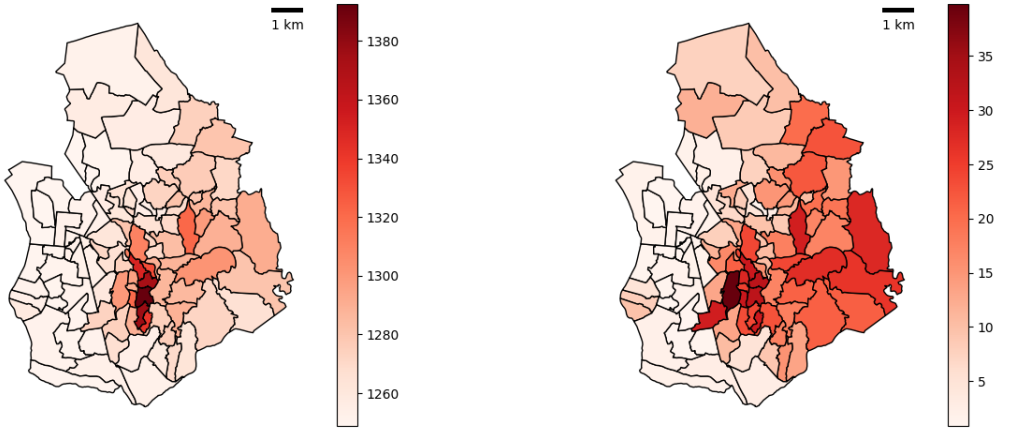
[10]https://www.earthdata.nasa.gov/



(a) NDVI moyen de chaque fokontany. (b) Variance du NDVI de chaque fokontany.

Figure 3.16: NDVI moyen (a) et variance (b) de chaque fokontany d'Antananarivo.

fokontany avec des pentes plus ou moins fortes. Les cartes d'altitude sont affichées en Figure 3.17. Nous pouvons observer une forte disparité entre l'Est de la ville, situé sur les collines, et l'Ouest, situé aux pieds de ces collines. Sur le point culminant, au centre de la ville, se situe le Palais de la Reine, entouré de quartiers riches. Les bas-quartiers, eux, se situent à l'ouest de la ville dans les altitudes les plus basses et sont soumis aux inondations. A l'est de la ville, les quartiers sont résidentiels.



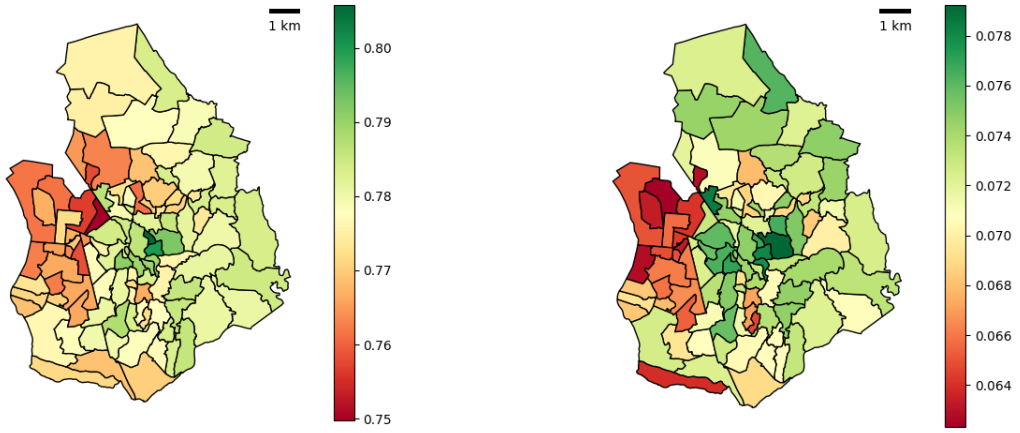
(a) Altitude moyenne de chaque fokontany. (b) Variance de l'altitude de chaque fokontany.

Figure 3.17: Altitude moyenne (a) et variance (b) de chaque fokontany d'Antananarivo.

Google Open Buildings

Google a généré des données de détection de bâtiments dans les pays du Sud, à partir d'apprentissage profond [77]. L'apprentissage des modèles a été fait sur un million d'images RGB de taille 600 × 600 et à très haute résolution (50 cm) autour des régions habitées du continent africain. Ces modèles ont ensuite été utilisés pour cartographier les bâtiments de tout

le continent. Pour ces travaux, nous utiliserons la confiance du modèle comme un indicateur de la structuration des bâtiments. L'hypothèse est que plus le bâtiment sera en matériaux légers, donc moins conventionnels, plus le modèle d'apprentissage profond aura du mal à le détecter, et donc sa confiance sera basse. Les valeurs de moyenne et de variance sont calculées par fokontany. La moyenne de la confiance nous donne une information, selon notre hypothèse, si les bâtiments dans le fokontany considéré sont construits "en dur" ou s'ils sont plutôt non conventionnels. Les fokontany d'Antananarivo ont des habitations très variées. Il n'est pas rare de voir des zones d'habitat informel à coté de zones plus pavillonnaires. En plus de la moyenne de la confiance, nous calculons aussi sa variance pour prendre en compte l'éventuelle diversité des zones d'habitations dans un même fokontany. Cette variable sera appelée Google O.B. dans le reste de ce manuscrit.



(a) Valeur de confiance moyenne.

(b) Variance de la valeur de confiance.

Figure 3.18: Valeur de confiance de Google Open Buildings : moyenne (a), variance (b) par fokontany.

Les trois cartes de la Figure 3.18 sont particulièrement intéressantes pour observer les tendances dans les différentes zones de la ville. La carte de droite, qui affiche le ratio du nombre de détections et de la surface du fokontany, indique que les quartiers les plus densément construits sont aussi les plus pauvres, au pied des collines. La confiance dans les bas quartiers, pauvres, est la plus faible, là où sont les zones de bidonvilles, ce qui valide l'hypothèse que nous avions faite. La variance est cependant aussi faible dans ces fokontany, ce qui indique qu'il y a peu de diversité dans les bâtiments.

Local Climate Zones

Dans cette sous-section, nous voulons mettre en valeur la morphologie de la ville en utilisant les LCZ. Pour cela, nous utilisons la carte générée par Demuzere *et al.* [59], qui a une résolution spatiale de 100 mètres, suffisante pour caractériser finalement l'environnement des fokontany. Pour pouvoir intégrer cette caractérisation environnementale dans une analyse de population, les fokontany sont groupés en clusters, déterminés en fonction de leur distribution de LCZ. Cette phase de regroupement permet d'associer deux fokontany qui ont une morphologie similaire, mais sont localisés à des endroits différents de la ville. Ces deux

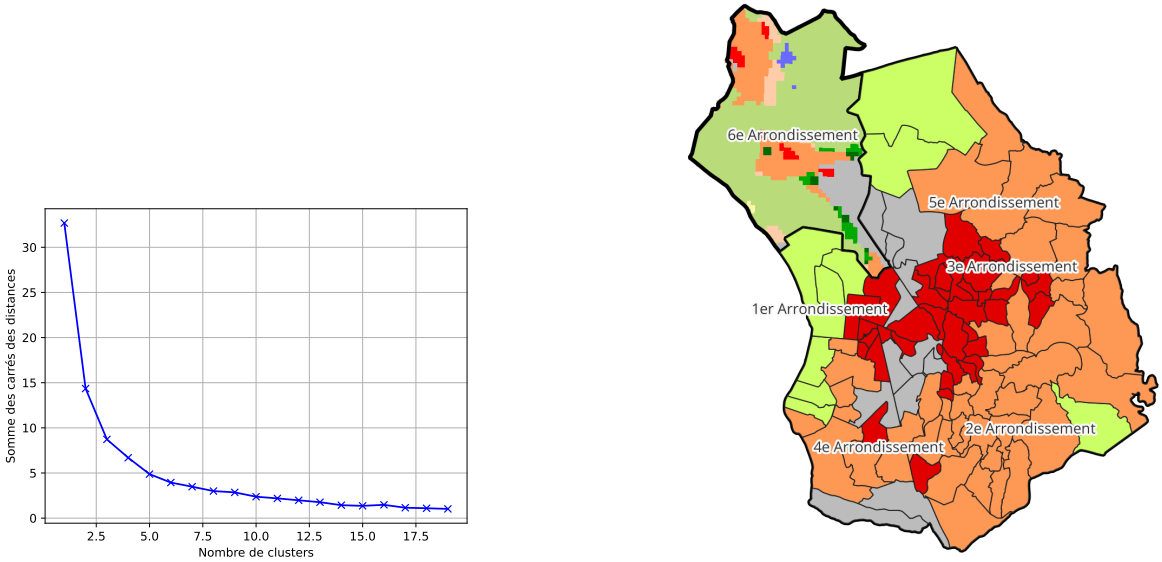
fokontany pourront cependant avoir des valeurs différentes en ce qui concerne les autres indicateurs environnementaux (NDVI, altitude). Séparer les différents indicateurs dans l'analyse démographique finale permettra de les étudier chacun indépendamment des autres.

Nous partitionnons les fokontany en fonction de leur distribution de LCZ. Ces distributions de LCZ sont obtenues après superposition des données de Global LCZ map [59] et du shapefile créé en section 3.5.1. Pour effectuer cette phase de clustering, nous appliquons l'algorithme *K-means* aux distributions des LCZ. La méthode du coude permet d'évaluer quels sont les nombres de clusters qui permettent de séparer au mieux les différents types d'environnements. Cette méthode consiste à faire une série de regroupement (*clustering*) en faisant varié le nombre de cluster. Pour chaque regroupement, un score de performance est calculé. Ici, la somme des distances quadratiques est calculée car donne un aperçu de la compacité des clusters. Après avoir calculé ces valeurs de performance, une évaluation visuelle permet le nombre de clusters à partir duquel le score de performance n'évolue plus significativement. La Figure 3.19 présente la courbe obtenue lors de recherche du nombre optimal de clusters, ainsi que les répartitions spatiales des fokontany pour le nombre optimal de clusters, quatre. La valeur du nombre de clusters optimale est donnée par la valeur au coude dessinée par la courbe, ici quatre. La Figure 3.20 présente les distributions de LCZ des centres des clusters, donc les distributions moyennes représentatives de chaque clusters. Les clusters représentent les quatre grandes catégories de morphologie présentes dans la ville. Le premier cluster, centré sur la classe *Compact low-rise* représente les zones les plus denses de la ville, incluant les bas quartiers et une partie du troisième arrondissement. Le second cluster est centré sur la classe *Open low-rise* et représente principalement les zones résidentielles des second, quatrième et cinquième arrondissements. Le troisième cluster est dominé par la classe *Large low rise*. Cette classe représente les bâtiments qui ne correspondent généralement pas à des habitations. Notons que dans ce cluster, il y a une importance certaine de la classe *Compact low rise*, qui peut correspondre à des zones résidentielles autour de ces grands bâtiments. Le dernier cluster est dominé par la classe *Low plants*. Ce cluster correspond aux périphéries de la ville, dans lesquelles il y a mélange d'habitations et de cultures, comme des rizières. Ces clusters sont très distincts les uns des autres et permettent de caractériser les différentes zones de la ville. Ils sont intégrés comme une variable catégorielle.

3.5.4 Corrélations des variables explicatives

Les cartes des figures 3.15, 3.16, 3.17, 3.18 ne nous indiquent pas que des disparités entre les quartiers de la ville. Certaines cartes sont très similaires : "Part cuisson améliorée", "Part véhicule motorisé", "NDVI moyen" et "NDVI variance", laissant supposer des corrélations entre les indicateurs affichés. Les modèles linéaires généralement utilisés en statistiques, pour établir des corrélations, requierent une indépendance entre les variables explicatives. Par conséquent, il faut comparer et trier les variables que nous possédons afin de valider cette hypothèse d'indépendance. La Figure 3.21 affiche la matrice de corrélation des variables explicatives présentées. Nous fixons arbitrairement le seuil de corrélation à **0,7**, le seuil à partir duquel nous considérons que deux variables sont trop fortement corrélées. Dans cette configuration, nous avons les corrélations suivantes :

- "Part Robinet" corrélée avec "Part électricité", "Part véhicule motorisé" et "Part cuisson",



(a) Méthode du coude

(b) Représentation spatiale des clusters

Figure 3.19: Présentation visuelle du choix des clusters avec la méthode du coude (a) et leur représentation spatiale (b). Les distributions moyennes des LCZ par cluster sont données en Figure 3.20. La couleur associée du cluster est celle de sa classe la plus représentée. La composition des clusters est présentée en Figure 3.20, et la légende couleur des LCZ en Figure 2.5.

- "Part niveau secondaire" corrélée avec "Part évacuation des eaux", "Part électricité", "Part Véhicule motorisé" et "Part cuisson".
- "NDVI moyen" corrélée avec "NDVI Variance", "DEM moyen" avec "DEM Variance" et "Google O.B. confiance" avec "Google O.B. Variance".

Les variances des variables environnementales sont souvent corrélées avec leurs moyennes : plus les moyennes sont élevées, plus les variances augmentent également. Prenons l'exemple du calcul du NDVI pour deux images d'une zone urbaine. Une augmentation du NDVI est principalement due à la présence de végétation, comme des arbres ou des espaces verts. Ces zones végétalisées augmentent localement les valeurs du NDVI, ce qui fait monter la moyenne du NDVI pour l'ensemble de la zone. En même temps, cela augmente la variance, car les pixels correspondant à la végétation se distinguent de leurs voisins, qui sont peu ou pas végétalisés. Les variances ne peuvent donc pas être utilisées. Dans le modèle mettant en lien les caractéristiques environnementales et la mortalité, nous utilisons :

- La part des individus habitant dans un logement pourvu d'un système d'évacuation des eaux usées,
- la part d'individus d'individus ayant un véhicule motorisé,
- la part d'individus vivant dans un ménage où le chef a au moins le niveau secondaire,
- la population ou nombre d'individus,

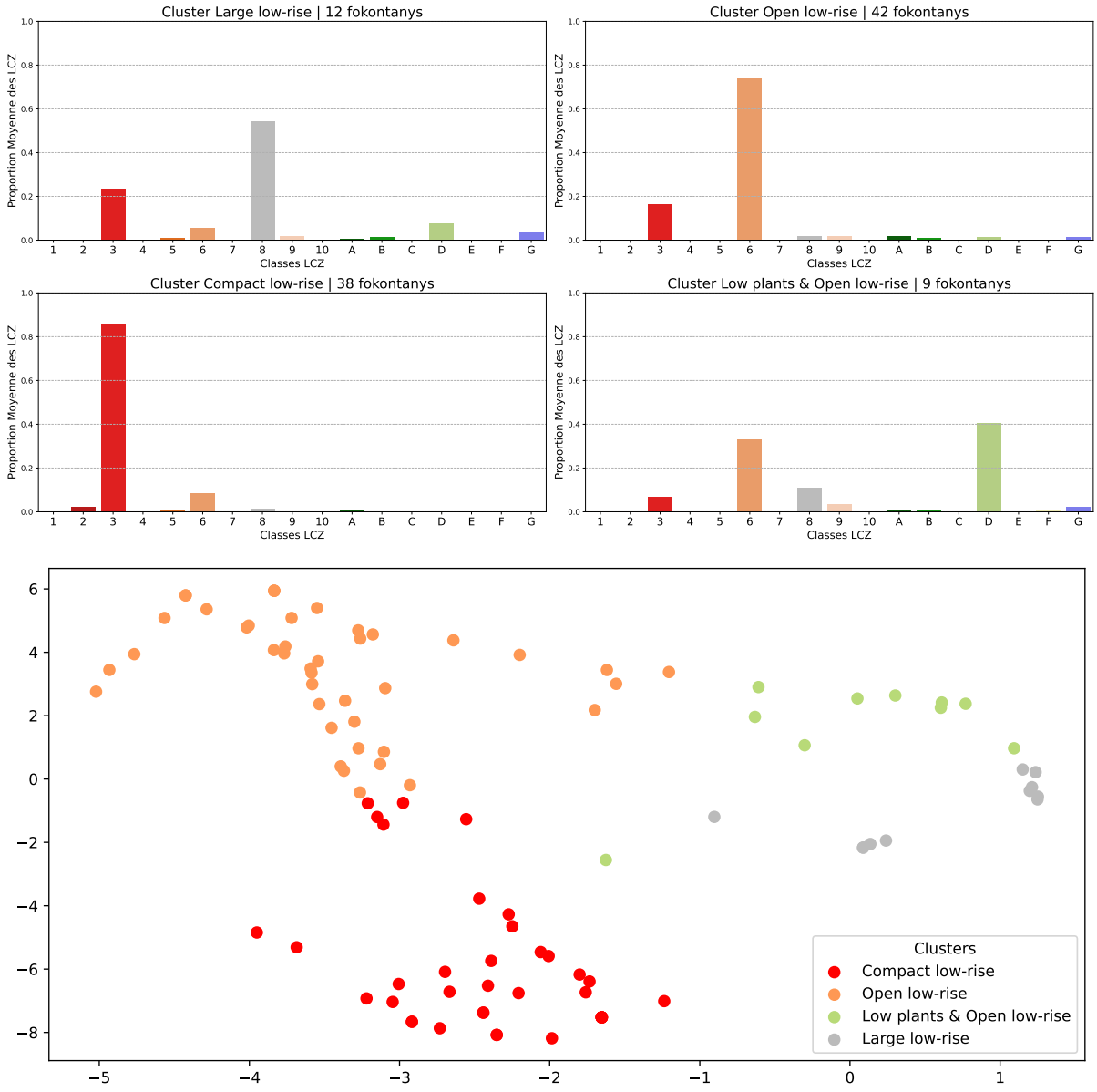


Figure 3.20: Distribution du centre pour chaque cluster de LCZ et représentation spatiale utilisant le t-sne [78]. Deux clusters sont dominés par une seule classe, et les deux autres sont plus diversifiés. La légende des LCZ est donnée en Figure 2.5.

- le NDVI moyen,
- l'altitude moyenne (DEM moyen),
- la confiance du modèle Google open building (Google O.B.),
- la morphologie de la ville représentée par les clusters LCZ.

3.5.5 Variables dépendantes

Le niveau de mortalité est appréhendé selon différents indicateurs estimés à partir des décès de la base du BMH et des effectifs de population par fokontany du recensement. Comme les effectifs de décès par âge peuvent être faibles au niveau des quartiers, nous avons eu recours à une méthode de lissage développée par Martin *et al.* (2024) [79], qui tient compte des niveaux de mortalité aux âges adjacents et des niveaux de mortalité des quartiers adjacents tout en laissant de la flexibilité pour permettre à chaque quartier d'avoir une mortalité distincte des quartiers voisins. A partir de cette méthode, ont été calculés pour chaque quartier:

- L'espérance de vie à la naissance i.e. durée de vie moyenne de la population du quartier,
- le quotient de mortalité des enfant de moins de cinq ans i.e. la probabilité pour un enfant né vivant de mourir avant son cinquième anniversaire.

Pour approcher au mieux la mortalité liée aux maladies les plus enclines à être dépendantes de l'environnement, nous avons estimé des taux de mortalité pour plusieurs groupes de causes de décès, en regroupant les codes CIM-10 des maladies suivantes : - les maladies liées à l'eau : choléra (A00), diarrhée (A01-A09), encéphalite liée aux moustiques (A83), zika, chikungunya et autre maladies liées aux moustiques (A92), fièvre jaune (A95), dengue (A97), paludisme (B50-B54), bilharzioses (B65), onchocercose, filariose (B73-B74) - les maladies respiratoires aiguës : tuberculose pulmonaire (A15-A19), diphtérie (A36), coqueluche (A37), grippe (J09-11), pneumonie (J12-J18), covid (U07 ou B97.2), SRAS (U04) - les maladies respiratoires chroniques : J40-J47 : bronchites, emphysèmes, BPCO, asthme Ces taux ont été standardisés par rapport à la structure par âge de la population de l'ensemble de la ville, ainsi les différences de niveaux entre quartiers ne dépendent que de la mortalité et non de différence dans la structure par âge.

A partir de cette méthode, ont été calculés pour chaque quartier:

- le taux de mortalité (standardisé) lié aux maladies liées à l'eau,
- le taux de mortalité (standardisé) lié aux maladies respiratoires aiguës,
- le taux de mortalité (standardisé) lié aux maladies respiratoires chroniques.

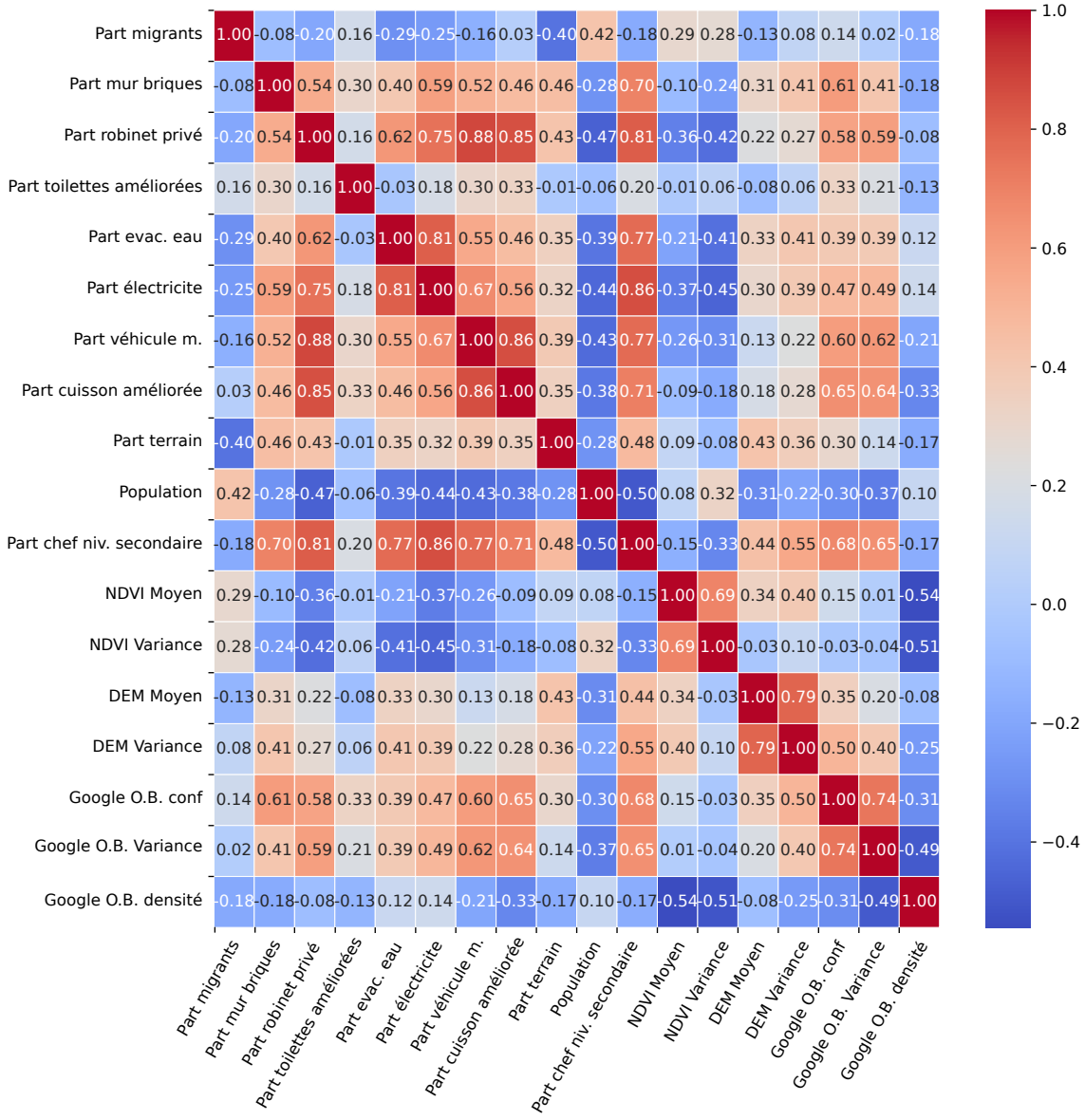


Figure 3.21: Matrice de corrélation des variables explicatives présentées à Antananarivo. En rouge sont les variables corrélées et en bleu les variables inversement corrélées.

3.5.6 Liens environnement et population

Modèles

Régression linéaire: Méthode des moindres carrés (OLS). Cette méthode permet de modéliser une relation linéaire entre une variable dépendante y et plusieurs variables explicatives indépendantes les unes des autres. L'objectif est de trouver la droite (ou l'hyperplan dans le cas multidimensionnel) qui minimise le carré de l'écart entre les valeurs observées et les valeurs prédites. L'intérêt de ce type de modèle est qu'il permet à la fois de prendre en compte des variables discrètes et continues : les clusters LCZ pourront être utilisés avec les autres variables continues. Pour cela, le modèle estime l'impact d'un changement de catégorie, par rapport à une catégorie de référence, sur la variable dépendante. Considérons un modèle de régression linéaire multiple avec p variables explicatives, comprenant à la fois des variables continues et des variables catégorielles encodées. L'équation du modèle est donnée par :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

où :

- y est la variable dépendante,
- β_0 est l'ordonnée à l'origine (intercept),
- $\beta_1, \beta_2, \ldots, \beta_p$ sont les coefficients de régression associés aux variables explicatives,
- $x_1, x_2, \ldots, x_p$ sont les variables explicatives, comprenant à la fois des variables continues et des variables indicatrices pour les variables catégorielles,
- ϵ est l'erreur aléatoire.

Régression logistique multinomiale. Cette régression est une extension de la régression logistique à lorsque la variable à expliquer possède plus de deux modalités. Elle permet de prendre en compte d'éventuels effets non linéaires, qui ne peuvent pas être considérés dans le modèle des moindres carrés. La régression logistique simple permet d'expliquer une variable binaire (comme la présence ou l'absence d'une caractéristique donnée) en fonction d'autres variables catégorielles indépendantes entre elles. De la même manière que pour les moindres carrés, le modèle va estimer l'impact d'un changement de catégorie, pour chaque variable, par rapport à une catégorie référence.

Pour chaque catégorie k (où $k = 1, 2, \ldots, K_{ref}, \ldots, K-1$), nous avons l'équation suivante, avec K_{ref} la catégorie de référence :

$$\log\left(\frac{P(y=k)}{P(y=K_{ref})}\right) = \beta_{0k} + \beta_{1k} x_1 + \beta_{2k} x_2 + \ldots + \beta_{pk} x_p$$

où :

- $P(y = k)$ est la probabilité que la variable dépendante y prenne la valeur k,
- $P(y = K_{ref})$ est la probabilité de la catégorie de référence K_{ref},
- β_{0k} est l'ordonnée à l'origine pour la catégorie k,
- $\beta_{1k}, \beta_{2k}, \ldots, \beta_{pk}$ sont les coefficients de régression associés aux variables explicatives pour la catégorie k,
- $x_1, x_2, \ldots, x_p$ sont les variables explicatives, comprenant à la fois des variables continues et des variables indicatrices pour les variables catégorielles.

Interprétation. Ces deux types de régressions permettent de mesurer l'impact "toute chose égale par ailleurs" de chaque variable environnementale et socio-économique dans un même modèle. Pris indépendamment les uns des autres, les coefficients associés à chaque variable modélisent son impact sur la variable dépendante, en ayant fixé les autres variables dépendantes. Le coefficient indique la force du lien, et la p-valeur indique si la relation est significative, ou non. En général, une p-valeur inférieure à 0.05 permet d'affirmer que la relation est significative. Ce type de modèle permet donc de comparer des ménages ayant des caractéristiques socioéconomiques similaires, mais avec des caractéristiques environnementales différentes, et inversement. Cela permet de confirmer ou d'infirmer l'importance de la relation entre population et environnement en prenant en compte la dimension socio-économique qui peut interagir entre les deux dans la mesure où les plus pauvres s'installent généralement dans les zones les plus défavorisées du point de vue environnemental.

Cette partie présente les résultats obtenus avec la méthode des moindres carrés. Les Tableaux 3.4, 3.5, 3.6, 3.7, 3.8 présentent les résultats de régression pour l'espérance de vie à la naissance, le quotient de mortalité des enfants de moins de 5 ans, le taux de mortalité des maladies liées à l'eau, le taux de mortalité des maladies respiratoires aiguës et le taux de mortalité des maladies respiratoires chroniques. Les variables explicatives dont le lien avec la variable dépendante est significatif sont indiquées en gras dans les tableaux. Les trois parties dans les tableaux représentent trois modèles : une régression multionimale avec uniquement les variables socioéconomiques, une régression multionimale avec les variables environnementales et le modèle linéaire avec toutes les variables explicatives.

Espérance de vie à la naissance

Cette sous-section présente les régressions de l'espérance de vie à la naissance. Les résultats sont affichés dans le Tableau 3.4. Cette variable est significativement corrélée à l'altitude moyenne des fokontany, et ce lien est positif d'après la méthode des moindres carrés : en contrôlant les autres facteurs, plus les quartiers sont situés en altitude, plus l'espérance de vie à la naissance a tendance à augmenter. Certains environnements LCZ sont aussi plus propices à une meilleure espérance de vie. Les populations des quartiers caractérisés par un environnement *Open low-rise* ont tendance à avoir une meilleure espérance de vie que les populations vivant dans des quartiers caractérisés par *Large low-rise*. Cette association est presque significative pour le changement d'environnement *Large low-rise* vers un mélange *Low plants / Open low-rise*. Mais elle n'est pas significative lorsque l'on passe de la catégorie de référence

	Tercile	Tercile 2 vs 1		Tercile 3 vs 1		Tercile 2 vs 1		Tercile 3 vs 1		OLS	
		Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur
Intercept	3.0	-2.6078	0.003	-1.9130	0.023	-2.9975	0.016	-2.6435	0.037	-95.0572	0.185
Part d'évac eau	2.0	-1.0522	0.240	-1.5979	0.078	-1.5999	0.153	-2.2400	0.053	3.8481	0.658
	3.0	-0.6066	0.553	-1.1278	0.251	-0.4890	0.690	-1.3550	0.272		
Part de véhicules motor.	2.0	0.5931	0.452	0.5379	0.500	0.4549	0.632	1.1565	0.273	-1.0174	0.904
	3.0	0.9034	0.385	0.3781	0.716	0.8330	0.494	1.0352	0.439		
Nombre d'individus	2.0	**1.6961**	0.036	**1.7705**	0.019	**2.1291**	0.028	**2.9495**	0.003	-2.971e-05	0.653
	3.0	**2.2509**	0.008	0.7863	0.361	**2.2415**	0.017	0.9717	0.322		
Part de chef niv. sec.	2.0	1.8844	0.057	**2.2639**	0.023	0.2440	0.850	1.1321	0.430	-3.181	0.784
	3.0	2.4686	0.058	**2.9896**	0.023	0.3539	0.827	1.5875	0.371		
NDVI moyen	2.0					0.2715	0.780	1.7647	0.098	11.3651	0.166
	3.0					1.0463	0.331	**2.7838**	0.026		
DEM moyen	2.0					1.5209	0.128	2.**1283**	0.070	**0.0997**	0
	3.0					1.3206	0.275	**3.6102**	0.007		
Google O.B.	2.0					1.3149	0.182	-0.1418	0.898	46.6101	0.588
	3.0					1.4422	0.249	-0.8111	0.539		
Cluster LCZ	LCZ 3					-0.2909	0.772	-1.3120	0.222	-4.5064	0.087
	LCZ 6					-0.6448	0.592	**-3.6441**	0.009	-2.9999	0.126
	LCZ 6 & LCZ 14					-1.5858	0.238	**-3.3897**	0.021	**-4.5466**	0.034

Tableau 3.4: Résultats de régressions logistiques et de la méthode des moindres carrés (OLS) pour expliquer l'espérance de vie à la naissance.

Large low-rise à la catégorie *Compact low-rise*. Aucune des associations avec les caractéristiques socio-économiques n'est significative.

Cependant, d'autres variables explicatives deviennent significatives lorsque l'on utilise la régression multinomiale, permettant de mettre en évidence des effets non linéaires. Le nombre d'individus des fokontany est significatif pour tous les quartiles et à presque tous les niveaux, prenant pour référence le premier tercile, et ces relations sont toutes positives. Les populations tendent à avoir une espérance de vie plus élevée dans les quartiers plus peuplés. De même, la part d'individus vivant dans un ménage dont le chef a atteint le niveau secondaire a une relation positive et significative pour le tercile 3 d'espérance de vie, lorsque les variables environnementales ne sont pas considérées. Les p-valeurs sont aussi très proches de la significativé (0.057 et 0.058) pour le tercile 2, ce qui indique un fort effet de cette variable. Ainsi, le modèle met en évidence que l'espérance de vie a tendance à être plus élevée quand la proportion d'individus vivant dans un ménage dont le chef a au moins le niveau secondaire est plus élevée aussi. Lorsque les variables environnementales sont intégrées au modèle, l'effet de l'éducation disparaît au profit du NDVI, de l'altitude et de la morphologie du fokontany représentée par les LCZ. Dans les hautes valeurs (tercile 3), l'altitude et la végétation ont des effets positifs sur l'espérance de vie. A l'inverse, les morphologies *Open low-rise* et *Low plants* ont un effet négatif. Les résultats des différents modèles donnent des résultats qui sont cohérents, en ce qui concerne le sens des relations entre les différentes variables explicatives et l'espérance de vie à la naissance.

Quotient de mortalité des enfants de moins de cinq ans

Cette sous-section présente les résultats des régressions pour le quotient de mortalité des enfants de moins de cinq ans, dans le Tableau 3.5. De la même manière que pour l'espérance de vie, la relation entre le quotient de mortalité et l'altitude est significative, mais négative (ce qui est attendu puisque cet indicateur de mortalité est inverse à celui de l'espérance de vie) selon la méthode des moindres carrés. Les quartiers les plus bas ont tendance à avoir des quotients de mortalité plus élevés, mais le coefficient est très faible, ce qui indique que la relation est

	Tercile	Tercile 2 vs 1		Tercile 3 vs 1		Tercile 2 vs 1		Tercile 3 vs 1		OLS	
		Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur
Intercept	3.0	-0.6820	0.433	2.0016	0.021	-0.4120	0.745	2.1049	0.094	0.2980	0.009
Part d'évac eau	2.0	-0.0190	0.983	1.9685	0.046	0.1690	0.867	2.2883	0.052	-0.0056	0.678
	3.0	0.7673	0.389	1.5719	0.142	1.1220	0.276	1.5096	0.232		
Part de véhicules motor.	2.0	0.1774	0.834	-0.6458	0.423	-0.1989	0.838	-0.8612	0.396	0.0159	0.229
	3.0	0.1033	0.917	-0.9482	0.364	-0.7276	0.533	-1.4708	0.252		
Nombre d'individus	2.0	0.1602	0.803	**-1.7640**	0.027	-0.3629	0.634	**-2.7701**	0.006	-3.948e-08	0.701
	3.0	1.2417	0.120	-0.9120	0.292	0.9720	0.288	-1.1824	0.224		
Part de chef niv. sec.	2.0	-0.3763	0.731	**-2.6236**	0.012	0.1197	0.932	-1.0937	0.442	-0.0112	0.538
	3.0	-0.1224	0.923	**-3.1674**	0.020	0.2758	0.860	-1.4503	0.410		
NDVI moyen	2.0					-0.8032	0.402	-0.9731	0.343	-0.0180	0.160
	3.0					-1.2089	0.295	-1.7253	0.148		
DEM moyen	2.0					-1.7013	0.163	-2.0498	0.098	**-0.0001**	0.005
	3.0					**-3.4448**	0.011	**-3.7032**	0.008		
Google O.B.	2.0					1.6280	0.137	-0.2304	0.834	-0.1559	0.247
	3.0					2.0851	0.107	0.3928	0.757		
Cluster LCZ	LCZ 3					0.8702	0.395	1.9553	0.078	0.0069	0.094
	LCZ 6					1.9408	0.129	**3.1928**	0.018	0.0048	0.117
	LCZ 6 & LCZ 14					1.4445	0.322	**3.1628**	0.031	**0.0068**	0.042

Tableau 3.5: Résultats de régressions logistiques et de la méthode des moindres carrés (OLS) pour expliquer le quotient de mortalité avant 5 ans.

faible. Les mêmes remarques pour les environnements LCZ peuvent être tirées de ces résultats: par rapport à *Large low-rise*, la relation est significative lors du passage à *Open low-rise* et presque significative lors du passage à *Low plants / Open low-rise*. Encore une fois, les caractéristiques socio-économiques ne sont pas significativement liées à la variable dépendante, lorsque le modèle est linéaire.

De nouvelles significativités apparaissent avec le modèle non linéaire. Lorsque les variables environnementales ne sont pas prises en compte, la part d'individus vivant dans un ménages dont le chef a un niveau secondaire a un effet négatif sur le quotient de mortalité, pour le tercile 3 : les fokontany où les habitants vivent dans les ménages dont le chef a un haut niveau d'éducation ont des quotients plus bas.

Lorsque l'environnement est pris en compte, encore une fois, la part d'individus vivant dans un ménage dont le chef a un niveau secondaire n'est plus significative, alors que les niveaux d'altitude les plus hauts le deviennent: les fokontany les plus hauts ont une mortalité plus faible. De même, pour ce même tercile, plus la population habite dans des quartiers en altitude élevée, plus la mortalité est forte. Les fokontany *Low plants* et *Open low-rise* ont une relation significative et positive pour le tercile de mortalité le plus haut, par rapport au tercile de mortalité le plus bas.

Taux de décès par maladies liées à l'eau

Cette sous-section présente les résultats des régressions pour le taux de décès par maladies liées à l'eau, dans le Tableau 3.6. Dans le modèle linéaire, seul le facteur éducation est significatif. Les fokontany où les chefs de ménage sont les plus éduqués tendent à être moins touchés par ce type de maladies. Nous sommes surpris d'observer que dans le modèle linéaire l'altitude n'est pas significative alors que les bas-quartiers, à basse altitude, sont souvent soumis aux inondations, et donc propices au développement de ce type de maladies.

Lorsque nous considérons un modèle non linéaire, et seulement les caractéristiques socio-économiques, le niveau d'éducation reste significatif.

	Tercile	Tercile 2 vs 1		Tercile 3 vs 1		Tercile 2 vs 1		Tercile 3 vs 1		OLS	
		Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur
Intercept	3.0	0.9011	0.349	1.6183	0.085	0.4363	0.736	1.0549	0.412	0.0003	0.312
Part d'évac eau	2.0	0.4008	0.659	2.3515	0.077	0.6612	0.540	2.8770	0.057	1.44e-05	0.665
	3.0	0.7300	0.412	2.0827	0.151	1.2640	0.233	2.6318	0.123		
Part de véhicules motor.	2.0	-0.2783	0.742	-1.0509	0.222	-0.0924	0.928	-0.9858	0.373	2.40-05	0.454
	3.0	0.0566	0.954	-0.5696	0.609	-0.3017	0.799	-0.8306	0.545		
Nombre d'individus	2.0	-0.7247	0.238	-0.4134	0.627	-0.9499	0.196	-0.6523	0.497	-1.784e-10	0.480
	3.0	-2.6943	0.005	-0.2871	0.751	**-3.3364**	0.005	-0.5751	0.583		
Part de chef niv. sec.	2.0	-0.3450	0.765	**-4.1183**	0.003	1.0959	0.531	-2.2592	0.174	**-0.0001**	0.005
	3.0	-0.8194	0.545	**-3.7329**	0.014	-0.4572	0.817	-2.4931	0.181		
NDVI moyen	2.0					1.5081	0.167	1.0389	0.352	5.398e-05	0.086
	3.0					2.4698	0.076	1.5956	0.236		
DEM moyen	2.0					-1.0443	0.448	**-2.6988**	0.022	-1.543e-07	0.104
	3.0					-3.**3897**	0.023	**-4.2471**	0.003		
Google O.B.	2.0					-0.5760	0.597	0.3902	0.728	0.0001	0.686
	3.0					1.0034	0.418	1.2850	0.388		
Cluster LCZ	LCZ 3					1.2883	0.263	1.2832	0.272	2.337e-06	0.815
	LCZ 6					-0.6396	0.666	-0.2575	0.843	-1.186e-06	0.873
	LCZ 6 & LCZ 14					-2.4511	0.260	0.5541	0.708	-1.292e-05	0.113

Tableau 3.6: Résultats de régressions logistiques et de la méthode des moindres carrés (OLS) pour expliquer la mortalité liée aux maladies liées à l'eau.

En revanche lorsque les variables environnementales sont ajoutées à l'équation, les mêmes effets que les résultats précédents sont observés. Le niveau d'éducation n'est plus significatif, tandis que l'altitude le devient, lorsque celle-ci est haute (tercile 3) et lorsque le niveau de mortalité est haut (tercile 3).

Taux de maladies respiratoires aiguës

Cette sous-section présente les résultats des régressions pour le taux de décès par maladies respiratoires aiguës, dans le Tableau 3.7. Le modèle linéaire indique qu'aucune variable n'est significative. Cependant, le cluster LCZ *Compact low-rise* semble avoir un léger effet positif (les fokontany de ce cluster ont une mortalité plus élevée, par rapport à la référence), mais la relation n'est que proche de la significativité (p-valeur de 0.094).

Pour les modèles multinomiaux, les résultats sont un peu différents. Sans les variables environnementales, comme pour tous les modèles précédents, seule l'éducation est significative, avec une relation négative. L'altitude devient significative lorsqu'on ajoute les caractéristiques environnementales à tous les niveaux de mortalité et pour tous les terciles d'altitude. Le NDVI, qui indique la présence de la végétation, n'est pas significatif, alors que sa présence est souvent associée à une meilleure qualité de l'air. Dans ce paramétrage, la morphologie de la ville n'est pas non plus significative.

Taux de maladies respiratoires chroniques

Cette sous-section présente les résultats des régressions pour le taux de décès par maladies respiratoires chroniques, dans le Tableau 3.8.

Aucune des variables indépendantes ne présente une p-valeur inférieure au seuil de 0.05 dans le modèle linéaire. Cependant, la confiance du modèle Google O.B. est proche de la significativité, avec une p-valeur de 0.062. Le coefficient associé est positif. Selon notre hypothèse par rapport à cette variable, cela indique que les fokontany avec le plus de bâtiments "conven-

	Tercile	Tercile 2 vs 1		Tercile 3 vs 1		Tercile 2 vs 1		Tercile 3 vs 1		OLS	
		Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur
Intercept	3.0	1.1515	0.201	2.5244	0.005	1.7424	0.237	3.4012	0.018	0.0042	0.034
Part d'évac eau	2.0	0.2408	0.783	1.0663	0.241	0.4884	0.640	1.3224	0.220	0.0001	0.858
	3.0	0.5221	0.556	1.1080	0.259	0.3654	0.736	0.9997	0.402		
Part de véhicules motor.	2.0	1.0031	0.284	-0.3816	0.632	1.2444	0.292	-0.2480	0.804	-0.0002	0.163
	3.0	1.1277	0.294	-0.7590	0.449	1.3694	0.295	-1.1403	0.334		
Nombre d'individus	2.0	-0.2019	0.748	-0.9608	0.196	-0.8086	0.274	-1.6626	0.063	3.546e-09	0.912
	3.0	-0.6927	0.376	-1.1897	0.149	-1.1361	0.261	-1.7697	0.088		
Part de chef niv. sec.	2.0	**-2.8580**	0.012	**-2.8237**	0.009	-1.7826	0.253	-1.0834	0.467	-0.0006	0.138
	3.0	-2.3795	0.064	**-3.7778**	0.007	-0.9899	0.559	-1.9759	0.262		
NDVI moyen	2.0					-1.0271	0.280	-0.2215	0.826	-0.0001	0.524
	3.0					-1.9288	0.101	-1.6081	0.167		
DEM moyen	2.0					**-5.4194**	0.002	-2.8311	0.056	-4.45e-07	0.504
	3.0					**-4.9908**	0.006	**-3.9503**	0.018		
Google O.B.	2.0					2.2798	0.084	0.2815	0.791	-0.0030	0.370
	3.0					2.0080	0.159	0.1126	0.928		
Cluster LCZ	LCZ 3					1.9122	0.109	1.3507	0.237	0.0001	0.094
	LCZ 6					2.6396	0.067	1.8100	0.190	5.271e-05	0.318
	LCZ 6 & LCZ 14					2.7270	0.121	2.5934	0.092	2.443e-05	0.670

Tableau 3.7: Résultats de régressions logistiques et de la méthode des moindres carrés (OLS) pour expliquer la mortalité liée aux maladies respiratoires aiguës.

	Tercile	Tercile 2 vs 1		Tercile 3 vs 1		Tercile 2 vs 1		Tercile 3 vs 1		OLS	
		Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur	Coefficient	p-valeur
Intercept	3.0	1.1979	0.158	1.6675	0.042	0.7477	0.562	2.1500	0.080	0.0101	0.000
Part d'évac eau	2.0	-0.5148	0.550	0.4811	0.552	-0.6240	0.554	0.3877	0.711	-0.0001	0.483
	3.0	0.3557	0.700	0.3346	0.724	0.1343	0.906	-0.8044	0.513		
Part de véhicules motor.	2.0	1.4007	0.084	-0.3874	0.633	1.3628	0.148	-0.7004	0.513	0.0001	0.190
	3.0	-0.0446	0.965	-1.4049	0.163	-0.2588	0.818	-1.7589	0.168		
Nombre d'individus	2.0	-1.3706	0.056	-1.2183	0.084	**-1.6677**	0.044	**-1.7054**	0.041	-1.816e-08	0.448
	3.0	**-2.0768**	0.011	**-1.6936**	0.037	**-2.7975**	0.006	-1.6189	0.092		
Part de chef niv. sec.	2.0	-0.3853	0.679	-0.9405	0.310	0.9868	0.484	1.8778	0.231	-0.0004	0.095
	3.0	-1.0599	0.346	-0.0531	0.962	0.3229	0.839	2.9359	0.101		
NDVI moyen	2.0					0.0705	0.944	-1.5776	0.095	-4.164e-05	0.319
	3.0					-0.5270	0.663	-1.1306	0.292		
DEM moyen	2.0					-0.4305	0.708	**-2.6608**	0.039	-2.32e-07	0.165
	3.0					-0.8957	0.493	**-3.1907**	0.026		
Google O.B.	2.0					-0.1847	0.853	-1.4503	0.191	0.0013	0.062
	3.0					-0.8688	0.469	-0.2465	0.849		
Cluster LCZ	LCZ 3					0.9245	0.386	1.9845	0.051	-3.06e-05	0.499
	LCZ 6					0.9843	0.426	1.8801	0.096	0.0001	0.091
	LCZ 6 & LCZ 14					2.6286	0.060	1.7545	0.200	5.071e-05	0.172

Tableau 3.8: Résultats de régressions logistiques et de la méthode des moindres carrés (OLS) pour expliquer la mortalité liée aux maladies respiratoires chroniques.

tionnels" tendent à avoir des taux de mortalité liés aux maladies respiratoires chroniques plus élevés.

Dans le modèle non linéaire ne considérant que les variables socio-économiques, seule la variable "Nombre d'individus" est significative à presque tous les niveaux (ou en est proche sinon), et est associée négativement au taux de mortalité. Les fokontany les plus peuplés ont tendance à avoir une mortalité liée aux maladies respiratoires chroniques moins élevée. Lorsqu'on ajoute l'environnement au modèle, le nombre d'individus dans le fokontany reste significatif, et l'altitude devient aussi significative avec les niveaux de mortalité les plus hauts.

3.5.7 Discussion

Les travaux dans cette partie permettent de montrer qu'il existe bien une corrélation entre les caractéristiques de l'environnement et la mortalité, à l'échelle des quartiers de la ville d'Antananarivo, y compris en tenant compte des variables socio-économiques.

Cette section discute des résultats obtenus à l'aide des modèles statistiques mobilisés. Plusieurs points et limites peuvent être soulevés, concernant les résultats (section 3.5.7) eux-mêmes mais aussi la qualité des données de population (section 3.5.7 et environnementales (section 3.5.7).

Régressions

Les modèles donnent des résultats contrastés, car peu de variables explicatives sont corrélées significativement à la variable dépendante, que ces données soient environnementales ou socio-économiques.

Les caractéristiques socio-économiques sont rarement associées significativement aux variables dépendantes, ce qui indiquerait, selon ces modèles, que toutes les populations sont touchées de manière similaire par ces causes de décès. Le nombre d'habitants des fokontany a souvent un effet significatif, et les résultats indiquent qu'une augmentation de la population diminue les risques. Ce résultat est à nuancer, car ne prend pas en compte la surface des fokontany. En effet, les fokontany des bas-quartiers sont densément peuplés, mais aussi plus petits que la moyenne, donc ont une population limitée. A l'inverse, les fokontany en périphérie sont eux très étendus, et ont une population supérieure en nombre à la moyenne.

Les résultats de la part de ménages ayant accès à une évacuation des eaux usées améliorées dans l'explication des maladies liées à l'eau sont inattendus. L'hypothèse était que les fokontany ayant le moins d'accès à des systèmes d'évacuation auraient d'avantage de décès liés à l'eau, car les populations sont à proximité des eaux salies ou contaminées. Par exemple, les bas-quartiers qui bordent le canal où sont vidées les eaux usées y sont plus exposés. L'effet de cette variable, bien que presque significatif pour les hauts taux de mortalité liés à l'eau, ne semble pas être l'une des principales causes de maladie selon nos modèles.

La part d'individus qui vivent dans des ménages dont le chef a un niveau d'éducation secondaire est significative pour l'espérance de vie, le quotient de mortalité, le taux de décès par maladies liées à l'eau et par maladies respiratoires aiguës lorsque le modèle ne contient pas de variables environnementales. Ce résultat était attendu, car les personnes les plus éduquées sont, en général, associées aux catégories socio-professionnelles supérieures, synonyme de meilleure qualité de vie, mais aussi d'une meilleure connaissance des risques liés aux maladies et d'un meilleur accès au système de soin. En revanche, l'effet de cette variable semble disparaître lorsque l'on ajoute les variables environnementales, ce qui montre la forte interaction entre les deux dimensions.

L'altitude est un critère qui revient plusieurs fois dans les résultats des modèles linéaires et multinomiaux en étant significativement corrélée à la mortalité, ou proche de la significativité. Ce résultat est attendu pour les décès via les maladies liées à l'eau, car ces quartiers ne subissent pas les inondations. Les bas quartiers bordent aussi le canal, souvent bouché, ce qui favorise le développement de ces maladies. Les quartiers les plus hauts (notamment autour du Palais de la Reine) sont aussi les quartiers historiquement les plus riches, et les populations qui y habitent possèdent une qualité de vie supérieure aux populations des autres quartiers. Les populations de ces quartiers se retrouvent favorisées face aux maladies, mêmes respiratoires chroniques pour les taux les plus hauts. La significativité de la part habitants vivant dans des

ménages dont le chef a un niveau secondaire semble disparaître au profit de l'altitude. Deux conséquences peuvent être tirées de cette observation. Premièrement, l'altitude et la part de niveau secondaire sont trop corrélées, ce qui ajoute un bruit supplémentaire aux modèles. D'après la matrice des corrélations 3.21, cette option peut être écartée. La seconde option est que l'effet de l'altitude est plus important que l'effet de cette variable socio-économique.

La morphologie des fokontany est significative pour les espérances de vie, les quotients de mortalité et les maladies liées à l'eau les plus élevés, mais pas pour les maladies respiratoires qui touchent tous les quartiers. Ces maladies peuvent être provoquées en particulier par la pollution causée par les transports, qui est intense dans toutes les parties de la ville. Ces maladies respiratoires semble être corrélées à aucune des variables environnementales et socio-économiques autre que l'altitude et la population. Ce résultat peut impliquer deux points. Premièrement, tous les quartiers sont touchés de la même manière par les facteurs pouvant donner des maladies respiratoires, comme la densité de la ville ou la pollution. Cette possibilité est peu probable, car certaines caractéristiques représentées par les LCZ ou l'altitude sont très différentes d'un quartier à un autre. Par exemple, l'impact de la pollution, plus fort dans les quartiers denses et où il y a peu de végétation, n'est pas le même dans la haute ville et dans les bas-quartiers où la circulation est très dense toute la journée. Deuxièmement, les populations se déplacent beaucoup dans la ville et ne sont pas soumis qu'a un environnement précis, celui de résidence mais aussi durant les trajets, ainsi que sur les lieux de travail. Les populations sont soumises aux conditions environnementales lors de tous ces déplacements, ce qui délocalise en particulier les causes des maladies respiratoires.

Limites : Données de population

Plusieurs sources de données de population sont utilisées pour générer ces résultats. Les variables dépendantes relatives aux causes de décès sont calculées à partir des données de mortalité présentées en section 3.2.2, une base de données exhaustives fondées sur les fiches du BMH de la ville d'Antananarivo. Les données socio-économiques sont issues du recensement général de la population et de l'habitat (RGPH3), effectué par l'INSTAT en 2018.

Plusieurs limites peuvent être soulevées concernant les données de mortalité. D'abord, ces données sont récupérées via des formulaires écrits à la main, qui sont ensuite retranscrits manuellement sous format numérique. Cette phase de retranscription est une source d'erreurs dans l'attribution des fokontany. La seconde source d'erreur dans l'attribution des fokontany provient de l'appariement effectué dans ces travaux et présentés en section 3.5.1. Bien que contrôlés par arrondissement, les associations entre les noms entrés dans la base de données de mortalité et la base de données officielle peuvent être incorrectes, ayant pour conséquence de noyer d'éventuels phénomènes et de masquer des corrélations. Enfin, la dernière source d'erreur concernant ces données provient du renseignement de la cause de décès sur le formulaire. Lorsqu'un décès a lieu et que les médecins ne sont pas directement sur les lieux pour constater le décès, ils interrogent les proches du défunt afin d'en déterminer la cause. Des erreurs de diagnostic, basées sur ces interrogatoires, sont possibles.

Les données de recensement peuvent aussi contenir des erreurs. Le décompte de la population est parfois compliqué, en particulier dans des quartiers comportant des bidonvilles et/ou des zones de constructions légères. A Antananarivo, les zones de constructions informelles

ne s'étendent pas sur les parties extérieures de la ville, mais comblent les espaces libres à l'intérieur des quartiers d'habitations conventionnelles. Cette structuration très variée de la ville, même dans des zones restreintes, complique le travail des enquêteurs dans le décompte. Selon cette observation, l'erreur de décompte ne fausse pas uniquement les chiffres de la population totale mais aussi la structure de la population. Les ménages les plus défavorisés seraient sous recensés, modifiant en particulier les terciles les plus bas des variables socio-économiques. Enfin, la délimitation des zones de dénombrement n'est pas celle des fokontany et nous n'avons pas eu accès à la cartographie utilisée pour le recensement. Nous n'avons pas pu vérifier si ces délimitations étaient les mêmes que celles mobilisées dans cette étude.

Nous considérons, dans ces travaux uniquement, le lieu de résidence pour définir l'environnement subit par les individus. Nous avons ainsi fait l'hypothèse que le temps d'exposition des individus dans d'autres fokontany était négligeable devant celui dans le fokontany de résidence. Cependant, les personnes et en particulier les actifs sont souvent mobiles dans une zone urbaine comme Antananarivo, et donc d'autres environnements pourraient être considérés afin d'affiner cette étude. L'inclusion de ces informations nécessiterais des données plus complètes sur les individus. De la même manière, nous avons agrégé l'information au niveau des fokontany, sans considérer la diversité socio-économique au sein de chacun d'entre eux. Affiner les données spatiales pourrait permettre de spécifier les interactions montrées en Section 3.5.6.

Limites : Données environnementales

Les limites des données environnementales sont étroitement liées aux limites des données de population. La morphologie de la ville est très diversifiée. Par exemple, dans les bas quartiers, des zones informelles sont intriquées dans des zones où le bâti est plus formel, comme en brique ou en parpaing. Ces zones pourraient être associées à deux classe LCZ différentes. Cette diversité ne peut pas toujours être modélisée, à cause de la résolution des données satellites utilisées, ou du système de classification du sol.

Dans un premier temps, les images Sentinel-2 ont une résolution de 10 mètres, qui n'est pas assez fine pour caractériser de petites zones de bidonvilles intriquées entre deux habitations conventionnelles.

Dans un second temps, le système de classification LCZ est structuré par rapport à des tableaux de pixels et non des pixels seuls. Les différentes cartes et jeux de données développés par la communauté scientifique, impliquant des LCZ, ont des résolutions entre 100 et 320 mètres. La résolution de la carte utilisée dans cette étude à une résolution de 100 mètres. Cette résolution ne permet de modéliser les zones informelles des bas-quartiers d'Antananarivo. Par conséquent, les bas-quartiers sont ici associés au cluster LCZ *Compact low-rise*, bien qu'ils contiennent des parties pouvant être associées à *Lightweight low-rise*. Ce mélange est illustré en Figure 3.22. Ce fokontany contient des zones qu'il n'est pas possible en pratique de séparer, mais qui pourraient appartenir à deux groupes différents. Dans ces travaux, nous avons apporté cette information sur les habitations conventionnelles et légères via les données Google Open Buildings, qui donnent la confiance d'un modèle de détection dans sa classification. Cette information est valable si et seulement si l'hypothèse que le score de confiance du modèle reflète la structure des bâtiments est valide.

Figure 3.22: Image satellite issue des données Google du fokontany "Manarintsoa Afovoany". En rouge sont entourées des zones de bâtis informels, et en bleu de zones de bâtis plus formels, mélangées avec de l'informel.

Enfin, les données de population n'indiquent que le lieu de résidence des personnes décédées, mais n'indiquent pas le lieu de travail. Dans un milieu urbain, les habitants, en particulier les adultes, sont très mobiles et peuvent passer leurs journées loin de leur lieu d'habitation. Ces données ne permettent donc pas de relier tous les environnements que traversent les habitants sur le long terme (quelques heures de la journée au moins). Ainsi, certaines causes de décès attribuées ici au fokontany de résidence pourraient renvoyer en pratique à d'autres fokontany avec des environnements différents.

3.6 Conclusion

Dans ce chapitre, l'objectif était d'estimer l'apport des images satellites pour l'analyse de données démographiques à petite échelle. Nous avons utilisé comme exemple applicatif la base de données de mortalité d'Antananarivo, entretenue par l'IPM, le BMH et l'INSTAT. Cette base donne accès, après traitement, aux causes de tous les décès ayant eu lieu dans la ville entre 2016 et 2020, ainsi qu'au quartier (ou fokontany) de résidence des défunts. L'environnement a été caractérisé au niveau de ces quartiers, à l'aide d'un jeu d'indicateurs : NDVI, DEM, détection du bâti (Google O.B.) et LCZ. Pour générer une carte LCZ, correspondant à l'intervalle de

dates des données (2018), une méthode d'adaptation de domaine a été développée, se basant sur l'intégration des définitions physiques des LCZ, appelés descripteurs, via un mécanisme d'attention. Cette méthode n'a cependant pas donné des cartes de suffisamment bonne qualité pour être intégrée dans un processus d'analyse démographique, et une autre carte rendue disponible par la communauté scientifique a été utilisée. Ces informations spatiales ont ensuite été combinées avec des informations socio-économiques des quartiers, tirées du recensement de la population en 2018, pour expliquer un jeu de variables dépendantes sur les causes de décès: l'espérance de vie, le quotient de mortalité des enfants de moins de 5 ans, le taux de décès dus aux maladies liées à l'eau, aux maladies respiratoires aiguës et chroniques. Les caractéristiques environnementales jouent effectivement un rôle important dans la mortalité, même en contrôlant les caractéristiques socio-économiques des quartiers.

Ce premier chapitre a montré des résultats remarquables pour les deux parties de télédétection et de démographie. L'utilisation des caractéristiques environnementales des LCZ n'a pas permis une meilleure adaptation de domaine. Notamment, certaines classes qui ont une fraction de bâti très faible, comme *Bush Scrub* restent très difficiles à classer. Les méthodes à l'état de l'art ne permettent pas non plus de générer des cartes suffisamment précises pour être utilisées dans des applications. Ce point souligne la difficulté de la tâche d'adaptation de domaine des LCZ, conséquence de la grande variété des morphologies des villes. D'autres part, les LCZ ne permettent pas de prendre en compte certains aspect de la ville, à cause de la taille des zones ainsi que la résolution des images Sentinel-2. En particulier les zones de bâtis informels (comme les bidonvilles) qui sont intriquées dans des zones de bâtis conventionnels ne sont pas prises en compte. La qualité des bâtiments a été estimée par la confiance du modèle Google O.B., mais n'a pas été significativement liée aux causes de mortalité. Du point de vue de la démographie, les résultats obtenus sont limités par la qualité des données. En revanche, cette partie montre l'effet important de l'environnement sur les causes de décès, en particulier de l'altitude et de la morphologie des quartiers représentée par les LCZ, même en contrôlant les données socio-économiques. Ces dernières perdent leur significativité avec l'ajout des variables environnementales. D'autres travaux, sur différentes villes et dans différents contextes devront être mis en oeuvre pour valider ces observations.

3.7 Note sur Google Open Building

Cette section ne présente pas de résultats en soi, mais plutôt une réflexion autour des données Google Open Buildings, et des observations qui ont pu être faites au cours de ces travaux. Lors des analyses introductives, nous avons analysé les corrélations entre toutes les variables explicatives, afin de vérifier le caractère d'indépendance nécessaire aux modèles de régression linéaire et logistique multinomial. Cette section présente succinctement les remarques issues des corrélations entre les données Google Open Buildings, et les données socio-économiques. Nous discuterons ensuite de l'usage de ce type de données pour la démographie dans le contexte de l'Afrique subsaharienne.

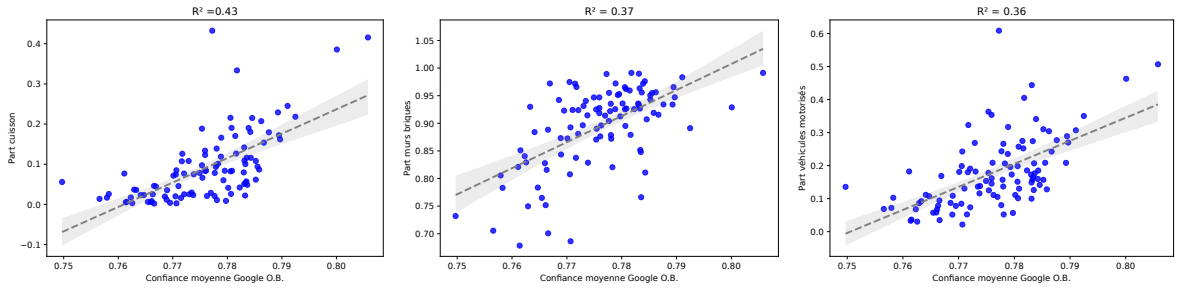


Figure 3.23: Corrélations entre "Part cuisson" (gauche), "Part mur brique" (centre) et "Part véhicule motorisé" (droite) avec Google Open Buildings.

3.7.1 Corrélations avec les variables socio-économiques

La Figure 3.23 affiche les valeurs, par fokontany, de trois indicateurs socio-économiques, la part d'habitants ayant accès à un système de cuisson amélioré (Part cuisson), la part d'habitants ayant un logement en briques (Part briques) ou en parpaings et la part d'individus ayant un véhicule motorisé (Part véhicules motorisés) en fonction de la confiance du modèle de Google Open Buildings dans ses prédictions. Ces trois variables sont corrélées avec les données de confiance du modèle, avec des R^2 respectivement de 0,43, 0,37 et 0,36.

La corrélation avec la variable "Part mur briques" était attendue, et aurait contredit notre hypothèse de départ sur l'utilisation de Google O. B. Nous avons supposé, dans les travaux précédents, que le modèle aurait plus de difficulté à détecter des bâtiments non conventionnels, et donc que les bidonvilles ne seraient pas détectés avec autant de confiance que d'autres bâtiments issus de quartiers plus riches. La variable "Part mur brique" est une image de cet aspect conventionnel des bâtiments, en utilisant le terme conventionnel pour les bâtiments faits de briques ou de parpaings. Nous trouvons une bonne corrélation avec la confiance du modèle Google O.B.: Les fokontany avec le moins de bâtiments avec des murs en briques ou en parpaings correspondent aussi aux quartier dans lequel la confiance du modèle est la plus faible. Sans totalement la valider, cette observation appuie notre hypothèse de départ sur l'utilisation du score de confiance.

Les observations pour les deux autres variables sont similaires. Les fokontany les plus riches qui tendent à avoir une part de logements en brique ou en parpaings, sont aussi les fokontany qui tendent à avoir des population les plus riches, avec des moyens de cuisson améliorés et des véhicules motorisés. Ces corrélations sont encore une fois attendues, et peuvent indiquer une corrélation entre le niveau socio-économique des ménages et la confiance de Google O.B.

Ces observations sont a priori très intéressantes pour la démographie, car permettent d'estimer des facteurs socio-économiques à partir d'un modèle, et pas seulement la détection de bâtiments. Cependant, elles soulèvent aussi de nombreuses limites qui sont discutées dans la sous-section suivante.

3.7.2 Discussion

Les limites de l'utilisation de cette caractérisation du bâti sont multiples. Les premières sont relatives au modèle en lui-même, et à son entraînement. Les données d'entraînement de ce modèle et le code de son entraînement ne sont pas disponibles, bien que publiés dans un rapport qui est lui disponible [77]. Il n'est par conséquent pas possible de reproduire le modèle décrit, ce qui limite notre capacité à évaluer pleinement la robustesse et la fiabilité du modèle. De plus, le score de confiance donné par le vecteur de sortie du modèle n'a pas valeur de probabilité. Ce score indique uniquement à quel point le modèle est sûr de sa prédiction. En particulier, ce qui arrive souvent dans l'apprentissage automatique, le modèle peut accorder un score de confiance très élevé pour une valeur aberrante (*outlier*) ou à une valeur jamais vue auparavant. Ces valeurs aberrantes faussent ainsi les corrélations qui sont observées, menant à des conclusions incorrectes. Dans le cas de notre étude sur Antananarivo, les valeurs de scores sont moyennées par fokontany, ce qui permet de les lisser et de réduire l'impact de ces valeurs aberrantes sur les résultats finaux, en supposant que la majorité des prédictions soient correctes.

D'autres limites sont plutôt liées au contexte de l'utilisation de ces détections. Ce projet permet la détection des bâtiments des pays du Sud, et est mis à jour une fois par an. Dans les villes africaines en particulier, plusieurs dynamiques viennent limiter la validité de ces prédictions. Dans un premier temps, l'urbanisation est rapide et mène à la formation de bidonvilles, dans les villes (comme dans le cas d'Antananarivo) ou dans leurs périphéries. Dans un second temps, les migrations sont dues à plusieurs facteurs économiques et climatiques, elles peuvent être de courte ou de longue durée. La question du temps devient donc importante, car les mouvements de population qui se font plusieurs fois dans l'année impliquent aussi la mise en place de logements, parfois informels, qui ne sont pas détectés par le modèle et peuvent entraîner des résultats incorrects.

L'analyse du contexte de la population à l'étude (mobilité, urbanisation) est primordiale avant l'utilisation de ces données Google O.B.. Par ailleurs, nos observations montrent qu'il y a certaines corrélations avec le niveau socio-économique des populations au niveau des fokontany dans le contexte précis de la ville d'Antananarivo. D'autres études sont nécessaires dans d'autres contextes (villes dans d'autres régions du monde avec des morphologies différentes) et à différentes échelles (quartiers et en dessous) pour déterminer s'il est possible ou non, d'utiliser la confiance du modèle comme un bon intermédiaire pour l'estimation de caractéristiques socio-économiques.

Chapitre 4

A l'échelle nationale

Oh cazzo mi sono dimenticata che non parli italiano.

– Lucrezia Tosato

4.1 Introduction

Le chapitre précédent s'est concentré sur une application de la démographie à l'échelle locale. Cette étude a nécessité une caractérisation très fine de l'environnement, afin d'en percevoir sa complexité au niveau d'une ville ou de quartiers d'une ville. Cependant, les analyses démographiques sont aussi effectuées à d'autres niveaux d'étude, comme celui d'un pays. Ces enquêtes, à plus grande échelle, permettent de déduire des caractéristiques ou tendances démographiques pour une population représentative d'un pays. Ce nouveau niveau d'étude ne possède pas uniquement des caractéristiques spatiales différentes, mais aussi temporelles différentes. En effet, certaines enquêtes au niveau local sont développées sur des temps longs, voire très longs. Par exemple, la saisie et le rassemblement des données de mortalité à Antananarivo étudié dans la première partie se déroule depuis plusieurs dizaines d'années, ce qui permet d'avoir des informations historiques importantes. Des sites de suivi démographiques sont aussi mis en place, et permettent de recueillir des données de manière exhaustive et avec un pas de temps régulier, sur un temps très long. Par exemple, l'observatoire Population Santé Environnement au Sénégal, financé par l'IRD avec le soutien de l'Ined, a permis la collecte de données depuis 1970. Les enquêtes à grande échelle ne peuvent pas s'étendre sur de longues périodes en raison de contraintes logistiques et de ressources, et durent généralement quelques mois. A l'inverse, pour les observatoires, les enquêteurs ne passent qu'une fois dans les ménages enquêtés. Une nouvelle contrainte, plus forte, est donc ajoutée sur les indicateurs environnementaux utilisés : ils doivent correspondre à la fois à une **zone** d'intérêt (contrainte **spatiale**) et une **période** d'intérêt (contrainte **temporelle**) relativement courte. L'intérêt d'utiliser des images satellites pour la génération des indicateurs se trouve renforcé par cette contrainte nouvelle. La fréquence à laquelle certaines données satellites sont disponibles permet de respecter ces deux contraintes.

De plus, les enquêtes à grande échelle sont assez rares dans les pays à faibles et moyens revenus, et les pays d'Afrique subsaharienne ne font pas exception. Moins de données de population sont disponibles, ce qui ralentit le travail de recherche, mais aussi l'action des pouvoirs publics. Pour combler ce manque, l'organisme *Demographic and Health Survey Program*, financé par l'Agence des États-Unis pour le développement international (USAID), réalise des Enquêtes Démographiques et de Santé[1] (EDS, ou Demographic and Health Survey en anglais) dans les pays du Sud pour collecter des données de fécondité, planification familiale, santé maternelle et infantile, genre, VIH/SIDA, paludisme et nutrition, représentative au niveau des pays d'étude. Ces enquêtes sont réalisées conjointement avec les autorités et organismes locaux.

Ce chapitre étudie l'usage des images satellites pour générer une caractérisation environnementale à l'échelle d'un pays, afin de la relier aux données d'une enquête démographique à large échelle. En particulier, cette étude porte sur le lien entre environnement et paludisme. La deuxième section de ce chapitre introduit le pays d'étude le Burkina Faso, ainsi que l'enquête sur le paludisme qui y a été effectuée en 2017/2018. Cette présentation va nous permettre d'appréhender les contraintes locales, ce qui est essentiel pour la génération d'indicateurs environnementaux. La troisième section de ce chapitre présente un processus de caractérisation environnementale, utilisant les LCZ, à l'aide d'apprentissage profond et d'adaptation de do-

[1] https://dhsprogram.com/

Figure 4.1: Zones climatiques du Burkina Faso.

maine. Cette méthode, basée sur les caractéristiques environnementales du pays cible, permet la génération d'une carte de ce pays tout entier. Cette génération, ainsi que la validation de la carte, est présentée en section 4.4. Enfin, la section 4.5 présente l'étude du lien entre la carte environnementale LCZ ainsi générée et la présence de paludisme dans les ménages de l'enquête MIS au Burkina Faso. Cette dernière section montre qu'une relation significative entre l'environnement et la présence de paludisme perdure en présence des caractéristiques socio-économiques des ménages.

4.2 Malaria Indicator Survey 2017-2018, Burkina faso

4.2.1 Présentation générale du Burkina Faso

Le Burkina Faso est un pays d'Afrique de l'Ouest, situé au sud du Sahel. Il possède un climat de type Soudano-Sahélien, qui est caractérisé par une différence de pluviométrie importante entre nord et le sud ainsi que d'une saison sèche et une saison des pluies. Ce pays est composé de trois principales régions climatiques, illustrées en Figure 4.1. Au nord du pays (zone Sahélienne), le climat est aride avec une saison sèche longue (environ 9 mois). Au sud de pays (zone Soudano-Guinéenne), le climat est plus humide avec une pluviométrie plus importante. Entre les deux se situe une région intermédiaire (zone Soudano-Sahélienne) dont la saison sèche ne dure pas plus de 8 mois et dont la pluviométrie est supérieure à la région Sahélienne mais inférieure à la région Soudano-Sahélienne. La topographie du pays est relativement plate, avec un point culminant à 739 mètres d'altitude. Par conséquent, les ménages burkinabè vivent dans des conditions environnementales très variées, dépendant de la région climatique dans laquelle ils habitent.

L'économie du pays est fragile, et dépend majoritairement du secteur primaire, notamment de la culture de coton. Sa population est estimée à plus de 20 millions d'habitants, dont 40% vivent sous le seuil de pauvreté. Le rapport 2021-2022 de l'Indice de Développement Humain, publié par le Programme des Nations Unies pour le développement (PNUD), positionne le Burkina Faso à la 184e place en termes de développement humain, sur 191 pays évalués. Les pouvoirs publics burkinabè font face à d'importants défis dans des domaines cruciaux tels que l'éducation, la santé et le niveau de vie. De plus, la situation politique du pays reste instable depuis un coup d'Etat en 2022, suivi par l'instauration d'un gouvernement de transition. La population subit aussi la présence de groupes terroristes, notamment au nord du pays, ce qui ralentit le développement du pays.

Le Burkina Faso fait face à de nombreuses menaces sur sa situation sanitaire. En particulier, le paludisme est très largement répandu sur le territoire. Il s'agit du dixième pays le plus impacté au monde. Au niveau national, le paludisme représenterait près de 22% des décès. En 2020, cette maladie était aussi la cause de 43% des consultations médicales[2]. Le paludisme est une maladie infectieuse, potentiellement mortelle, qui peut être transmise à l'homme par des piqûres de moustiques infectés par des parasites appartenant au genre *Plasmodium*. Parmi les 5 parasites responsables de la maladie, le *Plasmodium falciparum* est le plus répandu en Afrique, mais aussi celui qui cause le plus de décès[3]. Pour réduire l'impact du paludisme et l'éradiquer, les autorités de santé burkinabè ont établi un plan d'action entre 2016 et 2020 [80]. Ce plan souligne l'importance de l'environnement sur la propagation de la maladie, car certains environnements accélèrent la prolifération des vecteurs. Selon ce plan d'action, la pluviométrie, la température et le couvert végétal, qui peuvent être estimés depuis les images satellites, sont les facteurs les plus importants. Ils sont d'ailleurs amenés à évoluer et à être renforcés avec le changement climatique.

4.2.2 Données collectées sur les ménages

Les enquêtes sur les indicateurs du paludisme[4] (MIS) sont conçues pour surveiller le paludisme dans le cadre de l'initiative mondiale pour combattre cette maladie. Leur objectif est d'évaluer les indicateurs démographiques et de santé fondamentaux, ainsi que les connaissances et croyances de la population par rapport à cette maladie. Une enquête MIS a été menée au Burkina Faso entre novembre 2017 et mars 2018. La structure de ce type d'enquête est basée sur celle des EDS.

Données des ménages relatives au paludisme

Pour produire des indicateurs représentatifs au niveau d'une zone d'étude définie, par exemple une région administrative, les enquêtes EDS suivent une procédure d'échantillonnage en deux étapes : tout d'abord, les zones prospectées par un représentant de l'enquête, appelées grappes ou zones d'énumérations (ZE), sont échantillonnées de manière aléatoire. Ensuite, dans cha-

[2]https://www.severemalaria.org/fr/countries/burkina-faso
[3]https://www.who.int/fr/news-room/fact-sheets/detail/malaria
[4]https://dhsprogram.com/methodology/survey-types/mis.cfm

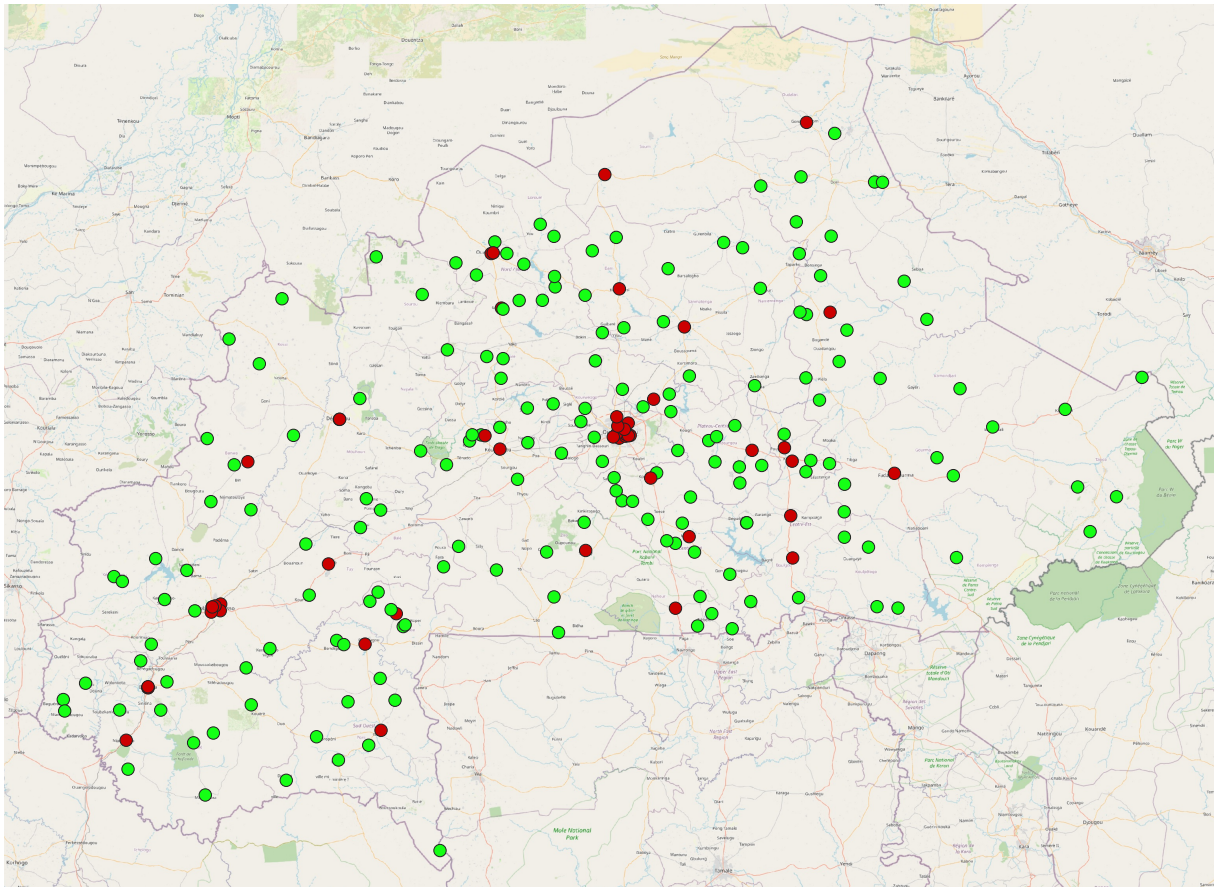


Figure 4.2: Position des zones d'énumération sondées dans l'enquête MIS 2017/2018 au Burkina Faso. Les zones rurales sont indiquées en vert, les zones urbaines en rouge.

cune des ZE échantillonnées, les ménages sont sélectionnés de manière aléatoire pour des entretiens [81]. Les enquêtes MIS suivent cette même méthode d'échantillonnage. Leur objectif est d'estimer la prévalence du paludisme pour les enfants de **6 à 59 mois**. Premièrement, des tests de dépistage rapide du paludisme (donnant des résultats en 15 minutes) sont réalisés sur tous les enfants de 6 à 59 mois dans les ménages échantillonnés, moyennant consentement de leurs représentants légaux. Les tests positifs sont ensuite confirmés par des tests de laboratoire, qui sont plus fiables et permettent de déterminer le parasite infectieux. Dans l'enquête que nous considérons, les résultats de la prévalence du paludisme sont représentatifs au niveau de 17 régions d'études, qui sont les 13 régions administratives du pays et les grandes villes, comme Ouagadougou ou Bobo-Dioulasso. Dans le reste de ce chapitre, nous définissons le taux de paludisme R_i de la ZE i comme le ratio du nombre d'enfants de 6 à 59 mois positifs au paludisme par le nombre total d'enfants âgés de 6 à 59 mois, parmi les enfants des ménages enquêtés de la ZE i. Pour des raisons de sécurité dans le Sahel, seulement 245 des 252 ZE échantillonnées ont pu être visitées, pour un total de 6322 ménages sondés. La répartition géographique des ménages est donnée en Figure 4.2

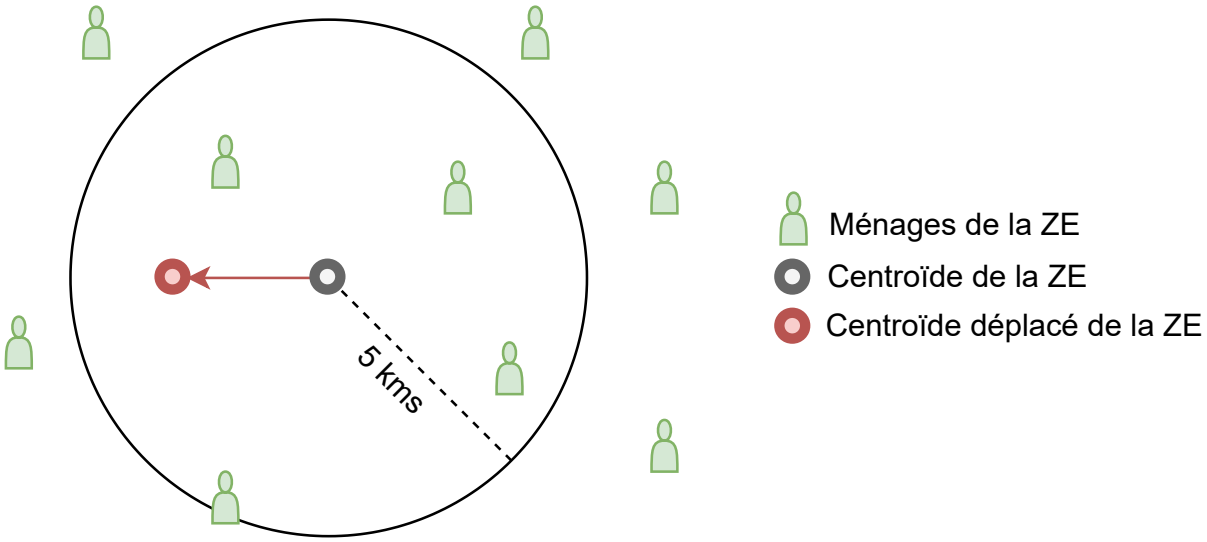


Figure 4.3: Processus d'anonymisation des positions réelles des ménages, pour une ZE rurale. Le processus est le même pour une ZE rurale, avec un cercle de rayon deux km.

Données géographiques et environnementales des ménages

Les positions GPS de chaque ménage sont enregistrées, mais ne sont pas rendues publiques pour conserver leur anonymat. Pour tout de même permettre une analyse spatiale, ces positions sont groupées par ZE et déplacées dans une zone circulaire de deux à 10 km dépendant du caractère urbain ou rural de la ZE. Ce processus d'anonymisation est composé de deux parties. Premièrement, le centroïde des positions des ménages de chaque ZE est calculé. Ensuite, ce centroïde est déplacé aléatoirement dans une zone circulaire de rayon variable selon la situation des ménages. Les centroïdes des ZE identifiées comme rurales dans l'enquête, c'est-à-dire dont les ménages sont situés en zone rurale, sont déplacés dans un cercle de rayon cinq km, sauf pour 1% d'entre eux qui sont déplacés dans un rayon de 10 km. Les ZE concernées par ce déplacement plus important ne sont pas indiquées dans l'enquête. Dans la suite de ce manuscrit, nous faisons l'hypothèse que l'omission de ce plus grand déplacement ne va pas altérer les résultats et nous considérons que tous les centroïdes ruraux sont déplacés dans un cercle de cinq km. Les ZE urbaines, plus petites, sont déplacées dans un rayon de deux km. Ce processus de d'anonymisation est illustré en Figure 4.3. Parmi les 245 ZE visitées, 21 ont des positions GPS corrompues, donc ne peuvent pas être utilisés dans cette étude. Finalement, 224 ZE peuvent être utilisées pour l'analyse.

En plus de données socio-économiques (détaillées en section 4.5.3), les enquêtes fournissent des données environnementales des ménages liées aux ZE. Certaines sont dérivées des images satellites comme le NDVI. Ces indicateurs environnementaux sont agrégés au niveau de la zone d'incertitude de la position GPS des centroïdes, ce qui limite fortement leur précision spatiale. De plus, elles ne sont pas disponibles avec une grande précision temporelle, voire sont données uniqquement pour des années précédant l'année de la collecte. Par exemple, des données sur les précipitations, la températures ou l'indice de végétation sont fournies jusqu'à 2015, avec un pas de 5 ans. Ce biais temporel pourrait masquer la potentielle influence de facteurs temporels, comme les changements saisonniers ou l'environnement climatique.

Ce manque de précision, à la fois spatiale et temporelle, peut être comblé par l'interprétation d'images satellites, très fréquentes et à haute résolution spatiale.

Choix des indicateurs environnementaux

Cette génération d'indicateurs environnementaux peut se faire avec de l'apprentissage profond. La section suivante traite du développement d'une méthode d'adaptation de domaine qui permet de générer une carte du Burkina Faso afin de l'intégrer dans notre démarche, à côté des données socio-économiques. Cette méthode permet de générer une carte LCZ du Burkina Faso respectant les contraintes spatiale et temporelle de l'enquête MIS. Contrairement au premier chapitre de ce manuscrit, uniquement les LCZ seront utilisés pour modéliser l'environnement. Ce système de classification du sol est à la fois le plus complet et permet de représenter le plus de caractéristiques de l'environnement. De plus, l'utilisation d'un unique indicateur, qui peut être illustré facilement, permet une meilleure lisibilité de l'étude au plus grand nombre de scientifiques et décideurs.

4.3 Adaptation de domaine saisonnière

Comme défini dans le premier chapitre, l'adaptation de domaine permet de transférer un apprentissage sur des données source vers des données cibles. Typiquement dans la télédétection, les données sources sont des données d'une région du monde pour lesquelles des données sont disponibles et les données cibles proviennent des régions d'intérêts, différentes des régions sources. Dans cette étude, le jeu de données source est So2Sat [13], présenté en Section 2.3.3. Il est composé de 400 673 images Sentinel-2 étiquetées, de taille $32\times32\times10$, de zones urbaines, dont seulement trois sont en Afrique: Le Caire (Egypte), Nairobi (Kenya) et Le Cap (Afrique du Sud). Les caractéristiques environnementales de l'Afrique, en particulier de l'Afrique de l'Ouest, ne sont pas incluses dans ces données. De plus, la construction du jeu de données autour des zones urbaines (villes et périphéries), ne prend pas en compte les zones rurales.

Cette section présente la méthode d'adaptation de domaine introduite pour générer une carte LCZ du Burkina Faso. Cette méthode doit donc respecter certains critères:

1. prendre en compte les caractéristiques environnementales et climatiques du Burkina Faso,
2. intégrer des régions rurales dans l'entraînement,
3. ne pas nécessiter d'annoter de nouvelles données, ce qui peut être fastidieux et sujets à des erreurs d'appréciation.

Pour cela, nous nous appuierons sur les changements environnementaux causés par l'alternance saison sèche et saison des pluies en Afrique de l'Ouest. Premièrement, nous illustrerons les changements saisonniers en Afrique de l'Ouest, avec l'exemple du Burkina Faso. Ensuite, nous détaillerons la prise en compte des saisons dans le processus d'entraînement d'un réseau de neurones. Cette méthode d'entraînement permet au modèle à extraire des

Figure 4.4: Images issues des satellites Sentinel-2, à Ouagadougou, pendant la saison sèche (gauche, mars 2019) et la saison des pluies (droite, septembre 2019)

.

caractéristiques résistantes aux variations saisonnières. Ces caractéristiques sont également spécifiques à l'environnement du Burkina Faso, facilitant ainsi le transfert des apprentissages du domaine source vers le Burkina Faso.

4.3.1 Caractéristiques climatiques du Burkina Faso

Étant situé en Afrique de l'Ouest, le Burkina Faso connaît une alternance très contrastée entre une saison sèche et une saison des pluies chaque année. La saison des pluies dure généralement quatre mois, entre juin et septembre, tandis que la saison sèche dure les huit autres mois. L'image de l'occupation du sol est fortement modifiée par ces variations saisonnières, comme le montre la figure 4.4, en particulier dans les zones rurales. La premier problème réside dans la question suivante: est-ce que les LCZ sont sujettes à ces changements saisonniers ? Si oui, un modèle devrait pouvoir attribuer deux LCZ différentes à une même zone géographique selon si l'image en entrée a été prise pendant la saison sèche ou la saison des pluies. Si non, ce modèle doit associer une classe unique cette même région géographique. Deux cas se présentent selon les deux types de LCZ définis dans [46] : les zones bâties (LCZ 1-10) et les zones de couverture terrestres (LCZ A-G). Premièrement, la classification des zones bâties est principalement dépendante de la présence de construction et de leur type (taille, matériau). La végétation, principale impactée par les changements saisonniers, n'est pas déterminante dans cette classification. Ainsi, les zones bâties doivent être classées de la même manière pendant la saison sèche et la saison des pluies. Cette hypothèse implique que nous supposons qu'il n'y a pas de migrations saisonnières. Au contraire, la classification des zones de couverture du sol est dépendante de la végétation et est donc soumise à son changement. Celles-ci *peuvent* changer au cours des saisons, avec le changement de végétation (par exemple *Bare soil or sand* vers *Low plants*).

De plus, le pays est grossièrement divisé en trois régions climatiques : le nord du pays a moins de précipitations et des températures plus élevées que le reste du pays. Le sud du pays connaît des précipitations plus importantes et des températures plus basses que le reste du

pays. La troisième région est une région intermédiaire en termes de localisation, de précipitations et de températures.

La méthode d'entraînement doit donc permettre de créer un modèle qui se généralise sur tout le territoire, et qui permet une flexibilité dans la classification de certaines classes. Cette méthode est détaillée dans la partie suivante.

4.3.2 Méthode

Nous définissons un réseau de neurones $F(.)$ qui prend en entrée une image x et renvoie un vecteur c de scores de prédiction pour chaque classe LCZ. La prédiction du modèle est la classe ayant le score le plus élevé. Comme précédemment mentionné, les bases de données globales existantes doivent être complétées par des données supplémentaires relatives au pays d'étude afin d'adapter le modèle à un pays d'Afrique subsaharienne. Cette méthode semi-supervisée vise à tirer parti des bases de données étiquetées existantes, comme So2Sat [13], et de la grande quantité d'images Sentinel-2 non étiquetées pour produire des cartes au niveau du pays. Nous définissons $D_S = (x_i, y_i)_{i \in [\![1, n_S]\!]}$ comme l'ensemble de données étiquetées où x_i est une image Sentinel-2, y_i son étiquette associée et n_S le nombre d'échantillons dans l'ensemble de données. Pour que le modèle soit robuste aux changements saisonniers, la méthode d'entraînement a été construite par rapport à ces changements. Nous complétons les données source D_S par un jeu de données non étiquetées $D_T = (z_i^{s1}, z_i^{s2})_{i \in [\![1, n_T]\!]}$ composé de n_T paires d'images z_i^{s1}, z_i^{s2}. Chaque paire d'images provient d'une même zone, à différentes saisons s_1 et s_2: z_i^{s2} peut être considérée comme une perturbation saisonnière de z_i^{s1}. Les seules informations que nous avons sur les images sont donc les saisons, ce qui est insuffisant pour la classification des LCZ. Pour intégrer ces images dans l'entraînement, l'idée est d'utiliser le principe de conservation des LCZ décrit en sous-section 4.3.1, couplé à de l'apprentissage contrastif qui permet d'extraire des informations de données non annotées.

Apprentissage contrastif

L'apprentissage contrastif est une méthode d'apprentissage *non supervisé* qui vise à entraîner un modèle à apprendre des représentations particulières des données en entrée. Ce type d'apprentissage est très utilisé en vision par ordinateur pour traiter la très grande quantités d'images non étiquetées disponible. Il est en particulier très pertinent en apprentissage semi-supervisé ou auto-supervisé [82], [83], qui permet de pré-entraîner un modèle à l'aide de données non annotées. Cette première étape permet d'améliorer les performances après un second entraînement supervisé. Ce processus permet de minimiser la distance entre les représentations de deux images similaires, paire dite *positive*, et de l'agrandir entre les représentations de deux images différentes, paire dite *négative*. Dans le cas de l'apprentissage profond, ces représentations sont générées par des réseaux de neurones. En général pour la vision par ordinateur, les paires positives sont formées avec augmentation de données, c'est-à-dire une modification de l'image d'origine pour créer une seconde image différente, mais fortement liée à la première. Ensuite, il est possible de minimiser la distance entre les représentations de ces images avec une fonction de coût adaptée, comme la fonction contrastive :

$$L_i = -\log \frac{\exp(sim(x_i, AD(x_i))/\tau)}{\sum_{k=1,k\neq i}^{2B} \exp(sim(x_i, x_k)/\tau)}, \tag{4.1}$$

Où $(x_i)_i \in [\![1, B]\!]$ une image, $sim(.)$ une fonction de similarité, $AD(.)$ une fonction d'augmentation de données, et τ est la température. B est la taille d'un *batch*, $2B$ est donc la taille d'un *batch* avec les augmentations. Le numérateur de cette fonction vise à réduire la distance entre les représentations de la paire positive $(x_i, AD(x_i))$, et à augmenter la distance entre x_i et toutes les autres images du *batch*. En pratique, si $AD(.)$ représente l'ajout de bruit, le modèle est entraîné à être robuste à ce bruit. En considérant que les changements saisonniers comme des augmentations temporelles, l'apprentissage contrastif peut permettre, dans notre cas, de rendre le modèle robuste à ces changements. La sous-section suivante présente le processus d'entraînement élaboré, qui utilise ce coût contrastif saisonnier.

Adaptation de domaine semi-supervisée saisonnière

Afin d'extraire des caractéristiques saisonnières des images saisonnières de la zone cible, sans nécessiter d'annotations supplémentaires, nous utilisons la fonction de coût contrastive pour rendre le modèle robuste aux changements saisonniers. L'équation 4.1 devient alors :

$$L_{i,j} = -\log \frac{\exp(sim(F(z_i), F(z_j))/\tau)}{\sum_{k=1,k\neq i}^{2B} \exp(sim(F(z_i), F(z_k))/\tau)}, \tag{4.2}$$

où B est le nombre de paires d'images, $sim(\cdot, \cdot)$ est la similarité cosinus, τ est la température, $(i, j) \in [\![1, B]\!]^2$, $(z_l)_{l\in[\![1,2B]\!]}$ échantillons d'un *batch* de B échantillons, et (z_i, z_j) une paire positive. Dans notre cas, la paire positive est constituée de deux images de la même zone à différentes saisons et la paire négative est constituée de deux autres images d'une zone différente.

Cette coût contrastif est intégrée dans un processus, appelé *seasonal Semi-Supervised Domain Adaptation* (s-SSDA), à deux chemins impliquant l'ensemble de données source étiqueté D_S et l'ensemble de données cible non étiqueté D_T. Ce processus est illustré à la Figure 4.5. La première piste est un processus d'apprentissage supervisé conventionnel utilisant D_S pour minimiser une perte d'entropie croisée supervisée (ou *Cross-Entropy*) L_S : le modèle $F(.)$ est entraîné à classifier les LCZs sur diverses régions du monde incluses dans D_S. Le deuxième processus, utilisant les échantillons non étiquetés de D_T, est effectué simultanément. $F(.)$ y est utilisé comme un réseau neuronal siamois, de manière similaire à [82], [84], pour classifier deux images différentes de la même zone spatiale mais à deux saisons différentes s_1 et s_2. Pour simplifier l'entraînement du modèle, nous faisons l'hypothèse qu'il n'y a pas de nouvelles installations ou de nouveaux bâtiments construits entre les dates d'enregistrement des deux images. Par exemple, nous supposons qu'une zone classée comme *Low plants* à la première saison restera *Low plants* à la deuxième saison, même si des maisons ont été construites. Ensuite, le coût contrastive L_T est calculée entre les sorties de la paire positive et de la paire négative. Il est à noter que les étiquettes LCZ rurales peuvent changer tout au long de l'année en raison des variations saisonnières. Cependant, les zones urbaines devraient rester inchangées par les

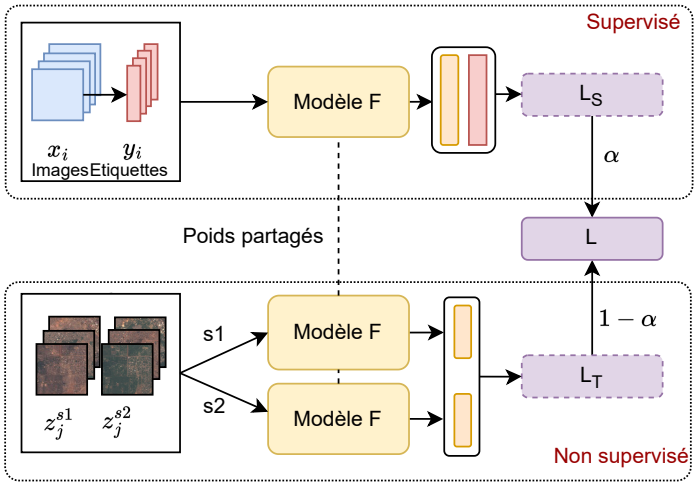


Figure 4.5: Processus d'entraînement comprenant une piste supervisée et une piste non supervisée. Ces deux approches sont menées simultanément pendant l'entraînement. Premièrement, un *batch* d'images étiquetées de taille 32×32x10 est donné au modèle pour calculer la perte supervisée via une entropie croisée L_S. Ensuite, la perte contrastive L_T est calculée à partir des vecteurs de prédiction des paires positives et négatives. Les deux pertes sont combinées en une seule perte totale L pour la rétropropagation.

changements saisonniers, malgré un aspect visuel différent. Nous introduisons cette information connue en ajoutant des poids au coût contrastif pour pénaliser fortement l'incohérence dans les prédictions des zones urbaines, et faiblement les incohérences pour les zones rurales. Cette pénalisation est introduite via un vecteur de poids qui, une fois multiplié avec les vecteurs de prédiction du modèle, donne plus de poids aux erreurs sur les zones urbaines dans la fonction de coût contrastive. Ainsi, nous respectons le postulat établi en section 4.3.1 sur la stabilité des LCZ urbaines par rapport aux saisons. Cette deuxième piste vise à renforcer la robustesse aux saisons ainsi qu'à transférer ses connaissances à des zones nouvelles, qui ne sont pas présentes dans D_S. La perte utilisée pour l'entraînement du modèle $F(.)$ est une combinaison des résultats des deux pistes avec un coefficient de régularisation $\alpha \in [0, 1]$:

$$L = \alpha \times L_S + (1 - \alpha) \times L_T. \tag{4.3}$$

Ce terme de régularisation est déterminé empiriquement.

Régularisation temporelle par chaine de Markov

Les tuiles Sentinel-2 sont disponibles à une très haute fréquence, avec un taux de rafraîchissement de maximum cinq jours. Cette haute fréquence permet d'apporter des informations supplémentaires dans la génération d'indicateurs environnementaux. Les cartes LCZ des zones cibles peuvent être générées non seulement au moment de la période de collecte des données d'enquête, mais aussi pendant des années précédant l'année d'intérêt. Ainsi, il est possible

d'utiliser des cartes générées à différentes périodes pour assurer une continuité temporelle dans notre prédiction. Cette information temporelle peut être intégrée via une chaine de Markov, définie en introduction.

Nous définissons :

- $I_N \in \mathbb{R}^{32\times32}$ une carte LCZ construite à partir de la sortie de $F(.)$ au temps N, où chaque coefficient de I_N est la classe avec le score le plus élevé donné par $F(.)$.
- LCZ_N la classe LCZ ($c_N \in [\![1, 17]\!]$) du patch au temps N.
- M $\in \mathbb{R}^{17\times17}$ une matrice où $m_{r,c\in[\![1,17]\!]^2}$ est le coefficient de M à la ligne r et à la colonne c. $m_{i,j}$ est la probabilité dans le processus de Markov du premier ordre de passer de $LCZ_{N-1} = i$ à $LCZ_N = j$, $(i, j) \in [\![1, 17]\!]^2$. Ces probabilités dépendent du contexte environnemental et politique de la zone cible.

Si la classification LCZ d'une zone spécifique au temps N (c'est-à-dire LCZ_N) suit un processus de Markov de premier ordre (de deux années consécutives à la même période), pour tout N :

$$P(LCZ_N = c_N) = m_{c_{N-1},c_N} \times P(LCZ_{N-1} = c_{N-1}), \tag{4.4}$$

alors, selon le théorème de Bayes :

$$P(LCZ_N = c_N|I_N) = \frac{P(I_N|LCZ_N = c_N)}{P(I_N)} \times m_{c_{N-1},c_N} \times P(LCZ_{N-1} = c_{N-1}). \tag{4.5}$$

Nous faisons l'hypothèse que $P(I_N|LCZ_N = c_N)$ est le score de prédiction du modèle. Après avoir prédit les scores LCZ mono-temporels avec F, la chaîne de Markov peut être appliquée pour obtenir les cartes LCZ finales. Ce processus de régularisation est montré dans la Figure 4.6. Dans la section suivante, nous expliquons comment cela a été utilisé pour générer une carte LCZ du Burkina Faso début 2018, pendant la période d'enquête.

4.4 Génération de l'indicateur à l'échelle d'un pays

4.4.1 Données cibles

Cette sous-section décrit la procédure utilisée pour générer l'ensemble de données cible pour la partie semi-supervisée du processus de formation. Pour réduire l'écart de domaine entre les données d'entraînement disponibles et le Burkina Faso, nous complétons l'ensemble de données source So2Sat par des images Sentinel-2 sur le Burkina Faso à la fin de la saison sèche et de la saison des pluies. Nous utilisons des images de niveau L1C afin de correspondre au modèle So2Sat. Cet ensemble de données cible a été créé en utilisant la procédure suivante :

1. **Téléchargement des tuiles Sentinel-2 de niveau L1C** liées à chacune des capitales régionales du Burkina Faso à la fin de la saison sèche et de la saison des pluies pour maximiser les variations entre les deux tuiles. Les deux tuiles de la même région à différents

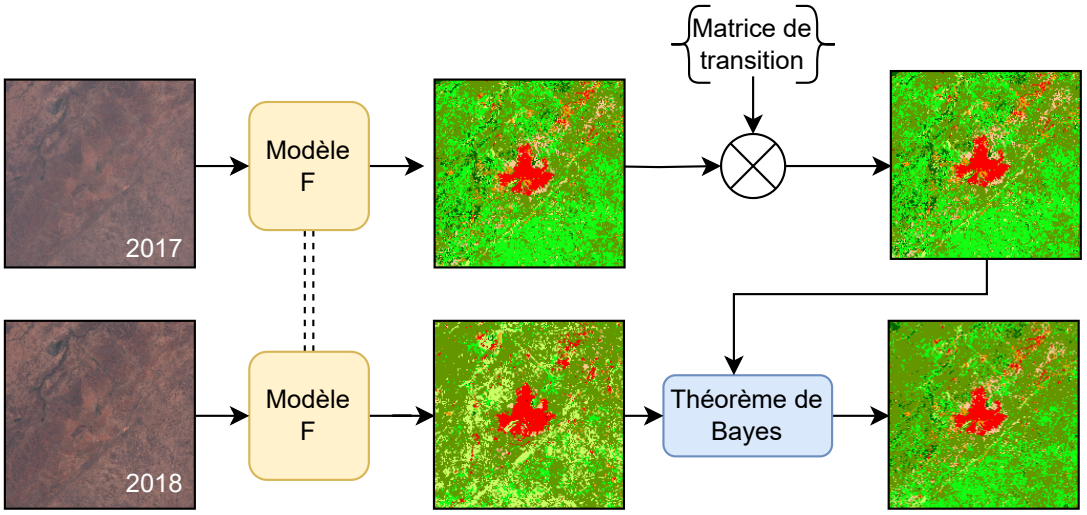


Figure 4.6: Processus de Markov appliqué à la régulation temporelle des cartes LCZ.

moments ont été sélectionnées pour avoir moins de 5% de couverture nuageuse afin de réduire les erreurs causées par les nuages.

2. **Sélection des régions** où des zones d'intérêt peuvent être trouvées : villes, villages, industries, parcs naturels, forêts, lacs ou rivières. Les tuiles sont découpées en formes rectangulaires centrées sur les zones d'intérêt et assez grandes pour inclure l'environnement. Les mêmes régions d'intérêt sont sélectionnées pour les deux tuiles de saison sèche et de saison des pluies.

3. **Division des régions d'intérêt** en images de 32 $\times$ 32 pixels pour correspondre à la taille des images de l'ensemble de données So2Sat. Des paires saisonnières d'images sont créées pour alimenter le réseau neuronal lors de l'entraînement.

Cette procédure donne lieu à 225 000 paires de patchs réparties dans tout le Burkina Faso, comme le montre la Figure 4.7. Comme on peut le voir sur la distribution spatiale des échantillons, toutes les régions et tous les climats sont inclus dans l'ensemble de données d'entraînement.

4.4.2 Paramètres d'entraînement et de génération de la carte

Nous utilisons l'architecture ResNet50 [64], pré-entraînée sur l'ensemble de données complet So2Sat en utilisant la version *block*. Cette étape de pré-entraînement est effectuée pour initialiser les poids du modèle pour la phase d'entraînement semi-supervisé ultérieure. Pour cette étape semi-supervisée, l'optimiseur Adam [85] est utilisé avec un taux d'apprentissage de 0,001, La taille du *batch* pour les phases supervisées et non supervisées est fixée à 256. Le paramètre de température τ est réglé à 0,5. En fonction de nos expériences, nous fixons

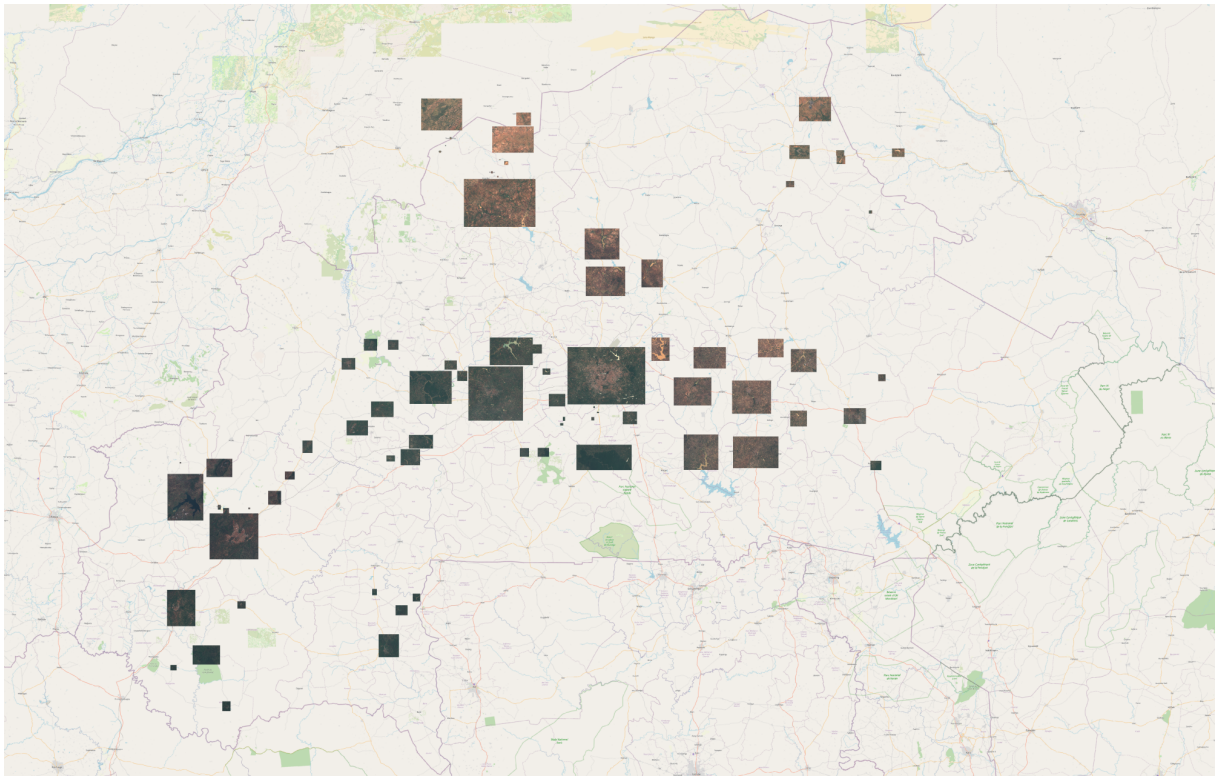


Figure 4.7: Régions d'intérêts utilisées pour la génération du jeu de données cible. Le fond de carte est OpenStreetMap.

le paramètre α de l'équation 4.3 à 0,9 dans ce travail. Les phases d'entraînement et de validation ont été effectuées à l'aide de cartes graphique NVDIA V100 16Go, disponibles sur le super-calculateur HPE SGI 8600 Jean-Zay.

L'hypothèse selon laquelle les LCZ suivent une chaîne de Markov nécessite la définition de coefficients de transition, qui représentent les probabilités de passer d'un état, c'est-à-dire d'une classe de LCZ, à un autre. Le contexte du pays influence fortement ces probabilités, car elles peuvent résulter de l'urbanisation (par exemple, la transition de *Open low-rise* à *Compact low-rise*), de la gestion forestière (par exemple, l'interdiction de la déforestation) ou de la situation géographique du pays. Par exemple, au Burkina Faso, l'urbanisation progresse rapidement et le terrain du pays est principalement plat. Ainsi, les villes ont tendance à s'étendre horizontalement plutôt que verticalement, ce qui suggère que les poids de transition vers *Compact high-rise* devraient être très faibles. Les poids de transition ont été définis empiriquement en tenant compte des caractéristiques spatiales et politiques du Burkina Faso, telles que le plan d'urbanisation. Le Tableau 4.1 résume ces coefficients.

4.4.3 Carte LCZ du Burkina Faso

Les données de l'enquête MIS ont été collectées au Burkina Faso de novembre 2017 à mars 2018, de la fin de la saison des pluies à la saison sèche. Cependant, la carte sera générée pour début 2018 car la majorité des entretiens a été menée en janvier et février. Des tuiles Sentinel-2 du

	1	2	3	4	5	6	7	8	9	10	A	B	C	D	E	F	G
1 - Compact high-rise	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 - Compact mid-rise	0.05	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3 - Compact low-rise	0	0.05	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4 - Open high-rise	0.05	0	0	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0
5 - Open mid-rise	0	0.05	0	0	0.95	0	0	0	0	0	0	0	0	0	0	0	0
6 - Open low-rise	0	0	0.1	0	0	0.9	0	0	0	0	0	0	0	0	0	0	0
7 - Lightweight low-rise	0	0	0.1	0	0	0	0.9	0	0	0	0	0	0	0	0	0	0
8 - Large low-rise	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0	0	0
9 - Sparsely built	0	0	0	0	0	0.1	0	0	0.8	0	0	0.1	0	0	0	0	0
10 - Heavy industry	0	0	0	0	0	0	0	0	0	1.0	0	0	0	0	0	0	0
A - Dense Trees	0	0	0	0	0	0	0	0	0	0	0.95	0.05	0	0	0	0	0
B - Scattered Trees	0	0	0	0	0	0	0	0	0.30	0	0.01	0.69	0	0	0	0	0
C - Bush, scrubs	0	0	0	0	0	0	0	0	0.10	0	0	0	0.9	0	0	0	0
D - Low plants	0	0	0	0	0	0	0	0	0.10	0	0	0	0	0.9	0	0	0
E - Bare rock or paved	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0	0	0
F - Bare soil or sand	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0.99	0
G - Water	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.0

Tableau 4.1: Coefficients de transition pour le processus de Markov lors de la génération de la carte LCZ. La correspondance entre les identifiants et les classes est disponible dans la Figure 2.5.

début de l'année 2017 et 2018 seront utilisées, pour effectuer le processus de Markov. Seules les images avec une couverture nuageuse de moins de 5% sont sélectionnées. Pour correspondre à la taille d'entrée du modèle, 32×32 pixels, chaque tuile Sentinel est divisée en images de 320m x 320m. Les classes de toutes ces images sont prédites avec le modèle $F(.)$, et le processus Markovien est appliqué. La carte, affichée en Figure 4.8 est produite en concaténant tous les résultats de classification. Une seule carte a été générée pour le début d'année 2018, période pendant laquelle le plus de ménages ont été sondés. Cette carte a une taille initiale de 320m × 320m, mais est suréchantillonnée à la résolution d'image Sentinel d'entrée : les pixels initiaux (320m × 320m) sont découpés en pixels plus petits (10m x 10m). Comme il pouvait être anticipé après observation des zones climatiques du Burkina Faso, la carte est divisée en trois parties principales (quatre si l'on considère les villes/zones urbaines). La partie sud plus humide est principalement couverte de zones *Scattered Trees*, et la partie plus tempérée est principalement couverte de zones *Bush/scrub*, sauf pour les réserves naturelles dans la partie est du pays. Les zones plus sèches du nord présentent un plus grand défi pour la classification LCZ. En effet, la précision de cette classification est limitée par la résolution des images Sentinel-2, 10m et 20m suréchantillonnées à 10m. Cela complique la détection de petits bâtiments ou maisons par le modèle, ceux-ci étant souvent de taille inférieure à la résolution de l'image. L'absence de cette détection conduit à une mauvaise interprétation des zones avec un très faible taux de bâti, par exemple la classe LCZ *Sparsely built*, et des zones sans aucun bâtiment, par exemple *Bare soil or sand* et *Low plant*, qui sont présentes dans cette partie nord. Cette confusion est renforcée par l'utilisation de méthodes d'apprentissage semi-supervisé qui ne performent pas bien lorsque les images d'entrée sont difficiles ou rares. C'est particulièrement le cas pour la classe *Sparsely built* dont la classification peut être ambiguë, comme indiqué par Bechtel *et al.* [55].

4.4.4 Comparaison avec d'autres cartes LCZ et étude d'ablation

Plusieurs cartes de classification du sol utilisant les LCZ ont été introduits ces dernières années. En particulier, So2Sat GUL [63] et Global LCZ map [59] ont permis une nouvelle vue d'ensemble mondiale de la morphologie de la surface terrestre.

- **So2Sat GUL** est un ensemble de 1642 villes du monde, aussi basé sur So2Sat LCZ42 qui est utilisé dans ces travaux comme jeu d'entraînement source. Le modèle utilisé pour la génération de ces cartes a été entraîné de manière supervisée avec des images Sentinel-1, et Sentinel-2 de plusieurs saisons. La prédiction finale est donnée par la moyenne des prédictions. Seulement deux cartes pour le Burkina Faso sont disponibles : la capitale, Ouagadougou, au centre du pays, et Bobo-Diolasso, la deuxième plus grande ville du

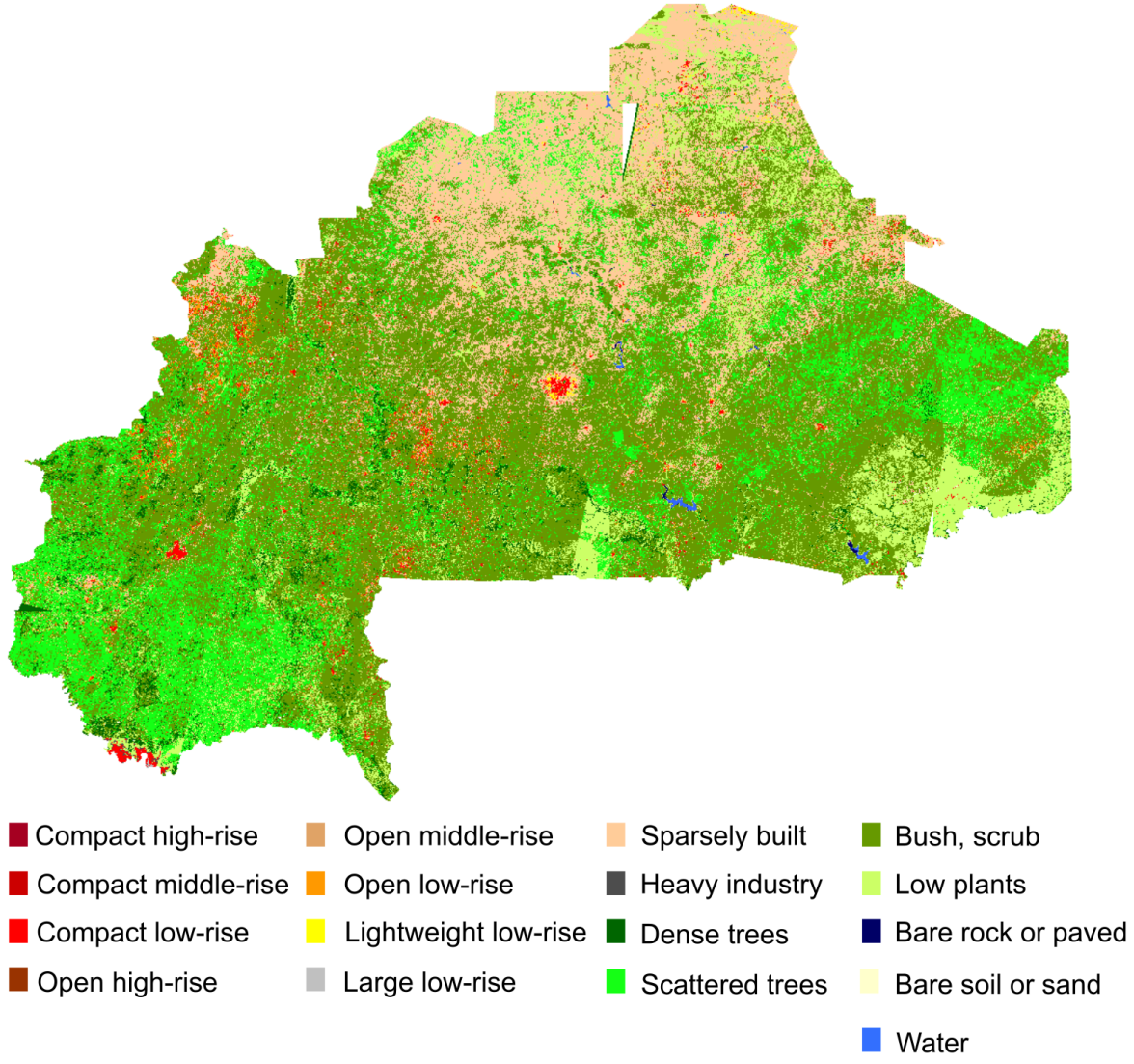


Figure 4.8: Carte LCZ du Burkina Faso pour début 2018 obtenu par concaténation des prédictions du modèle présenté dans cette section.

pays, dans le sud-ouest.

- **Global LCZ** couvre toute la surface de la planète [59] avec une résolution de 100m. Elle est le résultat de forêts aléatoires basées sur 46 caractéristiques spatiales de chaque pixel des zones d'entraînement précédemment réalisées par des experts urbains ou dans des plateformes de crowdsourcing [57]. Toute la surface du Burkina Faso a été cartographiée.

La validation des cartes d'occupation des sols en Afrique subsaharienne intertropicale présente des défis significatifs. Premièrement, les écosystèmes divers et complexes de la région contribuent à un large éventail de types de couverture terrestre. Cette variabilité rend difficile le développement d'approches de validation généralisées, en particulier pour les LCZ. Deuxièmement, peu de références sont disponibles en Afrique subsaharienne, ce qui a pour cause et conséquence que la majorité des jeux de données ont été faits pour les pays du Nord. Pour valider notre approche et la comparer aux cartes existantes, il était nécessaire d'étiqueter à la main certaines régions. Des tuiles Sentinel-2 de vastes zones autour de quatre villes des différentes zones climatiques du Burkina Faso : Ouagadougou, Bobo-Dioulasso, Fada-Ngourma et Ouahigouya ont d'abord été collectées [86]. Ces tuiles sont découpées en images et étiquetées manuellement à partir d'images d'images a très haute résolution à l'aide du logiciel Google Earth. Comme la carte So2Sat GUL n'est disponible que pour les zones situées au-dessus de Ouagadougou et de Bobo-Dioulasso, les mesures ont été calculées sur les 186 images qui se chevauchent cet ensemble de test introduit dans cette partie. Comme les cartes GUL et la carte Global LCZ sont disponibles à une résolution inférieure à celle des images de validation, la comparaison des cartes n'est pas triviale. Deux procédures d'aggrégation sont utilisées :

1. nous considérons que la zone est bien classée si la majorité des pixels de la zone ont la même étiquette que celle de notre ensemble de validation (MR),

2. nous considérons que la zone est bien classée si au moins un des sous-pixels a la même étiquette que celle de notre ensemble de validation (IS). Cette deuxième méthode d'agrégation est donc la plus favorable pour les cartes GUL et Global LCZ.

La Figure 4.9 est une comparaison visuelle des cartes LCZ des 3 villes (Ouagadougou, Bobo-Dioulasso et Fada-N'Gourma) générées à l'aide de différents modèles. Les différentes méthodes sont présentées : la méthode saisonnière proposée s-SSDA, l'entraînement supervisé sur So2Sat (Baseline), la carte Global LCZ [59] et les cartes GUL [63], de gauche à droite. Un fond de carte OpenStreetMap est affiché comme référence. Les défis liés à l'adaptation sont visibles. Le modèle *Baseline* ne parvient pas à correctement identifier les zones urbaines, car elle sont très différentes que les zones urbaines présentes dans le jeu d'entraînement: les géométries, matériaux et environnement sont différents. Les cartes GUL souffrent des mêmes difficultés sur Ouagadougou.

Les résultats de comparaison quantitatifs sont donnés dans le Tableau 4.2. En plus de la comparaison avec des cartes issues d'autres travaux de recherche, ce tableau contient les résultats des parties de la méthodes indépendemment les unes des autres. Cela permet d'évaluer l'apport de chaque partie de la méthode. Les deux méthodes s-SSDA surpassent So2Sat GUL dans tous les cas et pour toutes les mesures. Dans le meilleur des cas, la carte générée par

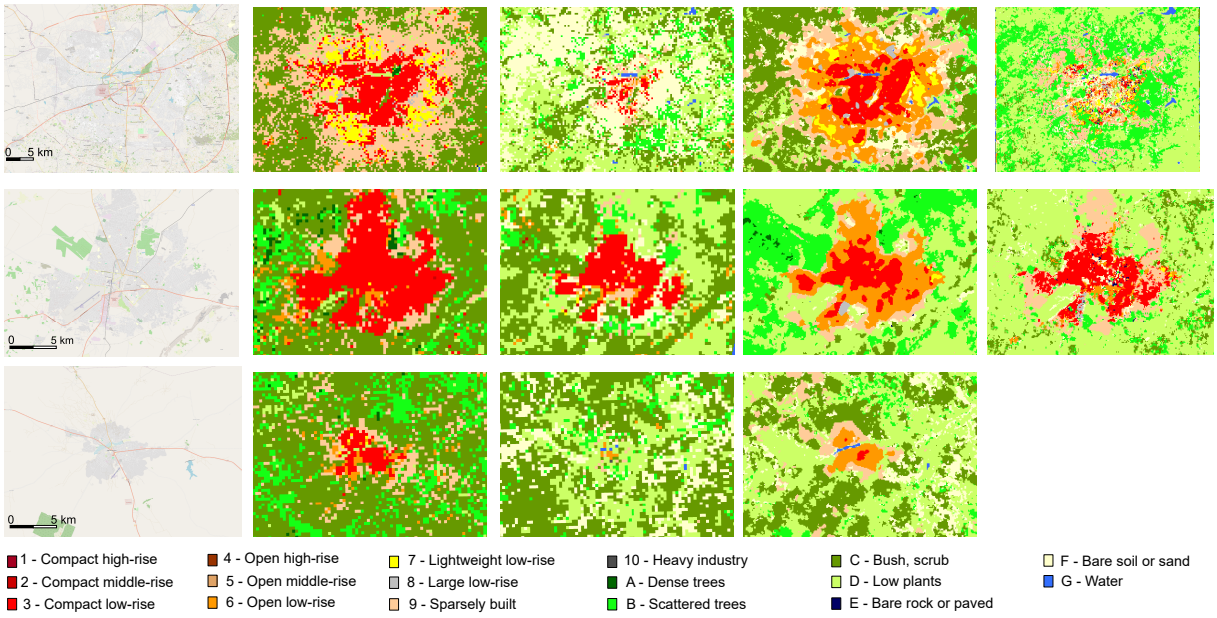


Figure 4.9: Comparaison visuelle des cartes LCZ de 3 villes (Ouagadougou, Bobo-Dioulasso et Fada-Ngourma). En plus d'une carte provenant d'OpenStreetMap, différentes méthodes sont montrées : s-SSDA, la méthode proposée dans cette partie, l'entraînement supervisé sur So2Sat (Baseline), la carte Global LCZ [59] et GUL [63], de gauche à droite.

s-SSDA avec régularisation temporelle surpasse GUL de plus de 30%. Ces résultats mettent en évidence la nécessité d'effectuer une adaptation de domaine pendant l'entraînement dans des zones non vues du monde, même après l'entraînement sur un ensemble de données, dit global. La carte Global LCZ, dans le meilleur des cas, a obtenu de nouvelles améliorations, obtenant une précision globale de 0,447, un score F1 de 0,481 et un IoU de 0,329, qui sont similaires mais légèrement meilleurs que la s-SSDA sans régularisation. La régularisation temporelle permet d'obtenir de meilleurs résultats de cartographie, avec une amélioration de près de 10 points pour chaque métrique. Les résultats sur notre ensemble de validation mettent en évidence le potentiel de tirer parti des caractéristiques saisonnières pour l'adaptation de domaine.

4.5 Lier environnement et population au Burkina Faso

La carte LCZ générée peut être combinée dans une analyse avec les données de population pour identifier le lien qui pourrait exister entre les caractéristiques environnementales et la présence du paludisme chez les enfants au Burkina Faso. Les deux prochaines sous-sections présentent les étapes qui permettent cette évaluation. D'abord, l'imprécision sur les données de position des ménages complique le lien entre la carte générée en section 4.4 et les données de l'enquête. La première sous-section présente le processus de caractérisation de l'environnement des ménages. La seconde sous-section présente l'intégration de cette caractérisation dans une démarche démographique, prenant en compte les caractéristiques socio-économiques des ménages.

Méthode	OA	F1	IoU
Baseline	0,245	0,270	0,175
GUL - MR	0,140	0,122	0,070
GUL - IS	0,194	0,203	0,119
Global LCZ - MR	0,360	0,397	0,258
Global LCZ - IS	0,447	0,481	0,329
s-SSDA (notre méthode)	0,427	0,402	0,278
s-SSDA + Markov (notre méthode)	**0,561**	**0,538**	**0,389**

Tableau 4.2: Résultats de comparaison. Nous comparons notre carte à d'autres produits LCZ sur 494 patchs étiquetés manuellement. OA est l'*accuracy* globale (*Overall accuracy*), F1 est le score F1 et IoU est le coefficient de Jacquard (*Intersection over Union*). Les meilleurs résultats sont indiqués en **gras**.

4.5.1 Lier la carte à l'enquête

Les coordonnées GPS fournies ne sont pas parfaites pour préserver l'anonymat des ménages enquêtés, comme indiqué en section 4.2.2. Par conséquent, la liaison ménages/environnements n'est pas triviale. Les experts de l'organisme *DHS program*, qui produisent les données MIS, recommandent de calculer les moyennes des indicateurs sur la zone d'incertitude autour des centroides des ZE [87]. Cette méthode a été remise en cause, notamment par Grace *et al.* 2019 [88], car elle induit une trop grande différence avec les variables issues des positions réelles. Par exemple, les indicateurs des ménages ruraux suffisamment proches des villes pourraient être modifiés et biaiser l'analyse des données de population en les associant à des environnements incohérents. Selon Grace *et al.* 2019 [88], les indicateurs doivent être sélectionnés manuellement pour correspondre à des zones habitées, où des installations sont construites. Ces zones peuvent être sélectionnées à partir d'images VHR, en utilisant par exemple Google Earth Engine, ou aussi à partir de données de population existantes telles que le GHSL [10] ou les données de population WorldPop[5]. Cependant, la méthode introduite par Grace *et al.* 2019 [88] nécessite d'avoir accès à de telles données de référence au moment de l'étude, ce qui n'est pas toujours le cas. Suivant la même idée que Grace *et al.* 2019 sur le problème de variabilité des données, nous nous basons sur la carte générée pour sélectionner nos valeurs.

Les zones d'incertitude sont modélisées autour de chaque centroïde de ZE par des disques C_e (avec e étant l'identifiant de la ZE) d'un rayon de deux ou cinq kilomètres selon le type urbain ou rural de la ZE. Pour chaque ZE e, nous échantillonnons semi-aléatoirement n_{random} zones carrées de taille A = 10×10 (100m×100m) à l'intérieur de C_e pour modéliser les positions réelles des ménages interviewés. Pour garantir la cohérence de l'échantillonnage avec les informations MIS sur le type urbain/rural de e, ces n_{random} zones ont été sélectionnées pour contenir au moins δ = 10% des pixels dont les classifications LCZ appartiennent à la zone urbaine ou rurale. Cette procédure d'échantillonnage semi-aléatoire donne un total de n_{random} zones qui couvrent toutes les ZE. Ces n_{random} zones donnent une vue globale sur les environnements locaux dans lequel les ménages ont été interviewés, excluant les valeurs contradictoires tout

[5]https://hub.worldpop.org/doi/10.5258/SOTON/WP00004

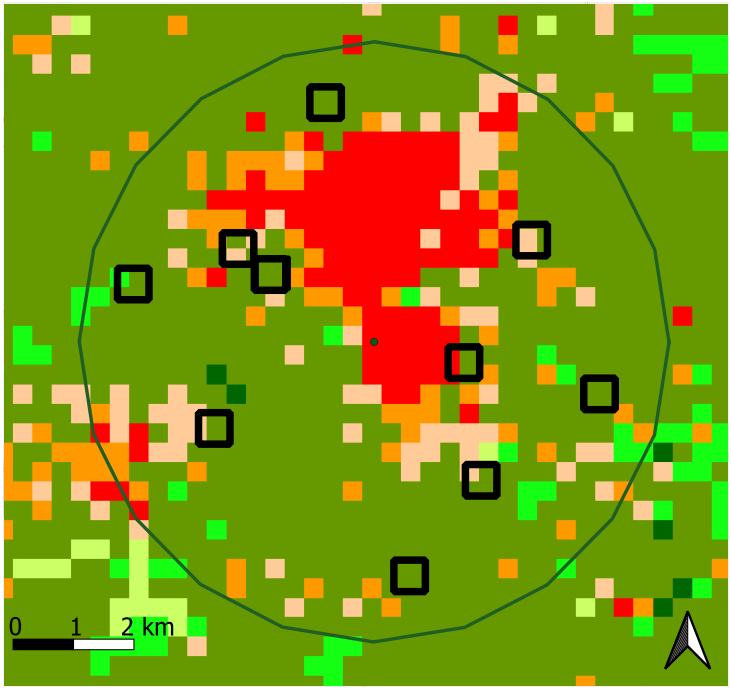


Figure 4.10: Sélection semi-aléatoire des positions potentielles des ménages sondés. Les zones noires sont les groupes de pixels sélectionnés pour être caractériser l'environnement des ménages de la ZE. L'exemple affiché montre la ZE urbaine autour de Dédougou au Burkina Faso.

en préservant les cas limites comme les frontières des villes et les parcs urbains. Un exemple visuel d'une sélection semi-aléatoire de zones est donné dans la Figure 4.10.

Cette caractérisation de l'environnement des ménages peut maintenant être reliée aux données sur le paludisme. Afin d'avoir une vue la plus globale, notre approche sera divisée en deux parties, pour deux niveaux d'analyse. Dans la prochaine sous-section, nous regarderons le lien entre les types d'environnement décrits pour les ZE. Une première visualisation simple du lien entre environnement et paludisme, sans prendre les caractéristiques socio-économiques des ménages de la ZE, sera présentée. Ensuite, nous irons à un niveau plus fin, celui des ménages, afin d'inclure leurs caractéristiques socio-économiques. Ces caractéristiques, qui peuvent aussi avoir un effet sur le taux de paludisme, sont essentielles à prendre en compte afin de conclure sur la qualité du lien entre environnement et paludisme.

4.5.2 Environnement et paludisme au niveau de la ZE

Pour une première visualisation des effets de l'environnement sur la présence du paludisme avec notre carte, nous cherchons des corrélations simples entre les types d'environnement et le taux de paludisme. L'idée n'est pas de prédire directement la prévalence du paludisme à partir des distributions de LCZ, car cela masquerait les disparités socio-économiques en traitant potentiellement deux foyers (l'un riche et l'autre pauvre) comme égaux, malgré des facteurs socio-économiques tels que les moustiquaires, les types d'installations sanitaires, et d'autres qui peuvent influencer les résultats. Pour faire ces observations, nous avons regroupé

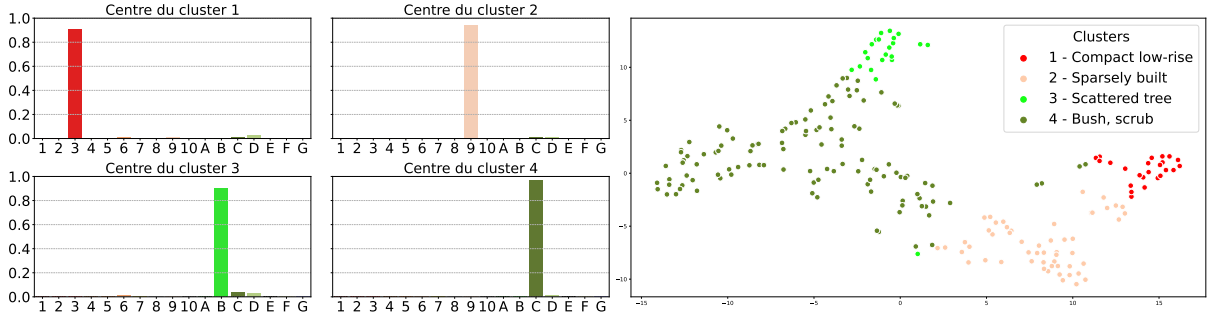


Figure 4.11: Distributions LCZ des centres des clusters et représentation 2D utilisant t-SNE [78]

la distribution LCZ des ZE en $n_E = 4$ types d'environnements en utilisant l'algorithme Fuzzy C-Means [89]. Les distributions LCZ des centres de chaque cluster, ou types d'environnement, sont représentées dans la Figure 4.11. La définition des clusters en utilisant la distribution LCZ va au-delà de la dichotomie urbain/rural et tire parti de la variété des classes dans le schéma de classification. Chaque cluster est fortement polarisé par une seule classe LCZ. Les clusters 1 et 2 tendent à être urbains, et les clusters 3 et 4 tendent à être ruraux. Le cluster 1 est très urbain et comprend principalement des LCZ *compact low-rise*. Ce cluster peut être associé aux villes. Le cluster 2 est moins urbain, principalement composé de *sparsely built*, comme on peut en trouver dans la périphérie des villes et dans la partie nord du pays. Les clusters 3 et 4 sont deux clusters ruraux dominés respectivement par *scattered trees* et *bush/scrub*. Les données MIS fournissent des résultats de tests rapides pour chaque enfant testé dans toutes les ZE. Dans cette sous-section, nous définissons le *taux de paludisme* comme le nombre de cas positifs parmi les enfants âgés de 6 à 59 mois divisé par le nombre total d'enfants âgés de 6 à 59 mois dans une ZE. Nous représentons dans la Figure 4.12 la distribution des taux de paludisme pour chaque type d'environnement. Visuellement, les taux de paludisme semblent associés au type d'environnement.

En moyenne, les ménages situés dans les clusters 1 et 2 présentent des taux de paludisme plus faibles que ceux des clusters ruraux 3 et 4. Au sein de ces clusters urbains, le plus urbain présente les taux les plus faibles. En milieu rural, le cluster 3 montre des taux légèrement plus élevés que le cluster 4, mais la différence n'est pas significative selon le test t de Student présenté dans le Tableau 4.3. Cette non-significativité peut s'expliquer par la surreprésentation de la classe *bush/scrub* sur la carte présentée dans la section 4.4.

P-valeurs	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1	1	$<$ 0,001	$<$ 0,001	$<$ 0,001
Cluster 2	x	1	0,004	0,033
Cluster 3	x	x	1	0,160
Cluster 4	x	x	x	1

Tableau 4.3: P-valeurs dans les résultats du test de Student.

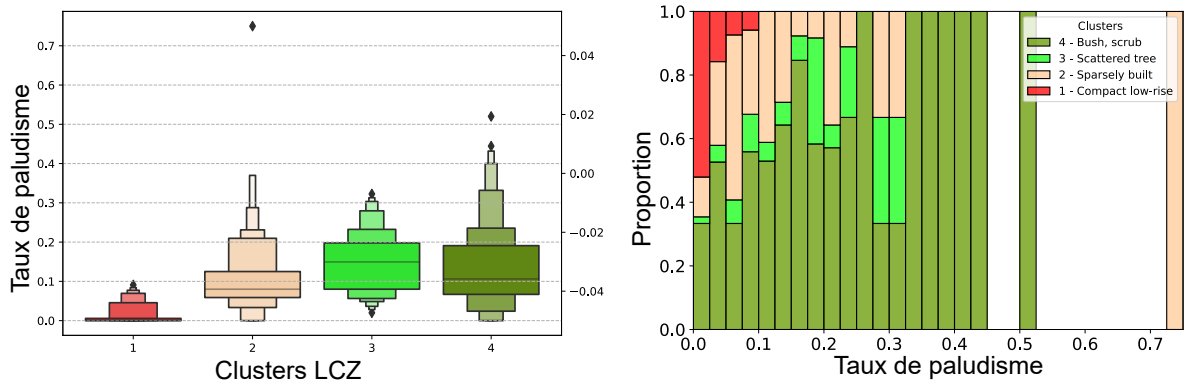


Figure 4.12: Distribution des taux de paludisme des ZE (gauche) et proportion des taux de paludisme par intervalles, regroupés par cluster.

4.5.3 Environnement et paludisme au niveau des ménages

Cette sous-section introduit l'étude du lien entre environnement et paludisme à un niveau plus bas, celui du ménage. Dans un premier temps, nous définissons la variable à expliquer : la présence de paludisme chez les enfants d'un ménage. Ensuite, nous montrons que l'influence de l'environnement sur le paludisme est significative, même en tenant compte des données socio-économiques des ménages sélectionnées pour cette étude. Ces caractéristiques sont décrites dans le Tableau 4.4.

Variable dépendante

Le Test Rapide de Diagnostic du Paludisme (RDT) pour tous les enfants testés est fourni dans les données MIS. Les RDT sont des tests sanguins qui détectent des antigènes spécifiques (protéines) produits par les parasites du paludisme dans le sang des personnes infectées. Pour étudier le paludisme au niveau des ménages, nous définissons maintenant notre variable dépendante comme la présence d'au moins un cas de paludisme positif, selon le RDT, parmi les enfants âgés de 6 à 59 mois dans le ménage. Par la suite, nous utilisons l'expression "ménage positif" pour désigner de tels ménages. Ainsi, cette variable est binaire et est expliquée à l'aide d'une régression logistique et de données liées au ménage. La régression logistique vise à estimer la probabilité de survenue d'un événement en fonction d'un ensemble donné de variables indépendantes. Dans ce travail, nous estimons la probabilité d'avoir au moins un cas positif dans un ménage en fonction des variables décrites ci-dessous. Ce modèle nous permet d'analyser les interactions de l'environnement avec d'autres variables au niveau du ménage.

Caractéristiques des ménages

Comme mentionné dans la description de l'enquête, les données MIS fournissent des informations socio-économiques sur les ménages en plus des résultats des tests de paludisme. Nos variables explicatives, au niveau du ménage, comprennent des informations sur la richesse, le

type de toilettes utilisées, la source d'eau potable, le niveau d'éducation de la mère des enfants et le nombre total d'enfants entre 6 et 59 mois dans le ménage :

- **L'indice de richesse** représente le bien-être financier. Dans cette étude, il est divisé en quintiles (plus pauvre, pauvre, moyen, riche, le plus riche).
- **Les types de toilettes** se répartissent en trois catégories selon le rapport des données MIS [90] : aucune installation, non amélioré (par exemple toilettes à seau, toilettes/latrines suspendues) et amélioré (par exemple toilettes à chasse d'eau reliées au système d'égouts, toilettes à compostage).
- **Les sources d'eau potable** ont également été réparties en catégories similaires : non améliorées (source ou puits non protégé, rivière), sources améliorées (par exemple, eau courante) et autres sources (par exemple, camion-citerne, puit protégé).
- **Le niveau d'éducation de la mère** ne peut pas être indiqué trivialement. Dans le contexte de l'étude, un ménage peut être composé de plusieurs enfants ayant différentes mères. Pour être aussi proche que possible de la réalité, nous définissons le niveau d'éducation des mères dans un ménage comme le niveau le plus représenté (pas d'éducation, primaire, secondaire et supérieur) parmi les mères présentes dans le ménage. Si plusieurs mères sont présentes dans le ménage, nous retenons le niveau d'éducation le plus élevé.
- **Le nombre d'enfants âgés de 6 à 59 mois** est également calculé et utilisé comme variable explicative.

Nous utilisons également les clusters décrits dans la section précédente comme notre variable environnementale. Le type d'environnement pour les ménages est déterminé par le cluster dans lequel leur ZE est classée. Notre carte générée a une résolution d'étiquette de 320m $\times$ 320m, ce qui peut être trop élevé pour classer les petits plans d'eau qui peuvent augmenter la population de moustiques dans la région. Pour compenser ce manque, nous ajoutons la présence d'eau dans la zone d'incertitude en utilisant les données OpenStreetMap. Le Tableau 4.4 donne un résumé des variables considérées. Dans ce tableau, "Taux+" est la proportion de ménages dans chaque catégorie où au moins un enfant entre 6 et 59 mois a été testé positif pour le paludisme.

Dans la suite de ces travaux, nous utilisons les données suivantes indiquées dans le Tableau 4.4.

Résultats

Après avoir supprimé les valeurs non valides, c'est-à-dire les ménages avec des données manquantes, il reste 4357 ménages avec au moins un enfant entre 6 et 59 mois dans l'ensemble de données de l'enquête. Leurs contextes socio-économiques sont donnés dans le Tableau 4.4. Tous ces ménages reçoivent une étiquette selon leur positivité au paludisme : 0 étant négatif et 1 étant positif. Les résultats de l'association univariée sont présentés dans le Tableau 4.5 et

Variables	Catégories	Nombre	Proportion (%)	Taux+ (%)	Prévalence (%)
Indice de richesse	Les plus pauvres	932	21,39	27,0	19,0
	Moins pauvres	939	21,55	25,0	18,0
	Moyen	897	20,59	23,0	17,0
	Plus riches	817	18,75	26,0	18,0
	Les plus riches	772	17,72	8,0	6,0
Type de toilettes	**Aucune installation**	1878	43,10	28,0	20,0
	Non amélioré	323	7,41	24,0	18,0
	Amélioré	2156	49,48	17,0	12,0
Source d'eau potable	**Aucune installation**	842	19,33	28,0	20,0
	Non amélioré	392	9,00	21,0	14,0
	Amélioré	3123	71,68	21,0	15,0
Niveau d'éducation	**Aucune éducation**	2934	75,68	26,0	18,0
	Primaire	512	13,21	21,0	16,0
	Secondaire +	431	11,12	10,0	8,0
Nombre d'enfants	**1**	2176	49,94	14,0	14,0
	2	1546	35,48	27,0	16,0
	3 et plus	635	14,57	40,0	20,0
Type d'environnement	Bush/scrub	2506	57.52	26.0	19.0
(Cluster LCZ)	Scattered Trees	386	8.86	26.0	18.0
	Sparsely built	991	22.75	20.0	14.0
	Compact low-rise	474	10.88	4.0	3.0
Eau dans la zone d'incertitude	**0**	1774	44,74	22,0	16,0
	1	621	15,66	24,0	18,0
	2 et plus	1570	39,60	22,0	15,0

Tableau 4.4: Caractéristiques socio-économiques et environnementales des ménages sélectionnés pour cette étude. Les catégories du ménage de référence sont en gras. Les noms des clusters LCZ sont donnés par la classe dominante.

les résultats de l'analyse multivariée, en utilisant toutes les variables explicatives, sont donnés dans le Tableau 4.6. Aucun poids n'a été utilisé pour équilibrer les foyers en fonction du nombre d'enfants de moins de 5 ans. L'effet du nombre d'enfants est contrôlé dans le modèle de régression final.

Association entre le paludisme et les variables socio-économiques

Pris indépendamment, la plupart des variables socio-économiques ont une association significative avec les ménages positifs. Le niveau d'éducation de la mère a une association négative avec la présence du paludisme : les ménages avec des femmes plus instruites sont moins susceptibles d'être des ménages positifs par rapport aux mères moins instruites. La source d'eau potable montre également une association similaire. Fait intéressant, seules les toilettes améliorées sont significativement associées aux ménages positifs, tandis que le fait d'avoir des toilettes non améliorées ne semble pas réduire les taux de paludisme par rapport à une absence de toilettes. De même, seul le quintile le plus riche de la population est associé à une présence moindre de paludisme. Contrairement à ce qui pourrait être attendu, les différences entre les autres quintiles ne sont pas significatives. Cela peut s'expliquer par la sélection des ZE pour l'enquête : les prévalences du paludisme indiquées dans Tableau 4.4 sont similaires pour les quatre quintiles les moins riches.

Association entre le paludisme et les variables environnementales

La présence d'eau dans la zone d'incertitude, c'est-à-dire la présence d'eau près des lieux de résidence des ménages, n'est pas significative. Les types d'environnements définis ci-dessus sont tous significatifs lorsque la variable de référence est le type *sparely built. Compact low-rise* est associé à des taux plus bas, *bush/scrub* a des taux plus élevés et *scattered trees* a les taux les plus élevés.

		5%	95%	Odds Ratio	P-valeurs	Significatif
Indice de richesse	Les plus pauvres	0,888	1,343	1,092	0,404	Faux
(ref. moins pauvres)	Moyen	0,740	1,134	0,916	0,420	Faux
	Plus riches	0,852	1,310	1,056	0,617	Faux
	Les plus riches	0,198	0,358	**0,266**	0,000	**Vrai**
Type de toilettes	Amélioré	0,437	0,592	**0,508**	0,000	**Vrai**
(ref. aucune installation)	Non amélioré	0,625	1,079	0,821	0,157	Faux
Source d'eau potable	Améliorée	0,556	0,785	**0,661**	0,000	**Vrai**
(ref. aucune installation)	Non améliorée	0,510	0,901	**0,678**	0,007	**Vrai**
Niveau d'éducation	Primaire	0,583	0,921	**0,732**	0,008	**Vrai**
(ref. aucune éducation)	Secondaire +	0,238	0,450	**0,327**	0,000	**Vrai**
Nombre d'enfants	1	0,364	0,507	**0,429**	0,000	**Vrai**
(ref. 2)	3+	1,521	2,243	**1,847**	0,000	**Vrai**
Type d'environnement	*Compact low-rise*	0,079	0,223	**0,133**	0,000	**Vrai**
(ref. *Sparsely built*)	*Scattered Trees*	1,091	1,860	**1,424**	0,009	**Vrai**
	Bush/scrub	1,169	1,669	**1,397**	0,000	Vrai
Eau dans la zone d'incertitude	0	0,914	1,404	1,133	0,256	Faux
(ref. 1)	2+	0,836	1,161	0,986	0,862	Faux

Tableau 4.5: Résultats des régressions logistiques univariées qui expliquent la positivité des ménages au Burkina Faso, 2017-2018 selon les données MIS, en considérant chaque variable séparément.

4.6 Discussion

Le travail présenté dans ce chapitre vise à estimer l'impact de l'environnement sur un problème majeur de santé en utilisant une enquête à grande échelle. Trois parties ont été détaillées : la caractérisation des données environnementales locales à partir de données satellites disponibles gratuitement, le lien de cette carte avec des données démographique collectées à partir d'enquêtes basées sur les ménages fournissant une géolocalisation approximative des ménages enquêtés, et l'évaluation de l'interaction entre l'environnement et la santé, en tenant compte des caractéristiques socio-économiques des ménages.

Dans la sous-section 4.6.1, nous discuterons des limites de la méthode de cartographie LCZ en utilisant l'adaptation de domaine présentée en section 4.4, et dans la sous-section 4.6.2 de l'impact de l'environnement sur le paludisme décrit en section 4.5.

		5%	95%	Odds Ratio	P-valeurs	Significatif
Interception		0,313	0,573	0,**423**	0,000	**Vrai**
Indice de richesse	Les plus pauvres	0,859	1,379	1,088	0,484	Faux
(ref. Moins pauvres)	Moyen	0,788	1,279	1,004	0,975	Faux
	Plus riches	0,863	1,420	1,107	0,422	Faux
	Les plus riches	0,490	1,048	0,716	0,086	Faux
Type de toilettes	Amélioré	0,620	0,918	**0,754**	0,005	**Vrai**
(ref. aucune installation)	Non amélioré	0,622	1,152	0,846	0,288	Faux
Source d'eau potable	Améliorée	0,650	**0,965**	0,792	0,021	**Vrai**
(ref. aucune installation)	Non améliorée	0,567	1,061	0,776	0,112	Faux
Niveau d'éducation	Primaire	0,863	1,427	1,110	0,416	Faux
(ref. aucune éducation)	Secondaire +	0,501	1,026	0,717	0,069	Faux
Nombre d'enfants	1	0,433	0,629	**0,522**	0,000	**Vrai**
(ref. 2)	3+	1,421	2,180	**1,760**	0,000	**Vrai**
Type d'environnement	*Compact low-rise*	0,162	0,513	**0,288**	0,000	**Vrai**
(ref. *Sparsely built*)	*Scattered trees*	1,191	2,128	**1,592**	0,002	**Vrai**
	Bush/scrub	1,187	1,777	**1,453**	0,000	**Vrai**
Eau dans la zone d'incertitude	1	0,842	1,345	1,064	0,603	Faux
(ref. 0)	2+	0,815	1,172	0,977	0,806	Faux

Tableau 4.6: Résultats d'une régression logistique multivariée. Le modèle explique la positivité du paludisme des ménages au Burkina Faso, 2017-2018 selon les données MIS en tenant compte de toutes les variables.

4.6.1 Cartographie LCZ

La méthode présentée dans ce chapitre vise à cartographier les LCZ dans un pays où aucune vérité terrain est disponible, en utilisant des méthodes d'apprentissage profond contrastif appliquées aux images Sentinel-2. Elle est basée sur les variations saisonnières du pays cible pour extraire des informations utiles pour la cartographie des LCZ. Pour cette étude, nous avons choisi de travailler sur le Burkina Faso où une enquête démographique et sanitaire récente a été menée et où des données collectées ainsi que des données de géolocalisation étaient disponibles pour une telle analyse. Un jeu de données, composé d'images Sentinel-2 des saisons sèches et des pluies, a alors été constitué. Cette inclusion d'un seul pays dans un tel jeu de donné limite la capacité de généralisation du modèle entraîné. En effet, sa performance de classification sur l'ensemble de données So2Sat diminue de 40%. Finalement, l'étape d'adaptation de domaine semble avoir inversé le problème d'origine : le modèle est désormais **uniquement** concentré sur le pays cible et ne peut pas être généralisé à d'autres domaines. Cependant, la création d'ensembles de données d'entraînement sur d'autres pays (zones ayant des caractéristiques climatiques similaires) devrait permettre de limiter cette dégradation des résultats d'adaptation de domaine.

Le déséquilibre de distribution entre les classes LCZ source et cible peut entraîner aussi un biais dans la classification résultante. Le jeu de données So2Sat a été construit sur des villes et leurs environs, ce qui peut limiter la caractérisation des environnements plus ruraux. Les zones plus rurales, donc loin des régions urbaines, ne sont pas suffisamment représentées dans le jeu de données et seront donc plus difficiles à classer. Pour essayer de limiter l'impact de ce déséquilibre, le jeu de données saisonnier créé pour ces travaux comprend des données

de toute la région du Burkina Faso, des zones très urbaines aux zones très rurales, mais les performances du modèles s'en trouvent limitées. Par ailleurs, certaines régions spécifiques au Burkina Faso, avec un climat extrême, ne sont pas prises en compte dans So2Sat. Par exemple, les zones désertiques sahariennes au nord ne sont pas présentes dans So2Sat et sont donc plus difficiles à classer avec précision. Les cartes générées à l'aide de ces données peuvent altérer le résultat de l'analyse multivariée dans les zones rurales, car la population au Burkina Faso est principalement rurale.

L'OA étant de 56% suggère que la précision des cartes LCZ pourrait être améliorée. Cependant, avec 17 classes, ces résultats sont bien supérieurs à un modèle aléatoire et aux autres cartes disponibles dans la communauté scientifique. Les erreurs de classification sont par ailleurs lissées par la sélection aléatoire des pixels effectuée dans 4.5.1. Les résultats finaux sur le lien entre l'environnement et le paludisme sont en accord avec nos attentes.

4.6.2 LCZ et paludisme

L'utilisation des LCZ comme indicateur environnemental et la génération d'une telle carte a permis la mise en évidence d'une association significative entre l'environnement et le paludisme au Burkina Faso. Cette association reste significative lorsque des facteurs démographiques socio-économiques, tels que le nombre d'enfants dans le ménage, le type de source d'eau potable, les types de toilettes, le niveau d'éducation de la mère et la présence de plans d'eau, sont inclus dans l'analyse comme variables de contrôle.

L'utilisation de certaines variables explicatives dans cette analyse a abouti à des résultats inattendus. Les populations les plus pauvres sont souvent associées à des taux de paludisme plus élevés en raison du manque d'infrastructures (centre médical, toilettes améliorées) et de connaissances [91], [92] sur la prévention de la maladie. Cependant, l'association entre l'indice de richesse et les taux de paludisme n'a pas été retrouvée, à part pour les ménages les plus riches. Une explication possible est l'échantillonnage des ZE. Comme indiqué dans le Tableau 4.4, celui-ci a conduit à des prévalences du paludisme similaire pour les populations des 4 quintiles de richesse les plus bas (les plus pauvres, les plus pauvres, moyennes et plus riches).

De plus, la présence de plans d'eau dans la zone tampon donne également des résultats paradoxaux. En général, les moustiques sont plus fréquents dans les régions humides, et où il y a de l'eau stagnante comme des flaques ou des étangs. En raison de l'incertitude sur la position des ménages, il n'est pas possible de calculer la distance des ménages aux petits plans d'eau, indiqués sur les données OSM par exemple, ce qui peut indiquer une plus grande présence de moustiques et donc de plus grande chances d'être en contact avec un moustique contaminé. Les données MIS fournissent également une variable "proximité à l'eau" pour chaque ZE, basée sur des bases de données internationales telles que l'ensemble de données des lacs et l'ensemble de données du littoral [93]. Cette variable ne comprend pas la proximité des petits plans d'eau comme, comme celles que l'on peut trouver dans les données OSM, donc ne donne pas d'information supplémentaire sur la proximité des plans d'eau.

Les associations entre les ménages positifs et les types de toilettes ou les sources d'eau potable conduisent à des résultats similaires, et qui peuvent être expliqués de la même

manière. Les associations avec les ménages positifs ne sont significatives que pour les variables "améliorées" en prenant comme référence l'absence d'installation. En effet, les sources d'eau améliorées sont protégées et associées à une présence moindre du paludisme. Dans le cas des sources d'eau potable, les installations améliorées limitent la quantité d'eau stagnante, réduisant la présence de moustiques. Cependant, les sources non améliorées (non protégées) ou l'absence de sources conduisent, à l'inverse, à une plus grande présence de moustiques. Le même raisonnement peut être appliqué aux types de toilettes utilisées, car les installations améliorées sont plus protégées et réduisent la présence de moustiques. Dans cette étude, il n'y a cependant aucune preuve que le fait d'avoir des installations non améliorées contribue à la réduction de la présence du paludisme dans les ménages, par rapport aux ménages sans installations du tout.

Les zones avec le plus fort risque de paludisme peuvent être trouvées dans la partie sud du pays, principalement couverte par des zones de *Scattered trees*. Cependant, nous avons également constaté que l'utilisation des régions climatiques (sud, centre, nord) ne conduit pas à des associations significatives. Les caractéristiques de l'environnement local des ménages restent invariantes lors de l'inclusion d'autres covariables explicatives. Il est intéressant de noter que dans les deux analyses de régression logistique univariée et multivariée, la relation entre les types d'environnements et la prévalence du paludisme est significative. Cette double significativité renforce la preuve de son association avec la présence du paludisme, indiquant une relation robuste et persistante même après avoir contrôlé les caractéristiques socio-économiques des ménages. Ce résultat suggère qu'il possède un effet véritable et indépendant sur la positivité au paludisme. Cette significativité est cependant perdue entre les deux types d'environnements ruraux (*Scattered Trees* et *Bush/scrub*) lorsqu'un des deux est utilisé comme référence. Comme indiqué dans le Tableau 4.3, ces deux clusters ne sont pas significativement différents, cette non-significativité était attendue. Les deux classes LCZ peuvent être trouvées dans des régions similaires du pays, et les zones de *Bush/scrub* sont très représentées sur la carte. De même, une autre classification du nord du Burkina Faso que les zones inhabituellement peu construites peut conduire à des clusters différents. En particulier, plus de zones *Bare soil or sand* ou *Low plants* étaient attendues. Une classification plus précise de certaines zones pourrait affiner les résultats de la régression logistique. Les conclusions de ce travail devraient rester les mêmes, car les régions du nord ne souffrent pas de taux élevés de paludisme, du fait du climat aride.

Cette étude suggère que l'utilisation du système de classification LCZ est appropriée pour l'analyse des données de population, après la génération d'une carte avec une résolution de 320 m. Passer à une résolution plus fine, ce qui donne une carte plus précise (comme pour la carte LCZ mondiale [59]), pourrait aboutir à des résultats sensiblement différents, sans modifier les conclusions. Des études supplémentaires sont nécessaires pour analyser l'effet de la résolution de la carte sur les résultats des données de population.

4.6.3 Végétation et paludisme

La présence de paludisme est souvent associée à la présence de végétation. Dans cette étude, nous avons dépassé cette première idée et avons utilisé une représentation beaucoup plus complexe de l'environnement, prenant aussi en compte les infrastructures humaines. Pour

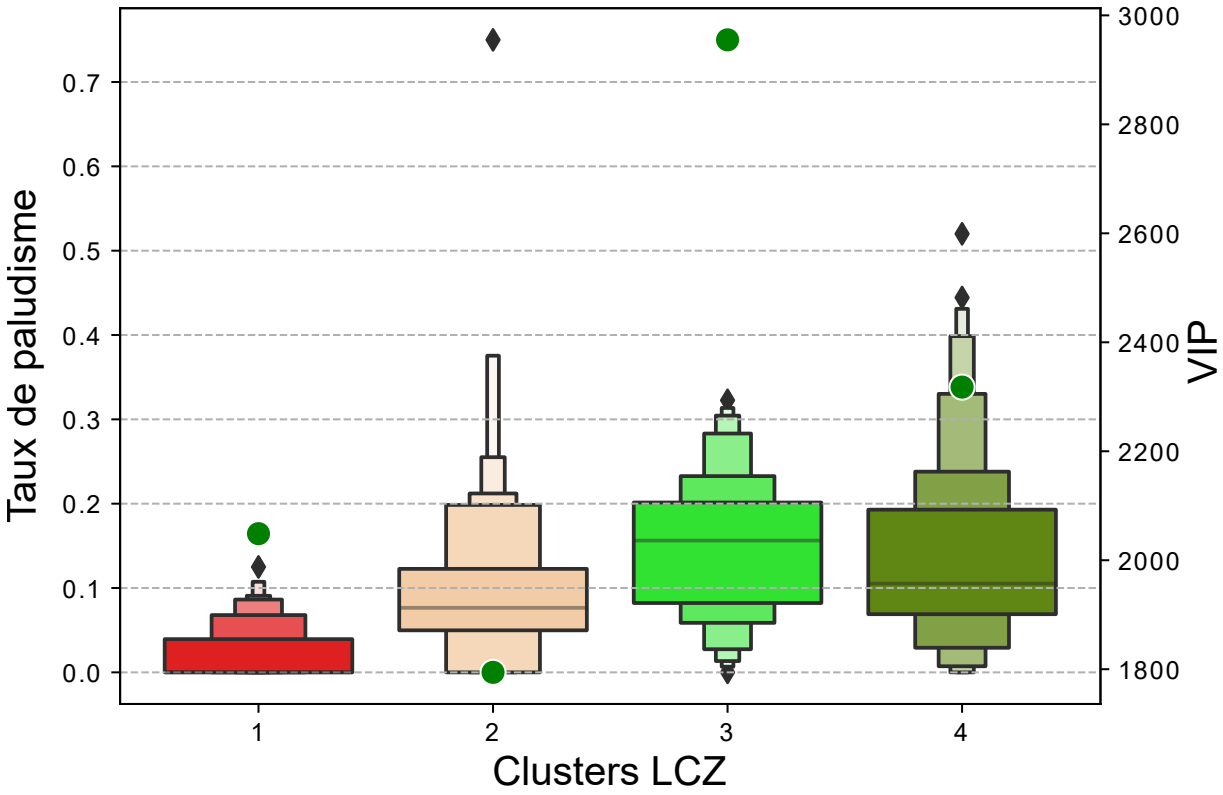


Figure 4.13: Taux de paludisme en fonction du type d'environnement, comme défini en sous-section 4.5.2. Les valeurs issues de VIP, les points verts, ont été ajoutées à ce graphe.

valider l'utilisation de cette représentation plus complexe, cette partie étudie l'interaction entre l'environnement et le paludisme, si l'environnement est uniquement modélisé par sa végétation. Pour se rapprocher le plus possible des données à disposition des chercheurs, nous utiliserons dans cette partie les données de végétation fournies dans l'enquête MIS, issues des données Vegetation Index and Phenology (VIP)[6]. Ces données de végétation sont dérivées du NDVI et du *Enhanced Vegetation Index* calculés à partir des données MODIS. Les valeurs indiquées dans les données de l'enquête sont données pour l'année 2015 et moyennées par zone d'incertitude autour du centroïde de chaque ZE.

Pour une première visualisation de ce nouvel indicateur environnemental, la Figure 4.13 affiche les valeurs de VIP par type d'environnement comme déterminés en sous-section 4.5.2, conjointement avec le taux de paludisme. La première chose à noter est l'absence de corrélation entre l'indice de végétation et le taux d'urbanisation qui apparaît avec les types d'environnement. Ce graphique confirme que la classification LCZ, dans ce contexte, dépasse la caractérisation unique de la végétation et fournit des informations supplémentaires.

Pour pousser plus loin l'analyse, nous remplaçons chaque type d'environnement par les quartiles de valeurs de VIP : (0) Le moins vert, (1) moins vert, (2) plus vert et (3) le plus vert. Ainsi, il sera possible d'estimer le liens entre paludisme est végétation à l'aide d'une analyse multivariée, donc en prenant en compte les caractéristiques socio-économiques des ménages. Les résultats de cette nouvelle analyse multivariée sont résumés dans le tableau 4.7. Cette

[6]https://lpdaac.usgs.gov/products/vipphen_evi2v004

		5%	95%	Odds Ratio	P-valeurs	Significatif
Intercept		0,334	0,600	**0,447**	0,000	**Vrai**
Indice de richesse	Les plus pauvres	0,834	1,343	1,058	0,641	Faux
(ref. Moins pauvres)	Moyen	0,800	1,301	1,020	0,873	Faux
	Plus riches	0,857	1,412	1,100	0,454	Faux
	Les plus riches	0,288	0,589	**0,412**	0,000	**Vrai**
Type de toilettes	Améliorée	0,603	0,897	**0,735**	0,002	**Vrai**
(ref. aucune installation)	Non améliorée	0,640	1,188	0,872	0,384	Faux
Source d'eau potable	Améliorée	0,666	0,993	**0,813**	0,042	**Vrai**
(ref. aucune installation)	Non améliorée	0,553	1,038	0,757	0,084	Faux
Nombre d'enfants	1	0,433	0,628	**0,521**	0,000	**Vrai**
(ref. 2)	3	1,441	2,210	**1,785**	0,000	**Vrai**
Niveau d'éducation	Primaire	0,824	1,359	1,058	0,656	Faux
(ref. aucune éducation)	Secondaire +	0,463	0,946	**0,662**	0,023	**Vrai**
Végétation	quartile 1 (<25%)	0,730	1,216	0,942	0,646	Faux
(ref. quartile 2 (25%< ... < 50%)	quartile 3 (50%< ... < 75%)	1,289	2,027	**1,617**	0,000	**Vrai**
)	quartile 4 (>75%)	1,165	1,840	**1,464**	0,001	**Vrai**
Eau dans la zone d'incertitude	1	0,824	1,320	1,043	0,726	Faux
(ref. 0)	2+	0,818	1,174	0,980	0,827	Faux

Tableau 4.7: Résultats d'une régression logistique multivariée en utilisant l'indice de végétation fournie dans l'enquête MIS comme indicateur environnemental. Le modèle explique la positivité du paludisme des ménages au Burkina Faso, 2017-2018 selon les données MIS en tenant compte de toutes les variables.

représentation ne permet pas d'avoir des résultats significatifs pour tous les environnements. En particulier, il n'est pas possible de percevoir les changements entre les deux quartiles avec le moins de végétation, correspondant aux zones d'énumérations avec des indices VIP les plus faibles. Ces deux observations permettent de valider l'apport de la nouvelle représentation de l'environnement, par les LCZ, pour l'étude du paludisme au Burkina Faso. Pour valider totalement cette approche, il faudrait l'appliquer dans différents contextes (donc différents pays) et sur des sujets différents.

4.7 Conclusion

L'objectif de ce chapitre était d'étudier l'utilisation des images satellites pour assister l'analyse des données d'enquête de population à grande échelle. Plus précisément, une carte LCZ du Burkina Faso a été générée pour intégrer l'environnement dans l'analyse des données de l'enquête MIS sur le paludisme, de 2017/2018. Deux contraintes, spatiales et temporelles, ont été identifiées. D'abord le modèle devait être adapté au Burkina Faso afin de produire une cartographie précise de son environnement. Ensuite, l'environnement devait être caractérisé pendant la période de l'enquête pour assurer la cohérence temporelle. Pour générer une telle carte, une méthode d'entraînement d'un réseau de neurones a été introduite. Cette stratégie d'adaptation de domaine, basée sur les caractéristique climatiques du pays cible et couplé à l'utilisation d'images satellites disponibles à une grande fréquence temporelle, a permis la génération d'une carte LCZ du Burkina Faso pour le début d'année 2018. Les informations environnementales des ménages sondés ont été extraites de cette carte, en consid-

érant l'incertitude autour de leurs positions réelles. Ces informations ont pu ensuite être intégrées dans une analyse multivariée pour démontrer le rôle significatif de l'environnement, représenté par les LCZ, dans la positivité des ménages au paludisme.

Deux conclusions principales peuvent être tirées de ces travaux. D'abord, d'un point de vue télédétection, la prise en compte des caractéristiques climatiques du pays cible a permis de transférer l'apprentissage des LCZ avec succès. Certaines classes restent cependant difficiles à classer, en particulier pour les classes impliquant de faibles taux de bâtis. Un questionnement sur la résolution des images en entrée d'un modèle permettrait de trouver un compromis concernant la précision de la classification. La résolution de Sentinel-2 est insuffisante, avec les méthodes actuelles, pour la classification de certaines classes LCZ. L'utilisation de ces images permet cependant de mettre à jour les cartes très fréquemment avec un coût de calcul réduit. L'accès à ces tuiles est en plus libre, permettant une meilleure reproductivité des résultats. Des images à hautes résolutions, comme les images PlanetScope[7] disponibles à trois mètres de résolution, permettraient une classification plus précise en augmentant le coût de calcul. Cette prise en compte pour l'extraction d'information pourrait être utilisée pour d'autres applications, comme pour la segmentation sémantique ou la détection d'objet. D'un point de vue démographie, l'utilisation de données environnementales plus complexes, générées avec de l'apprentissage profond, a permis une meilleure analyse de la relation entre environnement et paludisme. Plus précisément, le lien entre l'environnement modélisé par le système de classification LCZ et la positivité des ménages au paludisme est significatif, pour chaque type d'environnement. Ces types d'environnement vont au delà de la caractérisation de l'environnement uniquement par la végétation, qui est le plus souvent utilisé dans ce contexte. Cela souligne l'importance d'une considération globale de l'environnement pour ces types d'analyses. D'autres travaux sur d'autres applications devraient permettre de conclure sur l'apport général d'un système de classification complexe de l'environnement dans l'analyse de données de population.

[7] https://developers.planet.com/docs/data/planetscope/

Chapitre 5

Conclusion

P'tit Mario Kart ?

– Logan Servant

Bilan

Les travaux présentés dans cette thèse s'inscrivent dans le développement récent de l'utilisation des images satellites dans de nouveaux domaines d'application. En particulier, elles sont utilisées ici comme un moyen de caractériser l'environnement pour assister des études sur la population dans le contexte de l'Afrique subsaharienne. Deux études ont été menées, à différents niveaux d'analyse. La première, à l'échelle d'une ville, montre l'importance d'une représentation fine de l'environnement, générée à partir des images satellites, dans l'analyse des causes de décès. En particulier, nous montrons que le système de classification LCZ permet d'affiner l'analyse de données de population. La deuxième, à l'échelle d'un pays, montre l'intérêt de prendre en compte une représentation fine de l'environnement, au delà de la seule dichotomie urbain/rural, dans l'étude du paludisme.

La première partie traite du lien entre environnement et causes de mortalité à Antananarivo, la capitale de Madagascar. Elle propose deux contributions, une en vision par ordinateur et une autre en démographie. Une méthode d'adaptation de domaine a été développée pour cartographier les LCZ à Antananarivo. Cette méthode est basée sur la définition des LCZ basée sur des intervalles de valeurs acceptables pour des propriétés physiques et géométriques des zones à caractériser. Cette adaptation a pour but de transférer l'apprentissage du modèle à partir d'un domaine source étiqueté vers un domaine cible non étiqueté, ici Antananarivo. Après avoir prédit ces descripteurs physiques directement, ces derniers ont été utilisés comme prototypes, c'est-à-dire comme références pour les classes entre les deux domaines. La méthode intègre ces prototypes en calculant l'attention entre les caractéristiques extraites par le modèle et ces prototypes, qui sont donc fixes. Cette méthode, ainsi que la méthode à l'état de l'art de l'adaptation de domaine des LCZ, ne permettent cependant pas de cartographier

suffisamment précisément Antananarivo, car l'écart entre les deux domaines est trop important. Dans un deuxième temps, plusieurs caractéristiques de l'environnement ont été utilisées pour expliquer les données de mortalité de la ville d'Antananarivo, en contrôlant les données socio-économiques, avec comme unité spatiale les quartiers de la ville. Les modèles décrits dans ces premiers travaux montrent que l'altitude et les LCZ ont un effet significatif sur la mortalité par cause, notamment infectieuse et liée à l'eau.

La deuxième partie de cette thèse étudie la relation entre environnement et présence de paludisme au Burkina Faso. Deux contributions ont ici aussi été faites en vision et en démographie. Une autre méthode d'adaptation de LCZ a été développée pour correspondre au contexte d'un pays entier. Cette méthode d'entraînement utilise les caractéristiques saisonnières du pays pour extraire de l'information. Deux adaptations étaient ici nécessaires : l'adaptation d'une région d'une monde vers une autre comme pour la première partir (la région cible étant ici le Burkina Faso), mais aussi le transfert d'un modèle ayant été entraîné majoritairement sur les zones urbaines du jeu de données So2Sat vers des régions rurales du Burkina Faso. Les deux saisons sont comparées à l'aide d'un coût contrastif, qui ajoute une pénalité à l'apprentissage supervisé classique et permet d'effectuer le transfert. Un tel entraînement a permis la génération d'une carte LCZ du Burkina Faso en début d'année 2018, ce qui correspond à l'enquête démographique utilisée pour mesurer l'intensité et la répartition du paludisme. Pour la contribution démographique, seules les LCZ sont utilisées pour caractériser l'environnement. Comme l'enquête démographique sur le paludisme ne fournit pas les coordonnées précises des ménages, des pixels de la carte LCZ sont sélectionnés selon une stratégie semi-aléatoire dans la zone d'incertitude autour de ces coordonnées. De même que pour la première partie de la thèse, l'environnement est intégré dans un modèle conjointement avec les données socio-économiques des ménages, afin d'assurer l'indépendance des effets trouvés. L'environnement, caractérisé par les LCZ uniquement, permet d'expliquer la présence de paludisme dans les ménages Burkinabè. Cette contribution souligne l'importance d'une représentation détaillée de l'environnement, dépassant la simple dichotomie urbain/rural, et intégrant simultanément les éléments urbains et ruraux pour analyser les maladies vectorielles liées à l'environnement immédiat des ménages.

Ces deux études nous ont permis d'étudier différents indicateurs environnementaux pour la démographie, et de former une chaîne de traitement depuis les données satellites brutes jusqu'aux analyses de population. Pour traiter les données avec de l'apprentissage profond dans le contexte de l'Afrique subsaharienne, le développement de méthodes d'adaptation est essentiel étant donné qu'il y a peu de donnés disponibles. Ces caractéristiques extraites peuvent ensuite être utilisées dès lors qu'il existe une référence spatiale pour les données démographiques. Ces travaux montrent l'importance de prendre en compte l'environnement dans les analyses en démographie et indiquent que les LCZ semblent être appropriées pour cette application, même par rapport à des indicateurs plus classiques comme le NDVI.

Perspectives

Ces travaux étant à l'interface de domaines de recherche très distants, l'apprentissage automatique, la télédétection et la démographie, ils soulèvent des perspectives dans chacun d'entre eux. Cette partie présente les perspectives d'abord pour la partie vision par ordinateur, re-

groupant l'apprentissage automatique et la télédétection, et ensuite pour la démographie.

Concernant la partie vision par ordinateur, les perspectives sont principalement en adaptation de domaine pour la génération efficace de cartes environnementales. Nous l'avons vu dans la première partie, l'adaptation de domaine pour la classification des LCZ est une tâche compliquée. L'intégration des descriptions physiques des LCZ dans l'apprentissage n'a pas permis d'améliorer cette adaptation. Plusieurs points peuvent être modifiés dans les mécanismes utilisés, mais il convient premièrement de déterminer si les prototypes, même appris, permettent d'améliorer la classification des LCZ avec adaptation de domaine. De tels modèles d'adaptation pourraient permettre de créer des outils pour générer automatiquement des cartes de l'environnement, ce qui aiderait fortement les chercheurs en démographie. Les enquêtes démographiques étant limitées dans le temps, il est essentiel de pouvoir générer des cartes environnementales à des moments précis. Les modèles doivent donc être performants, quelle que soit l'année des images satellites utilisées en entrée, afin de produire des cartes suffisamment précises. Les cartes actuelles, comme le projet *Worldcover*[71] , sont générées tous les ans, mais cette fréquence est trop faible pour certaines applications, en particulier pour les applications de santé qui peuvent dépendre des saisons.

Concernant la démographie, ces travaux ont soulevé des perspectives nouvelles concernant l'usage des données environnementales, en particulier sur la santé des populations, appréhendée ici à travers différentes familles de causes de décès, dans le cas de la ville d'Antananarivo, et le paludisme, dans le cas du Burkina Faso. Cependant, d'autres domaines d'étude pourraient être affectés par des phénomènes environnementaux pouvant être caractérisés ou détectés à partir d' images satellites. Nous pouvons notamment penser à l'analyse de la migration des populations, notamment après des phénomènes climatiques extrêmes qui peuvent être très courts (inondations) ou beaucoup plus longs (sécheresse). Par ailleurs, maintenant que des données sont disponibles très fréquemment et partout dans le monde, des images satellites de sites dans lesquels un suivi démographique régulier existe déjà peuvent être utilisées et valorisées. Un suivi de l'environnement sur plusieurs années pourrait révéler des liens entre certains phénomènes et les changements environnementaux. D'autres analyses, plus méthodologiques sont aussi nécessaires. Dans ces travaux, nous avons utilisé une caractérisation LCZ avec une résolution de 320m et des images Sentinel-2 d'une résolution de 10m. D'autres études sur la résolution spatiale des cartes générées doivent être effectuées afin de trouver une résolution optimale, bien-sûr dépendant de l'application considérée.

Dans cette thèse, nous avons étudié deux phénomènes de population, l'un à l'échelle locale, sur un temps long (plusieurs années) et l'autre à l'échelle globale, sur un temps court (quelques mois). L'environnement d'Antananarivo a été décrit spécifiquement pendant la période d'étude, mais pour une seule date. L'évolution de l'environnement pendant cette période n'est donc pas prise en compte. Certaines données, issues des observatoires de population, sont disponibles depuis des années, voire plusieurs dizaines d'années comme nous l'avons vu dans l'introduction du chapitre 4. Des images avec des résolutions moyennes sont librement accessibles, comme les données Sentinel-2, mais leur résolution peut limiter les champs d'applications possibles, en particulier dans des zones restreintes, bien qu'elles soient accessibles sur des temps longs. Une première idée serait d'augmenter la résolution (tâche de super-résolution), domaine qui est déjà très étudié par la communauté scientifique et appliqué aux images Sentinel-2. Cela implique d'augmenter la résolution des images une à une, pendant

une longue période. Cependant, cette nouvelle dimension temporelle apporte avec elle une nouvelle contrainte : la cohérence temporelle. Cette cohérence est essentielle pour la caractérisation de l'environnement, au cours du temps et en relation avec les données de population. Augmenter la résolution des images séparément pourrait rompre cette cohérence. Des travaux préliminaires ont été menés, mais pas achevés par manque de temps, pour résoudre ce problème. L'idée était d'utiliser les réseaux de diffusion, introduits par Sohl-Dickstein *et al.* (2015) et Ho *et al.* (2020), comme moyen pour augmenter la résolution spatiale d'une série temporelle d'images. Dans notre application, nous considérons une série temporelle d'images à haute résolution comme un processus de diffusion, en nous reposant sur l'hypothèse que deux images d'un même endroit à des temps différents, sans changement, sont séparées par un bruit gaussien [94]. A chaque étape de la diffusion, nous conditionnons le débruitage par une image Sentinel-2. Ainsi, une image à l'étape t de la diffusion va dépendre d'une image réelle, à basse résolution, et de la reconstruction de l'étape de diffusion $t-1$, assurant la cohérence temporelle.

Les diverses applications et dimensions de la démographie, lorsque liées à l'environnement, créent de nouvelles opportunités pour la détection et la vision par ordinateur. Dans ces travaux, nous avons utilisé l'environnement comme une variable explicative au service de la démographie, et nous avons vu des corrélations intéressantes, y compris en prenant en compte des caractéristiques socio-économiques des populations considérées. En intégrant des données démographiques avec des techniques avancées d'intelligence artificielle, des modèles prédictifs pourraient être conçus pour estimer (et non plus expliquer) des données démographiques dans des zones dans lesquelles elles sont rares. Ces modèles devront être explicables, ce qui signifie qu'il sera possible, pour chaque décision prise, d'identifier les caractéristiques des données mobilisées et par quels moyens. Cela permettra de contrôler la valabilité et d'assurer la cohérence des décisions. Ces estimations pourront ensuite être combinées et ajustées avec les données d'enquêtes pour améliorer la précision des indicateurs et de leurs projections.

Retour d'expérience

Cette section est volontairement conçue pour être un peu plus ouverte et pour partager mon expérience personnelle du passage de l'informatique à la démographie. Les remarques et recommandations que je vais faire sont séparées en deux parties. D'abord, je discuterai de l'aspect lié au sujet développé ici, et dans un second temps, je traiterai de l'organisation d'un tel projet entre ces deux domaines.

La première chose est de se renseigner précisément sur le contexte du pays. J'avais étudié son contexte environnemental et la topographie du pays, mais il est important d'aller plus loin, en particulier dans certains contextes locaux. Le cas de la ville de Ouagadougou est un bon exemple. Comme indiqué dans la Figure 4.9, des zones de bâti informel sont situées en périphérie de la ville. La stratégie de développement urbain tend à construire de nouvelles infrastructures dans les périphéries, à la place et autour de ces zones informelles. Des mouvements de population dans la ville ont régulièrement lieu pour passer d'une zone informelle à une autre, ce qui renforce l'aspect temporaire de ces habitations légères. Ces connaissances sur un contexte plus précis permettent, dans un premier temps, une meilleure sélection du système de classification — au même titre que le choix de l'échelle utilisée — et, dans un second temps, donnent

des pistes d'amélioration du modèle entraîné et des classes les plus importantes du point de vue de la population. Cette connaissance a priori du terrain aurait permis, dès le début de l'étude sur le Burkina Faso, une attention particulière à la différenciation des zones urbaines denses, mais en bâti formel, des zones de bâti informel, ce qui aurait eu un effet important sur la population. Les LCZ étaient adaptées à ce cas de figure en considérant la classe *Lightweight low-rise*, mais d'autres classifications, ne considérant que les zones urbaines, auraient manqué une information importante sur le bâti. Cette attention particulière au contexte local a été faite pour la cartographie d'Antananarivo, dont le projet a commencé en deuxième partie de cette thèse après le travail sur le Burkina Faso. La période sur place et la visite des quartiers de la ville ont ouvert la réflexion autour des LCZ, qui caractérisent l'environnement par des zones de 320 m $\times$ 320 m ou 100 m $\times$ 100 m. Dans ces quartiers, les zones formelles et non formelles sont mélangées, dans des aires plus petites que les aires LCZ. Bien que donnant une information importante, les LCZ ne permettent pas une précision optimale. Ce manque d'information peut être comblé par l'utilisation de plusieurs indicateurs environnementaux pour l'analyse de données de population, qui est le deuxième point à retenir. Pour combler cette imprécision laissée par les LCZ, nous avons utilisé le score de confiance des cartes de détection de bâti proposées par Google, qui ont permis de représenter le caractère formel des habitations. De même, nous avons vu que le NDVI apporte une interprétation différente des résultats dans le travail sur le paludisme au Burkina Faso.

Ces travaux ont montré l'intérêt d'utiliser les images satellites pour caractériser finement l'environnement dans l'analyse des données de population, en plus des indicateurs classiques comme l'altitude et la végétation. Plusieurs questions se posent maintenant concernant la génération de ces données fines et leurs utilisations possibles. Les scientifiques issus de la démographie, qui n'ont a priori pas de compétences spécifiques en télédétection, peuvent facilement se former à l'utilisation de certaines données comme les indices spectraux. De nombreux sites gratuits permettent de télécharger ces indicateurs directement, sans contrainte temporelle. Il reste ensuite à utiliser un langage de programmation, comme R ou Python, pour relier les données ensemble, et ces usages se démocratisent de plus en plus grâce aux bibliothèques disponibles. Certaines caractérisations fines de l'environnement peuvent être utilisées de la même manière que les indices, comme des cartes de couverture et d'utilisation du sol déjà mises à disposition par la communauté scientifique [58], [59], [70], [71]. Cependant, ces données sont restreintes temporellement et peuvent ne pas prendre en compte des variations saisonnières (comme les saisons sèches et les pluies) ou des événements climatiques soudains (comme les inondations et les sécheresses). Dans ces cas, des cartes spécifiques devront être générées et nécessiteront de nouvelles compétences pour cartographier les zones « à la main » ou via des modèles, comme des modèles d'apprentissage. L'appel à un expert en environnement ou en apprentissage automatique sera nécessaire pour travailler conjointement avec les démographes. À l'inverse, les connaissances apportées par les chercheurs en démographie sont intéressantes pour l'élaboration de systèmes de classification adaptés. Les LCZ sont complètes car elles cartographient les zones urbaines avec précision. Elles sont donc orientées vers une utilisation en zone urbaine, ce qui peut limiter leur usage dans le contexte d'une enquête à l'échelle d'un pays, comme nous l'avons vu dans le chapitre sur le Burkina Faso. Une meilleure prise en compte du type de végétation, au-delà de la densité et de l'aspect global, serait un avantage. Ces conseils peuvent être apportés par les démographes, qui ont une meilleure connaissance du terrain et des caractéristiques locales qui peuvent jouer un rôle important dans les phénomènes étudiés. Vient maintenant la mise à disposition des cartes

générées. La communauté scientifique de la télédétection tend à générer des cartes globales, utilisant un système de classification unique. Ceci a l'avantage de permettre la comparaison avec une même référence pour plusieurs pays, lorsque cela est possible. En revanche, nous perdons les potentielles spécificités locales. Par exemple, comment définir la classe « zone urbaine » pour qu'elle englobe tous les cas ? Comment caractériser la végétation ? Les LCZ sont, à ma connaissance, le seul système dont le but est de produire des indicateurs indépendants de ces questions d'interprétation. Ces cartes globales sont ensuite mises à la disposition de tous. En revanche, la question se complique pour les modèles spécifiques à des pays. Lorsqu'un modèle ou une stratégie d'entraînement est sélectionnée, il faut pouvoir l'entraîner soi-même, sauf si un processus collaboratif est proposé, donc travailler conjointement entre informaticiens et démographes. Pour continuer ce projet sur la caractérisation du lien entre la population et l'environnement avec les images satellites, il faut envisager de mettre en place une équipe pluridisciplinaire impliquant des personnes spécialistes dans chacun des domaines mais ouverts aux autres.

D'un point de vue plus personnel, j'ai vraiment apprécié d'utiliser mes compétences en informatique et en télédétection pour servir la démographie. En plus de mes compétences techniques largement développées avec ce travail sur l'Afrique, j'ai appris à mieux intégrer le contexte local de la région étudiée, ce qui a permis d'améliorer à la fois la cartographie et de mettre en évidence des interactions intéressantes entre l'environnement et la population. Ce passage dans le domaine de la démographie a été plus qu'une simple découverte d'un nouveau domaine: il a également eu des retombées positives, parfois inattendues, sur mes compétences en télédétection et sur ma capacité à vulgariser des concepts complexes dans les deux domaines. En conclusion, je voudrais encourager les personnes issues de la partie informatique à s'engager dans des projets interdisciplinaires et à tirer pleinement parti des opportunités de découverte qu'ils offrent.

Chapitre 6

Annexes

Dans un premier temps, le tableau de correspondance entre fokontany officiels et groupés est donné dans l'annexe 6.1. Ensuite est présentée une piste qui a été explorée pendant cette thèse à l'occasion de l'encadrement d'un stage, pour la cartographie des LCZ. La dernière section liste les communications liées à ces travaux.

6.1 Tableau de correspondance des fokontany

Fokontany officiel	**Fokontany groupé**
Soarano Ambondrona Ambodifilao	Soarano Ambondrona
Andohatapenaka II	Andohatapenaka
Ankasina	Ankasina
Antohomadinika Avaratra Antani	Antohomadinika
Antohomadinika Antsalovana Faa	Antohomadinika
Antohomadinika Afovoany	Antohomadinika
Lalamby Sy Ny Manodidina	Lalamby Sy Ny Manodidina
Antohomadinika IIig Hangar	Antohomadinika
67ha Avaratra Atsinanana	67ha
67ha Avaratra Andrefana	67ha
Andohatapenaka IIi	Andohatapenaka IIi
Andohatapenaka I	Andohatapenaka
67ha Afovoany Andrefana	67ha
67ha Atsimo	67ha
Antohomadinika Atsimo	Antohomadinika
Antetezanafovoany II	Cite Ambodin'isotry
Cite Ambodin'isotry	Cite Ambodin'isotry
Avaratetezana Bekiraro	Ambalavao Bekiraro
Andranomanalina Isotry	Andranomanalina
Andranomanalina Afovoany	Andranomanalina
Andranomanalina I	Andranomanalina
Ambalavao Isotry	Ambalavao Bekiraro

Fiata	Fiata
Antanimalalaka Analakely	Fiata
Soarano Ambondrona Tsiazotafo	Soarano Ambondrona
Ampandrana Ankadivato	Faravohitra
Faravohitra Ambony	Faravohitra
Faravohitra Mandrosoa	Faravohitra
Ambatonakanga Ambohitsorohitra	Ambatonakanga Ambohitsorohitra
Amboasarikely Ambatomena	Fiata
Isoraka Ampatsakana	Isoraka Ampatsakana
Cite Ampefiloha	Cite Ampefiloha
Manarintsoa Isotry	Manarintsoa
Manarintsoa Afovoany	Manarintsoa
Manarintsoa Atsinanana	Manarintsoa
Manarintsoa Anatihazo	Manarintsoa
Anatihazo Isotry	Anatihazo Isotry
Antetezanafovoany I	Antetezanafovoany I
Andavamamba Anjezika I	Andavamamba 1
Andavamamba Anjezika II	Andavamamba 1
Andavamamba Anatihazo I	Andavamamba 1
Andavamamba Anatihazo II	Andavamamba 1
Amparibe Ambohidahy Mahamasina	Amparibe Ambohidahy Mahamasina
Ambatovinaky	Ambatovinaky
Faliarivo Ambanidia	Ambanidia
Antsahabe Ankorahotra Ankazoto	Antsahabe Ankorahotra Ankazoto
Antanimora Ampasanimalo	Antanimora Ampasanimalo
Tsiadana	Tsiadana
Ambolokandrina 5a	Ambolokandrina 5a
Ambohipo Tanana Ampahateza And	Ambohipo Tanana Ampahateza And
Andohan'i Mandroseza Ambohibat	Mandroseza
Ambatoroka	Ambatoroka
Ankazotokana Ambony	Ambohitsiroa Ambony
Ambohitsiroa Vn	Ambohitsiroa Ambony
Andafiavaratra Ambavahadimitaf	Andafiavaratra Ambavahadimitaf
Volosarika Ambanidia	Ambanidia
Ambohimiandra Fenomanana Antsa	Ambohimiandra Fenomanana Antsa
Mandroseza Afovoany Mandroseza	Mandroseza
Miandrarivo Ambanidia	Ambanidia
Manjakamiadana Antsahondra Amb	Manjakamiadana Antsahondra Amb
Ambohipotsy Ambohimitsimbina A	Ambohipotsy Ambohimitsimbina A
Andohamandry	Andohamandry
Manakambahiny Ankerakely Ambat	Manakambahiny Ankerakely Ambat
Morarano Andrangaranga Ambatol	Morarano Andrangaranga Ambatol
Ambohitsoa Ankazolava Antanamb	Ambohitsoa Ankazolava Antanamb
Androndrabe Ampamantanana Ambo	Androndrabe Ampamantanana Ambo
Mahazoarivo Ambohidraserika	Mahazoarivo Ambohidraserika
Androndrakely Saropody Antonta	Androndrakely Saropody Antonta

Antaninandro Ampandrana	Ampandrana 3
Ankazomanga Andraharo	Ankazomanga Andraharo
Ankorondrano Andranomahery	Ankorondrano
Ankorondrano Atsinanana Tsaramasay	Ankorondrano Tsaramasay
Ankorondrano Andrefana	Ankorondrano
Ambodivona Ankadifotsy	Ankadifotsy
Ankadifotsy Befelatanana	Ankadifotsy
Ambohibary Antanimena	Antanimena
Andravoahangy Tsena	Andravoahangy
Mandialaza Ambodivona	Mandialaza
Ankadifotsy Antanifotsy	Ankadifotsy
Ankaditapaka Avaratra	Behoririka
Mandialaza Ankadifotsy	Mandialaza
Antanimena	Antanimena
Mandialaza Ambatomitsangana	Mandialaza
Andravoahangy Andrefana	Andravoahangy
Mahavoky	Mahavoky
Behoririka Ankaditapaka	Behoririka
Ambatomitsangana	Ambatomitsangana
Andravoahangy Atsinanana	Andravoahangy
Behoririka Ambatomitsangana	Behoririka
Besarety	Besarety
Soavinandriana	Soavinandriana
Avaradoha	Avaradoha
Behoririka	Behoririka
Ampandrana Besarety	Ampandrana 3
Ampandrana Atsinanana	Ampandrana 3
Betongolo	Betongolo
Ambohitrakely	Ambohitrakely
Ampandrana Andrefana	Ampandrana 3
Ampahibe	Ampahibe
Ankadivato Ill	Ankadivato Ill
Antsakaviro Ambodirotra	Antsakaviro Ambodirotra
Andrefan'ambohijanahary IIIG-I	Andrefan'ambohijanahary
Andrefan'ambohijanahary IIIH-I	Andrefan'ambohijanahary
Mahamasina Atsimo	Mahamasina Atsimo
Tsimialonjafy	Tsimialonjafy
Ambanin'ampamarinana	Ambanin'ampamarinana
Ankadilalana	Ankadilalana
Tsarafaritra	Tsarafaritra
Fiadanana IIil	Fiadanana
Fiadanana IIin	Fiadanana
Mananjara	Mananjara
Andrefan'i Mananjara	Mananjara
Anosibe Ambohibarikely	Anosibe

Mandrangobato I	Mandrangobato
Mandrangobato II	Mandrangobato
Anosibe Andrefana I	Anosibe
Anosibe Andrefana II	Anosibe
Andavamamba Ambilanibe	Andavamamba Ambilanibe
Ivolaniray	Ivolaniray
Ambodirano Andrefana	Ambodirano
Anosibe Ambodirano	Anosibe Ambodirano
Ambohitrakely Andoharano	Andoharano
Ambohitrakely Atsimo	Ambohitrakely Atsimo
Behenjy Atsimo Andrefana	Behenjy Atsimo Andrefana

6.2 Sentinel-2 and OSM

Ces travaux se sont limités à l'utilisation de données satellites optiques pour la caractérisation de l'environnement. Il existent beaucoup d'autres sources permettant cette caractérisation, qu'elles soient d'autres types de données aéroportées (SAR, LiDAR et autres) ou qu'elles proviennent de sources "au sol". Ainsi, des données peuvent être générées à partir de capteurs au sol, comme pour la pollution, la température ou l'humidité, et nous apportent une information précieuse pour caractériser l'environnement des populations à l'étude. Les autres sources ne proviennent pas de capteurs mais de l'observation humaine. Des relevés peut être faits par des experts, ou des contributeurs, mais en général la zone couvertes par ces observations est faible si le nombre d'observateur est limité.

Le projet collaboratif OpenStreetMap (OSM) a pour objectif de créer une carte du monde, en libre d'accès et éditable par tous. Cette ouverture de la modification de la carte à tous permet cette couverture globale et des mises à jours rapides en cas de changement, sous réserve que des contributeurs soient sur place. Les contributeurs sont invités à donner la localisations et la caractérisation des éléments de la couverture terrestre (parc, routes, bâtiments, nombre d'étages, surface, ...). Dans les pays du Nord, les villes sont figées et l'urbanisation n'est pas rapide donc le besoin de mises à jour n'est pas très élevé. De plus, il n'y a pas (ou peu) de zones de bâti informel dans ces régions, qui sont difficiles à cartographier. En revanche, les villes d'Afrique sub-Saharienne évoluent dans un contexte différent. L'urbanisation est plus rapide et les zones informelles ne sont pas rares, et peu structurées comme nous l'avons vu dans les chapitres de ce manuscrit. A Antananarivo, les zones informelles et les zones formelles sont intriquées, alors qu'à Ouagadougou, ces zones sont dans la ville, mais sont séparées des zones formelles. Dans les deux cas, ces zones peuvent difficilement être cartographiées avec précision, et sont des fois totalement absentes dans les données OSM (Figure 6.1).

La grande fréquence des images satellites, par exemple Sentinel-2, pourraient permettre de combler ce manque d'information tout en conservant la précision des données OSM, qui servent de référence terrain qui a été faite par des contributeurs ou experts. Des liens entre les images satellites et OSM ont été explorés par la communauté scientifique, notamment pour la génération de cartes de couverture du sol [95] ou pour courriger les erreurs d'annotation [96]. L'idée de ces travaux est d'explorer la fusion de données Sentinel-2, qui ont une résolution moyenne qui ne permet pas toujours la cartographie des LCZ, et des données OSM incom-

(a) Données OSM (b) Image Google Maps (c) Image Sentinel-2

Figure 6.1: Données OSM (a), image Google Maps (b), et image Sentinel-2 (c) de Banfora au Burkina Faso. Certains bâtiments sont annotés, mais d'autres, éloignés du centre-ville, ne le sont pas.

plètes pour la cartographie des LCZ en Afrique sub-Saharienne. Ce sujet a fait l'objet de la proposition et l'encadrement d'un stage de trois mois effectué par Mathilde Bonin, étudiante en Master à l'Institut Polytechnique de Paris.

Plan des travaux

Le point de départ pour réaliser cette tâche est de prédire les LCZ en utilisant uniquement OSM dans un contexte connu, ici la France. Ces données permettent de calculer les caractéristiques géométriques physiques des LCZ, présentées dans le Tableau 2.1. Le modèle, basé sur des règles, permet d'avoir des cartes LCZ de référence pour une zone donnée. Dans ce stage, Dunkerque a été choisie comme la ville de référence pour sa diversité des morphologies dans les zones urbaines.

La seconde étape de ces travaux, qui n'a pas pu être réalisée pendant le stage est de fusionner les données OSM et Sentinel-2 pour la cartographie. L'idée est de se servir des données OSM, a priori complètes, et de les dégrader progressivement pour estimer l'apport des images Sentinel-2 dans ce contexte. Une fois le modèle proposé évalué, il pourra être transféré à des régions d'Afrique sub-Saharienne. Ce processus est illustré en Figure 6.2.

Discussion et défis

Ces travaux ont pour but d'utiliser le plus de données disponibles en Afrique sub-Saharienne pour la cartographie de l'environnement. Les données OSM sont fournis par des contributeurs sur place, ce qui nous apporte une connaissance terrain supplémentaire mais ces données sont incomplètes, en particulier dans les zones informelles. L'établissement d'un modèle basé sur des règles nécessite d'avoir toutes les informations nécessaires aux calculs des descripteurs géométriques des LCZ accessible via OSM. Cependant, même si tous les bâtiments et autres éléments sont bien indiqués, certains éléments manquent de descriptions, comme la présence d'arbres dans les parcs ou la hauteur des bâtiments, qui sont essentielles pour la classification des LCZ.

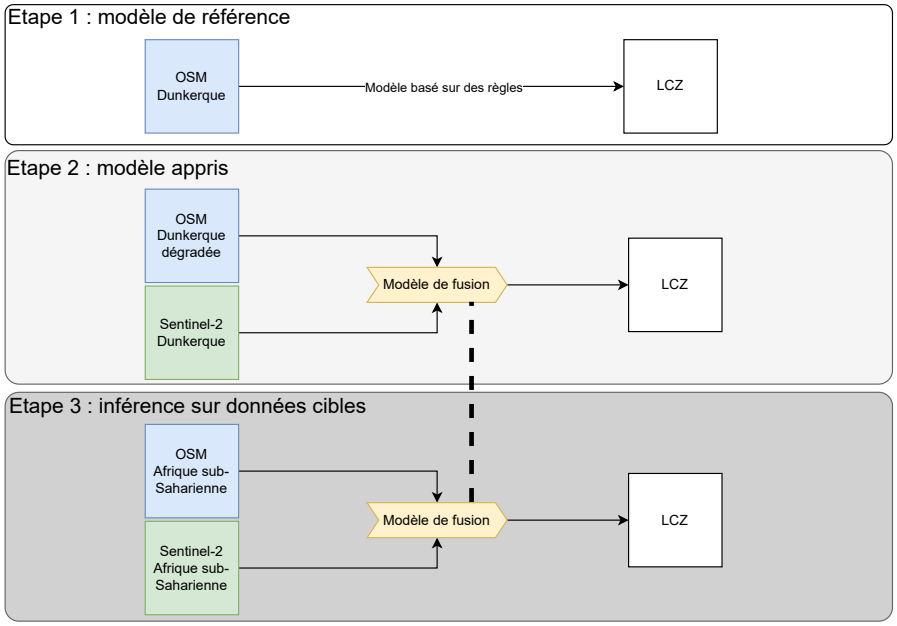


Figure 6.2: Processus global mis en place pour létude de la fusion entre OSM et Sentinel-2 pour la génération de cartes lCZ.

De plus l'organisation spatiale des villes d'Afrique sub-Saharienne est très différente des villes européennes. Dans la création de So2Sat [13], les ville sont d'ailleurs été séparées en région culturelles (version *Cultural 10*) pour évaluer les capacités de transfert des modèles. Dans le cadre d'un modèle appris avec les données OSM, le même phénomène d'écart entre les données d'entraînement (France) et de test (Afrique sub-Saharienne) pourrait apparaître, car les données OSM, bien que similaires, sont organisées différemment dans l'espace.

Enfin, l'a priori apporté par les données OSM peut être très différent de la réalité actuelle et biaiser le résultat final. Cette situation peut arriver fréquemment dans les pays en développement, avec une urbanisation rapide et des données OSM qui ne sont pas mises à jour. Dans ce cas, le modèle appris doit "prioriser" l'information extraite des images Sentinel-2, en prenant en compte que les données terrain sont trop anciennes.

6.3 Communications

Articles scientifiques

B. Rousse, S. Lobry, G. Duthé, *et al.*, "Seasonal semi-supervised domain adaptation for linking population studies and Local Climate Zones", in *2023 Joint Urban Remote Sensing Event (JURSE)*, 2023, pp. 1–4. DOI: 10.1109/JURSE57346.2023.10144163

B. Rousse, S. Lobry, G. Duthé, *et al.*, "Domain adaptation for mapping lczs in sub-saharan africa with remote sensing: A comprehensive approach to health data analysis", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 13 016–13 029,

2024. DOI: 10.1109/JSTARS.2024.3421284

Présentations lors de conférences

B. Rousse, L. Thay, S. Lobry, *et al.*, "Matching environmental data produced from remote sensing images to demographic data in Sub-Saharan Africa", in *ESA living Planet Symposium*, 2022

B. Rousse, L. Thay, S. Lobry, *et al.*, "Linking Malaria and Local Climate Zones: a geospatial analysis in Burkina Faso", in *Population Association of America Annual Meeting*, 2022

B. Rousse, S. Lobry, L. Wendling, *et al.*, "Linking population data to high resolution map: a case study in Burkina Faso", in *Machine Learning for Earth Observation @ ICLR*, 2023

B. Rousse, S. Lobry, G. Duthé, *et al.*, "Seasonal semi-supervised domain adaptation for linking population studies and Local Climate Zones", in *2023 Joint Urban Remote Sensing Event (JURSE)*, 2023, pp. 1–4. DOI: 10.1109/JURSE57346.2023.10144163

Bibliographie

[1] T. V. Ha, W. Kim, T. Nguyen-Tien, *et al.*, "Spatial distribution of Culex mosquito abundance and associated risk factors in Hanoi, Vietnam", *PLOS Neglected Tropical Diseases*, vol. 15, no. 6, 2021.

[2] R. Gibb, F. J. Colón-González, P. T. Lan, *et al.*, "Interactions between climate change, urban infrastructure and mobility are driving dengue emergence in Vietnam", *Nature Communications*, vol. 14, no. 1, p. 8179, 2023.

[3] O. Mudele, A. C. Frery, L. F. Zanandrez, A. E. Eiras, and P. Gamba, "Modeling dengue vector population with earth observation data and a generalized linear model", *Acta Tropica*, vol. 215, p. 105 809, 2021.

[4] A. Ayanlade, A. Oluwaranti, O. S. Ayanlade, *et al.*, "Extreme climate events in sub-saharan africa: A call for improving agricultural technology transfer to enhance adaptive capacity", *Climate Services*, vol. 27, p. 100 311, 2022.

[5] K. Grace and F. Davenport, "Climate variability and health in extremely vulnerable communities: Investigating variations in surface water conditions and food security in the West African Sahel", *Population and Environment*, vol. 42, no. 4, pp. 553–577, 2021.

[6] S. G. Wolde, P. D'Odorico, and M. C. Rulli, "Environmental drivers of human migration in sub-saharan africa", *Global Sustainability*, vol. 6, e9, 2023.

[7] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images", in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 210–223, ISBN: 978-3-642-15567-3.

[8] I. Harris, P. Jones, T. Osborn, and D. Lister, "Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset", *International Journal of Climatology*, vol. 34, no. 3, pp. 623–642, 2014.

[9] W. CIESIN, "Wildlife Conservation Society (WCS), and Center for International Earth Science Information Network (CIESIN)/Columbia University-Last of the Wild Project: Global Human Footprint Dataset (IGHP)", 2005.

[10] M. Pesaresi, D. Ehrlich, A. Florczyk, *et al.*, "GHS built-up grid, derived from landsat, multitemporal (1975, 1990, 2000, 2014)", *European Commission, Joint Research Centre (JRC)*, 2015.

[11] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12MS - a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion", vol. IV-2/W7, 2019, pp. 153–160.

[12] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding", in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5901–5904.

[13] X. X. Zhu, J. Hu, C. Qiu, *et al.*, "So2Sat LCZ42: A benchmark data set for the classification of global Local Climate Zones [software and data sets]", *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 3, pp. 76–89, 2020.

[14] S. Hafner, Y. Ban, and A. Nascetti, "Unsupervised domain adaptation for global urban extraction using sentinel-1 SAR and sentinel-2 MSI data", *Remote Sensing of Environment*, vol. 280, p. 113 192, 2022.

[15] X. Zhao, J. Hu, L. Mou, Z. Xiong, and X. X. Zhu, "Cross-city Landuse classification of remote sensing images via deep transfer learning", *International Journal of Applied Earth Observation and Geoinformation*, vol. 122, p. 103 358, 2023.

[16] D. Ienco and K. Ose, "Combine histogram matching and domain adaptation to cope with temporal transfer learning for the semantic segmentation of vhr images", in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 409–412.

[17] Y. Yang and S. Soatto, "FDA: Fourier Domain Adaptation for Semantic Segmentation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2020, pp. 4084–4094, ISBN: 978-1-72817-168-5.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative Adversarial Nets", in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27, Curran Associates, Inc., 2014.

[19] M. Mirza and S. Osindero, *Conditional generative adversarial nets*, 2014.

[20] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games", in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer International Publishing, 2016, pp. 102–118, ISBN: 978-3-319-46475-6.

[21] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3234–3243.

[22] M. Cordts, M. Omran, S. Ramos, *et al.*, "The cityscapes dataset for semantic urban scene understanding", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] J. Hoffman, E. Tzeng, T. Park, *et al.*, "CyCADA: Cycle-Consistent Adversarial Domain Adaptation", 2018.

[24] S. Ettedgui, S. Abu-Hussein, and R. Giryes, *ProCST: Boosting Semantic Segmentation Using Progressive Cyclic Style-Transfer*, 2022.

[25] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "ColorMapGAN: Unsupervised Domain Adaptation for Semantic Segmentation Using Color Mapping Generative Adversarial Networks", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7178–7193, 2020.

[26] N. Letheule, F. Weissgerber, S. Lobry, and E. Colin, "Automatic simulation of sar images: Comparing a deep-learning based method to a hybrid method", in *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 4958–4961.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1991.

[28] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A Kernel Method for the Two-Sample-Problem", in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19, MIT Press, 2006.

[29] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer Feature Learning with Joint Distribution Adaptation", in *2013 IEEE International Conference on Computer Vision*, IEEE, 2013, pp. 2200–2207, ISBN: 978-1-4799-2840-8.

[30] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks", in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, PMLR, 2015, pp. 97–105.

[31] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks", in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. ICML'17, JMLR.org, 2017, pp. 2208–2217.

[32] B. Sun and K. Saenko, *Deep CORAL: Correlation Alignment for Deep Domain Adaptation*, 2016.

[33] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation", in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16, Phoenix, Arizona: AAAI Press, 2016, 2058–2065.

[34] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, *Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning*, 2019.

[35] Y. Zhu, F. Zhuang, J. Wang, *et al.*, "Deep subdomain adaptation network for image classification", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1713–1722, 2021.

[36] S. Zhu, B. Du, L. Zhang, and X. Li, "Attention-based multiscale residual adaptation network for cross-scene classification", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[37] Y. Ganin, E. Ustinova, H. Ajakan, *et al.*, "Domain-Adversarial Training of Neural Networks", in *Domain Adaptation in Computer Vision Applications*, G. Csurka, Ed., Springer International Publishing, 2017, pp. 189–209, ISBN: 978-3-319-58346-4 978-3-319-58347-1.

[38] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun, "No More Discrimination: Cross City Adaptation of Road Scene Segmenters", in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, pp. 2011–2020, ISBN: 978-1-5386-1032-9.

[39] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation", in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3723–3732.

[40] Z. Chen, B. Yang, A. Ma, *et al.*, "Joint alignment of the distribution in input and feature space for cross-domain aerial image semantic segmentation", *International Journal of Applied Earth Observation and Geoinformation*, vol. 115, p. 103 107, 2022.

[41] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[42] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-Supervised Domain Adaptation via Minimax Entropy", in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, 2019, pp. 8049–8057, ISBN: 978-1-72814-803-8.

[43] X. Yue, Z. Zheng, S. Zhang, *et al.*, "Prototypical Cross-domain Self-supervised Learning for Few-shot Unsupervised Domain Adaptation", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2021, pp. 13 829–13 839, ISBN: 978-1-66544-509-2.

[44] J. Huang, D. Guan, A. Xiao, S. Lu, and L. Shao, "Category Contrast for Unsupervised Domain Adaptation in Visual Tasks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2022, pp. 1193–1204, ISBN: 978-1-66546-946-3.

[45] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–18, 2023.

[46] I. D. Stewart and T. R. Oke, "Local Climate Zones for urban temperature studies", *Bulletin of the American Meteorological Society*, vol. 93, pp. 1879–1900, 2012.

[47] P. J. Alexander, G. Mills, and R. Fealy, "Using LCZ data to run an urban energy balance model", *Urban Climate*, vol. 13, pp. 14–37, 2015.

[48] J. Geletič, M. Lehnert, P. Dobrovolnỳ, and M. Žuvela-Aloise, "Spatial modelling of summer climate indices based on Local Climate Zones: Expected changes in the future climate of brno, czech republic", *Climatic Change*, vol. 152, no. 3-4, pp. 487–502, 2019.

[49] H. Wouters, M. Demuzere, U. Blahak, *et al.*, "The efficient urban canopy dependency parametrization (sury) v1. 0 for atmospheric modelling: Description and application with the cosmo-clm model for a belgian summer", *Geoscientific Model Development*, vol. 9, no. 9, pp. 3027–3054, 2016.

[50] A. G. Davenport, S. B. Grimmond, T. R. Oke, and J. Wieringa, "Estimating the roughness of cities and sheltered country", in *12th Conference on Applied Climatology*, Asheville, NC: Amer. Meteor. Soc., 2000, pp. 96–99.

[51] I. D. Stewart, T. R. Oke, and E. S. Krayenhoff, "Evaluation of the 'local climate zone' scheme using temperature observations and model simulations", *International Journal of Climatology*, vol. 34, no. 4, pp. 1062–1080, 2014.

[52] M. Lehnert, J. Geletič, J. Husák, and M. Vysoudil, "Urban field classification by "local climate zones" in a medium-sized Central European city: The case of Olomouc (Czech Republic)", *Theoretical and Applied Climatology*, vol. 122, no. 3, pp. 531–541, 2015.

[53] L. Liu, Y. Lin, Y. Xiao, *et al.*, "Quantitative effects of urban spatial characteristics on outdoor thermal comfort based on the lcz scheme", *Building and Environment*, vol. 143, pp. 443–460, 2018.

[54] B. Bechtel, P. J. Alexander, J. Böhner, *et al.*, "Mapping Local Climate Zones for a worldwide database of the form and function of cities", *ISPRS International Journal of Geo-Information*, vol. 4, no. 1, pp. 199–219, 2015.

[55] B. Bechtel, P. J. Alexander, C. Beck, *et al.*, "Generating WUDAPT Level 0 data – Current status of production and evaluation", *Urban Climate*, vol. 27, pp. 24–45, 2019.

[56] M. Demuzere, J. Kittner, and B. Bechtel, "Lcz generator: A web application to create local climate zone maps", *Frontiers in Environmental Science*, vol. 9, 2021.

[57] M. Demuzere, S. Hankey, G. Mills, W. Zhang, T. Lu, and B. Bechtel, "Combining expert and crowd-sourced training data to map urban form and functions for the continental us", *Scientific data*, vol. 7, no. Article 264, pp. 264–1 –264–13, 2020.

[58] M. Demuzere, B. Bechtel, A. Middel, and G. Mills, "Mapping Europe into local climate zones", *PLOS ONE*, vol. 14, no. 4, pp. 1–27, 2019.

[59] M. Demuzere, J. Kittner, A. Martilli, *et al.*, "A global map of local climate zones to support earth system modelling and urban-scale environmental science", *Earth System Science Data*, vol. 14, no. 8, pp. 3835–3873, 2022.

[60] J. Rosentreter, R. Hagensieker, and B. Waske, "Towards large-scale mapping of local climate zones using multitemporal sentinel 2 data and convolutional neural networks", *Remote Sensing of Environment*, vol. 237, p. 111 472, 2020.

[61] C. Qiu, M. Schmitt, L. Mou, P. Ghamisi, and X. X. Zhu, "Feature Importance Analysis for Local Climate Zone Classification Using a Residual Convolutional Neural Network with Multi-Source Datasets", *Remote Sensing*, vol. 10, no. 10, p. 1572, 2018.

[62] C. Qiu, L. Mou, M. Schmitt, and X. X. Zhu, "Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network", *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 154, pp. 151–162, 2019.

[63] X. X. Zhu, C. Qiu, J. Hu, *et al.*, "The urban morphology on our planet – global perspectives from space", *Remote Sensing of Environment*, vol. 269, p. 112 794, 2022.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[65] B. C. Thiede and C. Gray, "Climate exposures and child undernutrition: Evidence from Indonesia", *Social Science & Medicine*, vol. 265, p. 113 298, 2020. (visited on 01/27/2022).

[66] K. Nicholas, L. Campbell, E. Paul, G. Skeltis, W. Wang, and C. Gray, "Climate anomalies and childhood growth in Peru", *Population and Environment*, vol. 43, no. 1, pp. 39–60, 2021.

[67] S. J. Macfarlan, R. Schacht, I. Bourland, *et al.*, "NDVI predicts birth seasonality in historical Baja California Sur, Mexico: Adaptive responses to arid ecosystems and the North American Monsoon", *Biodemography and Social Biology*, vol. 66, no. 2, pp. 145–155, 2021. (visited on 01/27/2022).

[68] A. Sikarwar, R. Rani, G. Duthé, and V. Golaz, "Association of greenness with COVID-19 deaths in India: An ecological study at district level", *Environmental Research*, vol. 217, p. 114 906, 2023.

[69] A. L. Buczak, P. T. Koshute, S. M. Babin, B. H. Feighner, and S. H. Lewis, "A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data", *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 124, 2012.

[70] D. Zanaga, R. Van De Kerchove, W. De Keersmaecker, *et al.*, *Esa worldcover 10 m 2020 v100*, version v100, Oct. 2021.

[71] D. Zanaga, R. Van De Kerchove, D. Daems, *et al.*, *Esa worldcover 10 m 2021 v200*, version v200, Oct. 2022.

[72] O. Brousse, S. Georganos, M. Demuzere, *et al.*, "Can we use local climate zones for predicting malaria prevalence across sub-saharan african cities?", *Environmental Research Letters*, 2020.

[73] T.-H. K. Chen, H. T. Horsdal, K. Samuelsson, *et al.*, "Higher depression risks in medium- than in high-density urban form across denmark", *Science Advances*, vol. 9, no. 21, eadf3760, 2023.

[74] P. I. D. Lin, M. Qi, S. Hankey, *et al.*, "Associations of local climate zones with cardiovascular disease: Findings from the us-based nationwide nurses' health study from 2000 to 2016", *ISEE Conference Abstracts*, vol. 2023, no. 1, pp. 1–1, 2023.

[75] B. Masquelier, D. Waltisperger, O. Ralijaona, G. Pison, and A. Ravélo, "The epidemiological transition in Antananarivo, Madagascar: An assessment based on death registers (1900–2012)", *Global Health Action*, vol. 7, no. 1, p. 23 237, 2014. (visited on 08/22/2024).

[76] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[77] W. Sirko, S. Kashubin, M. Ritter, *et al.*, *Continental-Scale Building Detection from High Resolution Satellite Imagery*, 2021.

[78] L. van der Maaten and G. Hinton, "Viualizing data using t-sne", *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[79] J. Martin and C. G. Camarda, "Modelling age-space mortality dynamics in small areas", *38th International Workshop on Statistical Modelling*, pp. 190–193, 2024.

[80] Health Ministry of Burkina Faso, "Plan stratégique national de lutte contre le paludisme du burkina faso 2016–2020", Tech. Rep., 2016.

[81] C. Burgert, J. Colston, T. Roy, and B. Zachary, "Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys", 2013.

[82] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations", in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 1597–1607.

[83] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[84] O. Mañas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodríguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data", in *2021 IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9394–9403.

[85] D. Kingma and J. Ba, "Adam: A method for stochastic optimization", in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[86] B. Rousse, S. Lobry, G. Duthé, V. Golaz, and L. Wendling, "Seasonal semi-supervised domain adaptation for linking population studies and Local Climate Zones", in *2023 Joint Urban Remote Sensing Event (JURSE)*, 2023, pp. 1–4.

[87] C. Perez-Heydrich, J. Warren, C. Burgert, and M. Emch, "Influence of demographic and health survey point displacements on raster-based analyses", *Spatial Demography*, vol. 4, pp. 1–19, 2015.

[88] K. Grace, N. N. Nagle, C. R. Burgert-Brucker, *et al.*, "Integrating environmental context into dhs analysis while protecting participant confidentiality: A new remote sensing method", *Population and Development Review*, vol. 45, p. 1, 2019.

[89] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm", *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[90] INSD, "Enquête sur les indicateurs du paludisme au burkina faso", 2018.

[91] L. S. Tusting, J. Rek, E. Arinaitwe, *et al.*, "Why is malaria associated with poverty? Findings from a cohort study in rural Uganda", *Infectious Diseases of Poverty*, vol. 5, no. 1, p. 78, 2016.

[92] A. Degarege, K. Fennie, D. Degarege, S. Chennupati, and P. Madhivanan, "Improving socioeconomic status may reduce the burden of malaria in sub saharan africa: A systematic review and meta-analysis", *PLOS ONE*, vol. 14, no. 1, pp. 1–26, 2019.

[93] P. Wessel and W. H. F. Smith, "A global self-consistent, hierarchical, high-resolution shoreline database", *Journal of Geophysical Research*, vol. 101, no. C6, pp. 8741–8743, 1996.

[94] L. Bruzzone and D. Prieto, "Automatic analysis of the difference image for unsupervised change detection", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

[95] N. Audebert, B. Le Saux, and S. Lefevre, "Joint Learning from Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps", in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI: IEEE, 2017, pp. 1552–1560.

[96] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, "Openstreetmap: Challenges and opportunities in machine learning and remote sensing", *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 184–199, 2020.

[97] B. Rousse, S. Lobry, G. Duthé, V. Golaz, and L. Wendling, "Domain adaptation for mapping lczs in sub-saharan africa with remote sensing: A comprehensive approach to health data analysis", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 13 016–13 029, 2024.

[98] B. Rousse, L. Thay, S. Lobry, G. Duthé, L. Wendling, and V. Golaz, "Matching environmental data produced from remote sensing images to demographic data in Sub-Saharan Africa", in *ESA living Planet Symposium*, 2022.

[99] B. Rousse, L. Thay, S. Lobry, G. Duthé, L. Wendling, and V. Golaz, "Linking Malaria and Local Climate Zones: a geospatial analysis in Burkina Faso", in *Population Association of America Annual Meeting*, 2022.

[100] B. Rousse, S. Lobry, L. Wendling, G. Duthe, and V. Golaz, "Linking population data to high resolution map: a case study in Burkina Faso", in *Machine Learning for Earth Observation @ ICLR*, 2023.