Combien d'habitants ? De la détection du bâti à l'estimation de la population pour différentes régions d'Afrique

Léo Lipovac

amU
Aix Marseille Université

NNT : 0000AIXM0000

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 20 juin 2025 par

Léo Lipovac

Combien d'habitants ?

De la détection du bâti à l'estimation de la population pour différentes régions d'Afrique

Discipline
Démographie

École doctorale
ED 355 – Espaces, Cultures, Sociétés

Laboratoire/Partenaires de recherche
Diginove
Laboratoire Population Environnement Développement (LPED)
Institut National d'Etudes Démographiques (INED)

Composition du jury

Didier Breton	Rapporteur / Président du jury
Professeur des universités, Université de Strasbourg	
Catherine Linard	Rapportrice
Professeure, Université de Namur	
Hervé Bassinga	Examinateur
Maître-Assistant, ISSP	
Valérie Golaz	Directrice de thèse
Directrice de recherche, INED, LPED	
Nicolas Pech	Co-directeur de thèse
Maître de Conférences, AMU	
Ali Ahmad	Membre invité
Directeur technique (PhD), Diginove	
Géraldine Duthé	Membre invitée
Directrice de recherche, INED	
Laurence Reboul	Membre invitée
Maîtresse de Conférences HDR, AMU	

Diginove
ined INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES
LPED Laboratoire Population Environnement Développement www.lped.fr

Affidavit

Je soussigné, Léo Lipovac, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Valérie Golaz, Nicolas Pech et Ali Ahmad, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.
Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Aubervilliers, le 4 avril 2025

Affidavit

I, undersigned, Léo Lipovac, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Valérie Golaz, Nicolas Pech et Ali Ahmad, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.
This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Aubervilliers, 4 avril 2025

Liste de publications et participations aux conférences

1) Liste des publications et/ou brevet [1] réalisées dans le cadre du projet de thèse :

Article de revue

Lipovac L, Golaz V, Pech N, « Localiser les nouvelles communes malgaches en l'absence de cartographie officielle : une méthode pour la mise en cartes des décomptes de population du RGPH-3 à Madagascar », 32 pages (en cours de soumission).

Chapitre d'ouvrage

Artadji A., Lipovac L., Andriamanantena H. N. et Rousse B., 2024. Can we estimate sub-Saharan Africa's population from remote sensing images and land cover mapping? In : Cardi J., Favrot M., Gastineau B., Genin D., Golaz V. et Robles C. (éds.), Digressions. Marseille, France : Laboratoire Population Environnement Développement (LPED). Les Impromptus du LPED, n° 8. p. 186-194. Disponible à l'adresse : https://lped.fr/?LesImpromptusDuLped8Digressions.

Mémoire/Travaux académiques

Lipovac L., 2021, Détection du bâti et modèles d'estimation démographiques de la population des zones géographiques africaines, une analyse de la région d'Abuja, capitale du Nigéria, Rapport d'alternance de 2ème année du master MASS, Aix-Marseille Université – Diginove, 55 pages.

Lipovac L., 2020, Estimation de la population par images satellites, Rapport de stage de Master 1 de MASS, Aix-Marseille Université – Diginove, 42 pages.

[1] Cette liste comprend les articles publiés, les articles soumis à publication et les articles en préparation ainsi que les livres, chapitres de livre et/ou toutes formes de valorisation des résultats des travaux propres à la discipline du projet de thèse. La référence aux publications doit suivre les règles standards de bibliographie et doit être conforme à la charte des publications d'AMU.

2) Participation aux conférences[2] et écoles d'été au cours de la période de thèse :

Communications dans un colloque ou un séminaire scientifique

Lipovac L., 2024. Localiser les nouvelles communes malgaches en l'absence de cartographie officielle : une méthode pour la mise en carte des décomptes de population du RGPH-3 à Madagascar. Présenté à : séminaire jeunes chercheurs, Ined, 04 décembre 2024.

Lipovac L., 2023. Disaggregating Census Population Using Satellite Imagery: A Case Study of Itasy Region of Madagascar . Présenté à : Environmental and Climate Mobilities Network (ECMN), session "new datasets and modelling approaches", Conference - 2023. Vienne, Autriche, 10–12 juillet 2023. *http://hdl.handle.net/20.500.12204/15zsi4wB-5e4nGnw15vZ*

Lipovac L., 2023. Désagrégation des décomptes de population de recensement : une étude de cas de la région d'Itasy à Madagascar. Présenté à : Séminaire Journée doctorale 2023, Ined, 30 mai 2023 *http://hdl.handle.net/20.500.12204/nE_ki4wBU9Wft_qNVGx8*

Lipovac L., 2023. Caractéristiques environnementales à prendre en compte et leur rôle dans l'estimation de la population des zones bâties. Présenté à : Journée d'étude « Populations et environnement ». Ined, Aubervilliers, France, 9 février 2023. *http://hdl.handle.net/20.500.12204/kU_Ji4wBU9Wft_qNMGxQ*

Lipovac L., 2022. Disaggregating census population using satellite imagery : A case study of Itasy and Androy regions of Madagascar. Présenté à : Journée Demosud-LPED, Marseille, 8 décembre 2022

Lipovac L., 2022. Des images satellites aux estimations de population. Présenté à : Séminaire 6ème Journées scientifiques de l'École Doctorale 355 « Circulations des savoirs et interdisciplinarités en SHS », MMSH, 16 juin 2022

Lipovac L., 2022. Des images satellites aux estimations de population. Questions d'échelles et méthodes d'analyse spatiale. Présenté à : Séminaire Des outils d'analyse spatiale au service de la recherche en démographie, Ined, 11 mai 2022.

[2] Le terme « conférence » est générique. Il désigne à la fois « conférence », « congrès », « workshop », « colloques », « rencontres nationales et/ou internationales » … etc.
Indiquer si vous avez fait une présentation orale ou sous forme de poster.

Lipovac L., 2021. How TeleCense helps Companies, Authorities, International Organizations and Labs to assess and anticipate population growth and migration in emerging countries. Présenté à : Séminaire Des images satellite aux estimations de population, Ined, 20 mai 2021.

Communications diverses (médias et autres)

Golaz Valérie., Lipovac Léo., Mandamule Uacitissa, 2022. Quand les humains fuient le climat. Science En Direct 2022, Muséum National d'Histoire Naturelle, 8 octobre 2022. https://www.ined.fr/en/everything_about_population/videos/consequences-of-climate-and-environmental-changes-on-the-movement-of-populations/

Poster dans un colloque

Lipovac L., 2023. Disaggregating Census Population Using Satellite Imagery: A Case Study of the Itasy Region of Madagascar. Présenté à : PAA 2023 Annual Meeting. New Orleans, Louisiana, États-Unis, 12–15 avril 2023. *http://hdl.handle.net/20.500.12204/yZzRi4wB-5e4nGnw85th*

Ecole d'été

École d'été Quantilille – méthodes quantitatives en sciences sociales – 20-24 juin 2022.

Résumé

L'objet de cette thèse est de développer et d'évaluer des modèles démographiques pour estimer la population à partir des caractéristiques du bâti. Le plus souvent, les estimations de population s'appuient sur un carroyage, découpant l'espace en cellules de grille homogènes, comme celles produites et popularisées par le projet WorldPop. Nous proposons ici une méthode innovante, utilisant des zones bâties identifiées par le projet TeleCense et détectées à partir d'images des satellites Sentinel, comme unité principale d'analyse. L'objectif est d'obtenir une estimation fine de la population, tout en tenant compte de la diversité des zones bâties et d'évaluer la fiabilité des prédictions démographiques obtenues.

Pour atteindre ces objectifs, deux approches sont explorées. La première consiste en une méthode descendante, qui répartit la population des zones administratives dans des unités plus fines, en s'appuyant sur des données complémentaires pour affiner les estimations. Cette méthode est d'abord développée et testée à Madagascar, un pays choisi pour sa diversité agro-climatique et la qualité de son dernier recensement (RGPH-3, 2018). Son application à six régions permet de valider les estimations et d'en évaluer la pertinence. Elle est ensuite mise en œuvre au Bénin, ce qui révèle certaines limites des méthodes descendantes.

La partie suivante de la thèse se concentre sur la création d'une méthode ascendante novatrice, permettant d'estimer la population de zones géographiques sans recensement récent et totalement indépendante de données de recensement. Cette méthode ascendante repose uniquement sur des données issues de la télédétection et des informations socio-démographiques facilement disponibles, afin de rendre le modèle applicable à d'autres régions. Cette nouvelle approche est appliquée à Mayotte, grâce à des données démographiques précises, puis dans un contexte de fortes densités urbaines, dans la sous-préfecture d'Abidjan, en Côte d'Ivoire, posant un véritable défi en termes de modélisation.

Enfin, le dernier chapitre de la thèse est consacré au développement d'un modèle d'estimation exploitant les données très récentes « Open Buildings 2.5D » de Google Research, qui offrent une représentation détaillée et à haute résolution des surfaces bâties, incluant une estimation de la hauteur des bâtiments. En reprenant l'approche globale développée dans cette thèse, l'objectif est de mesurer dans quelle mesure une identification plus fine du bâti – quasiment bâtiment par bâtiment – permet d'améliorer la qualité des modèles démographiques.

Mots clés : Afrique, population, télédétection, bâti, recensement.

Abstract

The aim of this thesis is to develop and evaluate demographic models for estimating population on the basis of building characteristics. Most often, population estimates are gridded, dividing space into homogeneous cells, such as those produced and popularised by the WorldPop project. Here we propose an innovative method, using built-up areas identified by the TeleCense project and detected from Sentinel satellite images as the main unit of analysis. The aim is to obtain accurate population estimates, while taking into account the diversity of built-up areas, and to assess the reliability of the demographic predictions obtained.

To achieve these objectives, two approaches are explored. The first is a top-down method, which divides the population of administrative zones into finer units, using complementary data to refine the estimates. This method was first developed and tested in Madagascar, a country chosen for its agro-climatic diversity and the quality of its most recent census (RGPH-3, 2018). By applying it to six regions, we were able to validate the estimates and assess their relevance. It is then implemented in Benin, revealing some of the limitations of top-down methods.

The next part of the thesis focuses on the creation of an innovative bottom-up method, completely independent of census data, for estimating the population of administrative areas in the absence of a recent census. This bottom-up method relies solely on data from remote sensing and readily available socio-demographic information, in order to make the model applicable to other regions. This new approach is being applied to Mayotte, using precise demographic data, and then in a context of high urban densities, in the sub-prefecture of Abidjan, in Côte d'Ivoire, being a real challenge in terms of modelling.

Finally, the last chapter of the thesis is devoted to the development of an estimation model using the very recent ‘Open Buildings 2.5D’ data from Google Research, which provides a detailed, high-resolution representation of built-up areas, including an estimate of building height. Using the global approach developed in this thesis, the aim is to measure the extent to which finer identification of the built environment - virtually building by building - can improve the quality of demographic models.

Keywords : Africa, population, remote sensing, buildings, census.

Remerciements

Mes premières pensées vont vers mes encadrants de thèse, Valérie Golaz et Nicolas Pech. Merci pour votre soutien et votre encadrement depuis le début du master et jusqu'à la fin de la thèse. Nicolas, j'ai adoré faire mes premiers pas en tant qu'enseignant co-encadrant à tes côtés. Cela avait quelque chose de vraiment symbolique alors que j'étais à la place des étudiants peu de temps auparavant. Valérie, tu m'as vraiment encouragé et soutenu lors de moments plus difficiles et tu n'a cessé de m'encourager le reste du temps, sans oublier ta bienveillance et tes invitations sincères à discuter d'autres sujets que ceux liées à la thèse. Votre direction et vos avis scientifiques et humains ont toujours été très pertinents, et il y aurait tellement à dire. Je ne pense pas que je serais où j'en suis aujourd'hui sans votre soutien et je n'aurai pu espérer meilleur encadrement et suivi que le votre ces dernières années. Pour tout cela, merci à vous deux !

Merci à l'entreprise Diginove qui a financé ce projet de thèse. Mes premiers mots vont vers Michel, dirigeant de Diginove mais surtout premier penseur du programme TeleCense. Nous avons pensé ce projet de thèse ensemble et tu m'as poussé à le débuter et à le continuer, merci pour tes conseils et ta bienveillance. Merci tout d'abord à Steven, puis à Ali pour leur suivi institutionnel qui a permis d'assurer un bon déroulement du doctorat. Puis, merci à toute l'équipe de Diginove en général, vous êtes une équipe vraiment accueillante et j'ai toujours beaucoup aimé mes moments passés avec vous lorsque je venais à Aix-en-Provence. Un mot tout particulier à Maximilien qui en plus d'avoir été un super camarade, m'a fortement soutenu durant la deuxième année de thèse et qui a fortement contribué à l'avancement de cette dernière.

Merci à Didier Breton, Catherine Linard, Géraldine Duthé, Laurence Reboul et Hervé Bassinga d'avoir accepté de composer mon jury et de lire ce travail. J'ai hâte de discuter de ce travail avec vous le 20 juin prochain.

Mon parcours de démographe a débuté en septembre 2019 avec mon entrée dans le master mass pop à Aix-Marseille-Université. Je tiens à remercier toute l'équipe enseignante, vraiment géniale, qui m'a donné des bases solides à travers les différents apprentissages. Je me rappelle surtout d'un réel soutien de la part des professeurs et une envie sincère de pousser les étudiants à poursuivre ce qu'ils apprécient faire. Merci également à tous mes camarades, avec un petit mot particulier pour Nono et Nico pour leur soutien et la prise de nouvelles ces dernières années. Et je l'écris ici Nico comme ça c'est promis : je finirais par venir te voir à la Réunion !

D'un point de vue institutionnel, je suis inscrit à l'ED 355 et affilié au LPED où j'ai toujours apprécié les moments passés lors de mes passages. Puis j'ai surtout passé la plus grande partie de mon temps au sein de l'Ined. Je tiens tout d'abord à remercier mon unité d'accueil – l'unité 15, démographie des pays du sud –, pour votre

bienveillance et pour vos efforts sincères d'inclusion de chacun et chacune, notamment les doctorants à toutes les activités, qu'elles soient professionnelles et académiques mais aussi plus ludiques. J'ai été ravi de faire part de cette unité ! A ce propos je veux remercier ma team « remote sensing », Basile et Ankit, c'était tellement super déjà de vous avoir comme camarades, mais aussi comme collègues travaillant sur les mêmes thématiques, c'était un réel soutien !

Ces trois dernières années, j'ai eu la chance d'enseigner dans le master sociologie d'enquêtes à Paris-Cité, merci à Éric Dagiral pour ton accueil et ta confiance. Un grand merci aussi à Edith Darin de l'équipe de WorldPop pour tous les conseils de pré-thèse et de début de thèse qui m'ont beaucoup aidé.

Durant la thèse, j'ai travaillé au sein du bureau partagé 2.057 : merci Constance pour le super accueil, la découverte de l'INED et les rencontres ! Nancy, c'était super de partager de nombreuses discussions avec toi. Paul, Inès, je pense que vous savez à quel point vous avez été un grand soutien : je suis tellement heureux d'avoir pu partager une grande partie de ma thèse avec vous deux, merci ! Et Paul, on en aura passé du temps ensemble eheh, je me sens chanceux du temps partagé et surtout, n'oublie pas que c'est super ce que tu fais ! Mathis, merci de ta bonne humeur communicative mais aussi de ton sérieux, c'était super de partager cette dernière année et demie avec toi.

J'ai croisé et rencontré tellement de gens à l'Ined avec qui j'ai apprécié passé des moments, que ce soit au restaurant ou autour d'un café : toutes nos conversations m'ont beaucoup appris, vous êtes toutes et tous tellement intéressants et drôles. Je suis reconnaissant d'avoir pu partager ces supers moments, qui ont conduits à une belle bande d'amis et d'amies, j'en suis très heureux. Merci à vous d'avoir de ce doctorat un moment riche, joyeux et épanouissant.

Plus généralement, j'ai vraiment adoré mon passage à l'Ined en tant que doctorant. Les conditions d'accueil étaient pour moi super, un grand merci à tout le personnel et notamment à celui du restaurant : on s'est vraiment régalé tous les midi grâce à vous ! Merci aussi à Manu pour sa bonne humeur et sympathie !

Jujustar, merci à toi pour ton soutien, ta gentillesse et ta présence en général ces dernières années. Tu as rendu ces années de thèse beaucoup plus douces lorsqu'elles étaient plus compliquées, puis surtout tu m'as fait rigoler – tout le temps – et tu m'as soutenu jusqu'au bout. Tu sais que j'en serais pas ici sans toi non plus, juste la meilleure et très heureux de continuer mon après thèse avec toi !

Merci aussi à ma famille pour la présence et le soutien, je suis très heureux de pouvoir enfin partager mon travail avec vous lors de la soutenance. Nina et Victor, j'suis très fier de vous et de ce que vous faites. Merci aussi à mes colocataires, Mathilde et Ronan, qui ont rendu la vie joyeuse et drôle ! Et enfin, merci à mes copaines d'avoir été là, vous vous reconnaitrez et vous êtes super importants pour moi.

Table des matières

Affidavit 3
Affidavit 4
Liste de publications et participations aux conférences 5
Résumé 8
Abstract 9
Remerciements 10
Table des matières 13
Préambule 18
Introduction générale 19
Les recensements en Afrique : une répartition homogène sur le territoire et dans le temps ? 20
Estimer la population à partir d'images satellites, une possibilité récente 22
Questions de recherche 25
L'identification du bâti à partir d'images satellites 26
Les données démographiques 32
Positionnement de la thèse 33
Démarche et structure de la thèse 34
PARTIE 1 : Etat de l'art et approche méthodologique pour l'estimation de population à partir d'images satellites 39
Chapitre 1 – De l'estimation démographique aux projets de cartographie : un état des lieux 41
1. Méthodes descendantes 41
2. Méthodes ascendantes 46
3. Révolution cartographique : les projets de mise à disposition de données de population géoréférencées 50
4. Fournisseurs de données sur le bâti : diversité et spécificités 52
5. Africapolis et les agglomérations : cartographier l'urbanisation africaine 55
6. Conclusion 57
Chapitre 2 – Présentation du programme TeleCense et enjeux méthodologiques 59

1. Identification du bâti : première étape du programme TeleCense 59
2. Caractérisation du bâti et caractéristiques propre aux shapes 64
2.1 Variables liées aux shapes 64
2.2 Variables liées aux Zones Administratives 72
3. Méthodologie permettant l'estimation de population à l'intérieur du bâti 74
3.1 Approche descendante – La désagrégation 74
3.2 Approche ascendante 81
4. Conclusion 81
PARTIE 2 : Estimation de population par méthode descendante 83
Chapitre 3 – Désagrégation de la population à Madagascar 85
1. Préparation des zones sources : les communes malgaches 86
1.1 Les six régions du cas d'étude 86
1.2 Des changements administratifs fréquents et nombreux à Madagascar 88
1.3 Les sources disponibles et mobilisées 89
1.4 Méthodologie permettant la jointure des deux informations 92
1.5 Mise en place de la méthode et exemples de présentation dans les régions d'Itasy et de Diana 94
1.6 Méthode supplémentaire pour les communes difficiles à localiser 97
1.7 Récapitulatif et ultime vérification 102
1.8 Création des variables au niveau des communes 105
2. Cartographie des zones bâties (les zones cibles) 106
2.1 Gestion des shapes appartenant à plusieurs zones administratives 107
3. Etude descriptive des deux bases de données utilisées 109
3.1 La population des communes 109
3.2 Densité d'occupation du sol par le bâti 110
3.3 Récapitulatif et distribution des variables pour les deux couches de données 112
4. Calcul des pondérations Ws par modélisation linéaire et forêt aléatoire 115
4.1 Modélisation linéaire des communes de Madagascar 115
4.2 Forêts aléatoires pour la modélisation des communes de Madagascar 118
4.3 Comparaison des performances des deux modèles : 119

5. Mise en place des différentes méthodes de désagrégation..... 120
5.1 Evaluation des méthodes de désagrégation..... 120
5.2 Comparaison des résultats de désagrégation avec les estimations de WorldPop 125
5.3 Distribution de la population au niveau des shapes 129
5.4 Vérifier les résultats au niveau des shapes : un défi inatteignable à relever 130
6. Estimer la population après 2020 : quelles solutions pour Madagascar et au-delà ?..... 131
7. Conclusion 135
Chapitre 4 – Cas d'application sur plusieurs années de la désagrégation de la population du Bénin..... 137
1. Préparation des zones sources : les arrondissements du Bénin..... 140
1.1 Le recensement de la population du Bénin 142
1.2 Le tracé des limites administratives du Bénin 143
2. Identification des zones bâties des années 2017 à 2023 144
2.1 Un problème d'identification des shapes au fil des années ?..... 145
3. Projection des décomptes de population des arrondissements du Bénin..... 148
4. Répartition de la population à partir de la méthode d'interpolation surfacique entre 2017 et 2023 dans les shapes 152
5. Proposition de projections démographiques en prenant en compte l'évolution de la surface bâtie 154
6. Conclusion 157
PARTIE 3 : Estimation de population par méthode ascendante 159
Chapitre 5 – Développement du modèle ascendant à partir des données de bâti TeleCense et évaluation à Mayotte puis à Abidjan..... 161
1. Création du modèle ascendant à Madagascar 162
1.1 Modèle ascendant par modélisation linéaire..... 163
1.2 Modèle ascendant par forêt aléatoire 166
1.3 Comparaison et discussion des deux modèles ascendants..... 167
2. Application du modèle ascendant dans le département de Mayotte 168
2.1 La population et les zones administratives de Mayotte 172

2.2 La cartographie des zones bâties en shapes 175
2.3 Méthodes de validation employées 175
2.4 Résultats de l'approche ascendante à Mayotte 176
3. Application du modèle ascendant dans la sous-préfecture d'Abidjan 185
3.1 La population et les zones administratives d'Abidjan 186
3.2 Cartographie des zones bâties d'Abidjan 189
3.3 Résultats de l'approche ascendante à Abidjan 190
3.4 Un modèle ascendant pas encore universel 191
4. Conclusion 195
Chapitre 6 – Estimation de population à partir des nouvelles données de bâti à très haute résolution issues de la détection Google 2.5D 197
1. Présentation et objectif des données Google 2.5D 198
1.1 Téléchargement des données 200
1.2 Utilisation et préparation des données 201
1.3 Réinterpréter les modèles avec une identification de bâti plus fine 204
2. Répartition de la population dans les shapes Google 2.5D : une nouvelle désagrégation à Madagascar 206
3. Développement du modèle ascendant avec les shapes Google 2.5D de Madagascar 208
3.1 Création du modèle et analyse des coefficients 209
3.2 Prédictions et erreurs sur les données de validation 211
4. Exportation du modèle dans les communes d'Abidjan 213
4.1 Une nette amélioration par rapport à l'identification du bâti TeleCense 214
4.2 Des problèmes d'estimations qui persistent 216
5. Bilan et perspectives finales 219
5.1 Perspectives concernant la désagrégation 221
5.2 Perspectives concernant l'approche ascendante 223
5.3 Perspectives concernant l'approche globale 223
6. Conclusion 226
Conclusion générale 229
Liste des tableaux 237

Liste des figures 239
Annexes 243
Annexe 1 : Requête OSM 243
Annexe 2 : Mise en place de la méthode permettant de localiser les nouvelles communes dans les régions de Melaky, Atsinanana, Analamanga et Androy ... 248
Annexe 3 : Comparaison des estimations de la désagrégation TeleCense avec les estimations WorldPop dans les régions de Madagascar 253
Annexe 4 : Détection de Meta (data for good) en 2020 à Madagascar 256
Annexe 5 : Cartographie et surface TeleCense de 2017 à 2023 au Bénin 257
Annexe 6 : Projection des décomptes de population des arrondissements du Bénin 258
Bibliographie 265

Préambule

Sauf mention contraire, tous les traitements statistiques, figures et cartes ont été réalisés par l'auteur.

Sauf mention contraire, les analyses ont été réalisées avec le logiciel R. Les packages ggplot2 (Wickham, 2016) et mapsf (Giraud, 2024) ont été utilisés pour la réalisation des figures et des cartes.

Introduction générale

Selon les estimations des Nations Unies en 2024, la population mondiale compte aujourd'hui 8,2 milliards d'habitants, et devrait atteindre environ 9,7 milliards en 2050 avant de se stabiliser autour des 10 milliards d'habitants d'ici la fin du 21ème siècle (World Population Prospect, 2024). Cependant, cette croissance globale masque de profondes disparités régionales et c'est en Afrique, continent aujourd'hui peuplé de 1,4 milliard d'habitants, que se concentrera la majeure partie de cette augmentation démographique.

D'un point de vue strictement quantitatif, il semble logique que l'Afrique, avec ses 30,3 millions de km² — une superficie équivalente à celle de la Chine, de l'Inde, de l'Europe occidentale et des États-Unis réunis (Magrin, Dubresson, Ninot, Boissière, 2022) — absorbe une grande part de la croissance démographique mondiale. Cette tendance s'explique par sa densité de population, bien inférieure à celle de l'Asie du Sud, de l'Asie orientale ou de l'Union européenne, et par le fait que la transition démographique y est encore en cours dans de nombreux pays (Tabutin, Schoumaker, 2020), où l'on observe une baisse de la mortalité en décalage avec celle de la natalité.

Avec un Indice Synthétique de Fécondité (ISF) moyen de quatre enfants par femme, l'Afrique affiche le plus fort taux de fécondité parmi tous les continents, dépassant l'Océanie (2,1), l'Asie (1,9), l'Amérique latine et les Caraïbes (1,8), l'Amérique du Nord (1,6) et l'Europe (1,4) (Pison, Poniakina, 2024). Des disparités similaires existent au sein même de l'Afrique : alors que les pays d'Afrique du Nord et d'Afrique australe ont pour la plupart achevé leur transition démographique avec une forte croissance de la population entre 1950 et 1980 et désormais des ISF respectifs de 2,9 et 2,3, les autres régions de l'Afrique subsaharienne restent marquées par des taux de natalité plus élevés, contribuant à la forte dynamique démographique actuelle (Magrin et al., 2022; Pison, Poniakina, 2024). C'est notamment le cas pour le Tchad, la République Démocratique du Congo ou le Niger, qui ont des indices de fécondité qui avoisinent les six enfants par femme.

Finalement, les projections prévoient une population africaine atteignant 2,5 milliards d'habitants d'ici 2050, avec une stabilisation autour de 3.8 milliards vers 2100 (World Population Prospect, 2024). Pour ces pays d'Afrique en pleine croissance, il est essentiel de suivre et d'évaluer non seulement l'effectif total de la population mais aussi sa répartition spatiale. La connaissance de la population africaine est essentielle pour des enjeux de planification économique et sociale, notamment de la part des gouvernements nationaux dans l'optique de politique

publiques pertinentes concernant les infrastructures, les services de santé, l'éducation ou d'autres services publics.

Les recensements en Afrique : une répartition homogène sur le territoire et dans le temps ?

Afin de connaitre la population et de suivre son évolution à l'échelle nationale et ensuite à des plus petites échelles géographiques, l'outil le plus commun utilisé est le recensement de la population. Le recensement d'après la définition des Nations Unies est « un ensemble d'opérations qui consistent à planifier, recueillir, grouper, évaluer, analyser et diffuser des données démographiques, économiques et sociales se rapportant, durant un moment précis, à tous les habitants d'un pays ou d'une partie bien déterminée d'un pays » (Nations Unies, 2020). Il s'agit également d'une collecte statistique indispensable qui fournit une base de sondage pour la réalisation d'enquêtes ultérieures (Gendreau, 2019; Nations Unies, 2020).

Par ailleurs, Gendreau et Dackam-Ngatchou (2024) précisent que le recensement, en raison de son caractère exhaustif – englobant l'intégralité du territoire national, jusqu'au niveau administratif le plus détaillé – fait des données collectées une ressource précieuse pour de nombreux secteurs de développement, ainsi que pour les partenaires de développement, tels que les organisations internationales et les ONG, qui s'appuient sur ces informations pour leurs initiatives locales et nationales.

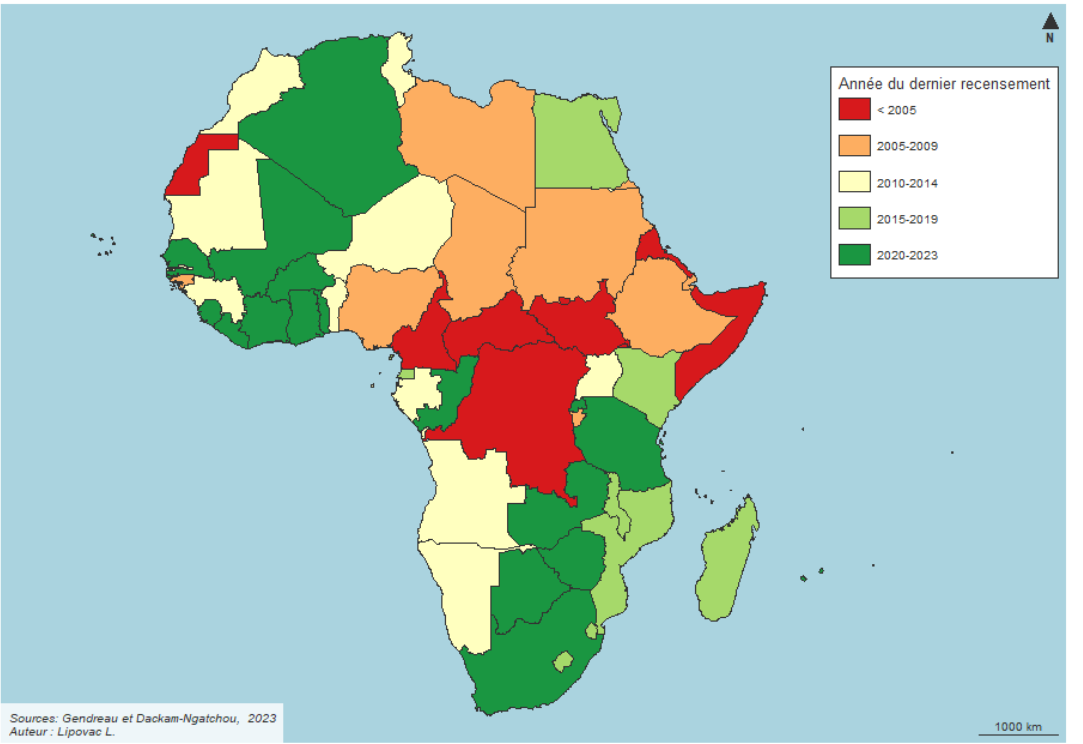
Les Nations Unies recommandent d'entreprendre un recensement tous les dix ans. D'après les informations de la figure 1, on observe que plusieurs pays africains ont récemment réalisé un nouveau recensement, c'est notamment le cas de deux des pays faisant partie de l'analyse de la thèse : la Côte d'Ivoire en 2021 et Madagascar en 2018. Le Bénin dont le dernier recensement date de 2013 est en train de réaliser son RGPH-5 et a terminé la cartographie censitaire le 31 mai dernier[3].

En revanche, plusieurs pays ont des recensements plus anciens. Le Nigeria, l'un des pays les plus peuplés d'Afrique, n'a pas mené de recensement depuis 2006, et son recensement prévu en 2016 a été régulièrement repoussé. Le Tchad, quant à lui, dispose de données datant de 2009. Dans certains cas extrêmes, des pays n'ont pas pu recenser leur population depuis plusieurs décennies : la République Démocratique du Congo, par exemple, n'a eu qu'un seul recensement, en 1984. Madagascar, avant son

[3] https://rgph5.instad.bj/2024/06/24/rgph5-linstad-et-la-banque-mondiale-satisfaits-du-deroulement-de-la-phase-de-la-cartographie-censitaire/

recensement de 2018, n'avait pas recensé sa population depuis 1993, ce qui représente un écart de 25 ans.

Figure 1 : Période du dernier recensement pour l'ensemble des pays africains



Une question essentielle est de savoir si, même après avoir mené un recensement récent, un pays sera en mesure de le reproduire dans les dix prochaines années. Le recensement représente une opération complexe et coûteuse, dont la réalisation n'est pas anodine. Les ressources financières et organisationnelles nécessaires rendent difficile sa pérennisation à intervalles réguliers. De plus, la stabilité politique nécessaire pour mener à bien cette opération tous les dix ans n'est pas toujours garantie, et des facteurs extérieurs, comme les épidémies d'Ébola ou de Covid-19, peuvent aussi entraver sa réalisation, comme cela a été observé dans plusieurs pays ces dernières années. Nous avons notamment l'exemple récent de la Côte d'Ivoire qui a repoussé de 2019 à 2021 son recensement, en plus de devoir s'assurer de la vaccination obligatoire des 36 000 agents avant déploiement sur le territoire (Gendreau, Dackam-Ngatchou, 2024).

Parmi les 54 pays d'Afrique analysés pour la période 1945-mi 2023, Gendreau et Dackam-Ngatchou (2024) indiquent que 26 pays ont été « rarement ou peu recensés », avec moins de cinq recensements, tandis que 28 ont été "régulièrement ou souvent recensés" avec six ou sept recensements ou plus. En répertoriant tous les

recensements réalisés depuis 1945 et en omettant les pays sans recensement ou avec un seul recensement, l'intervalle moyen entre chaque recensement s'élève à 11,1 ans. Il apparaît cependant que les pays ayant des recensements plus fréquents sont également ceux qui ont achevé ou presque achevé leur transition démographique, notamment en Afrique du Nord et australe. De fait, si l'on considère seulement l'intervalle entre les deux derniers recensements, 36 pays dépassent le seuil de dix ans entre deux campagnes de recensement (Gendreau, Dackam-Ngatchou, 2024).

La répartition intercensitaire est inégale et les difficultés de recensement sont particulièrement marquées dans les pays à faible et moyen revenu où des difficultés économiques, des conflits internes, des infrastructures inadéquates ou parfois un manque de volonté politique compliquent la collecte de données fiables (Wardrop et al., 2018). En l'absence de recensements réguliers, ces pays recourent souvent à des estimations de croissance démographique ou à des enquêtes socioéconomiques pour ajuster leurs données, limitant donc la précision des analyses géographiques et temporelles et risquant ainsi de masquer les variations locales et les dynamiques récentes.

Estimer la population à partir d'images satellites, une possibilité récente

Le besoin de données démographiques fiables et actualisées ne se limite pas à combler l'absence de recensements récents. Même lorsqu'un recensement a été mené au cours de la dernière décennie, les transformations rapides que connaissent certains pays, notamment en Afrique, rendent ces données rapidement obsolètes. Croissance démographique soutenue, urbanisation accélérée, fortes mobilités internes ou transfrontalières : autant de dynamiques qui modifient profondément la répartition spatiale de la population sur des temporalités bien plus courtes que les cycles de recensement traditionnels. Disposer de données à jour, est d'autant plus marqué dans les zones difficilement accessibles, en raison de contraintes politiques, logistiques ou de contextes de conflits, où l'accès au terrain est limité, voire impossible. Face à ces enjeux, de multiples recherches ont été développées afin d'estimer la population à partir d'images satellites. Ces travaux, s'appuyant sur les avancées des programmes d'observation de la Terre visent à produire des estimations datées et précises, notamment pour les pays du continent africain.

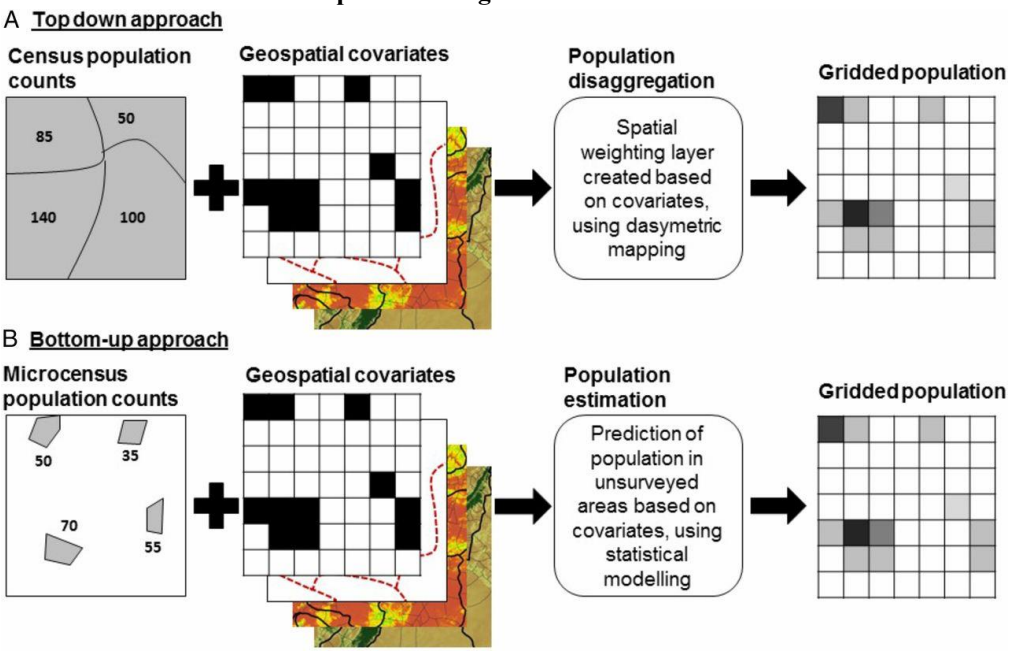
A ce jour, deux grandes familles de méthodes coexistent et ont été principalement popularisées par le programme de population WorldPop[4]. Tout d'abord, les méthodes descendantes – développées par plusieurs équipes de recherche depuis les années 2010, notamment dans le cadre des programmes AfriPop et WorldPop (Linard, Gilbert, Snow, Noor, Tatem, 2012 ; Stevens, Gaughan, Linard, Tatem, 2015) – sont les plus répandues et consistent à désagréger les effectifs communaux dans des unités plus petites et plus précises (Neal, Seth, Watmough, Diallo, 2022). Pour cela, il faut partir d'une unité administrative – généralement les zones administratives des pays concernés – et utiliser les données du dernier recensement ainsi qu'un certain nombre de données géographiques pour entraîner un modèle prédictif capable d'estimer la densité de population au niveau administratif et ensuite d'appliquer cette relation au niveau du carreau (Darin, Kuépié, Bassinga, Boo, Tatem, 2022).

Le second ensemble regroupe les approches ascendantes, conçues pour être employées en l'absence de recensement fiable ou disponible. Ces techniques reposent souvent sur des micro-recensements ou des enquêtes locales, qui permettent de créer des modèles à partir de données très précises. L'objectif est de généraliser les relations identifiées concernant la densité de population à cette échelle fine, afin de couvrir des zones plus vastes, sans nécessiter de données de recensement en entrée (Metzger et al., 2022). Ce type d'approche, plus récent, s'est surtout développé à partir du milieu des années 2010.

Les deux approches sont représentées schématiquement dans la figure 2 dans laquelle la population est distribuée de manière carroyée (ou encore maillée). Il s'agit, au lieu de représenter les données selon les découpages administratifs classiques (qui peuvent être compliqués à utiliser pour cause d'imprécisions ou d'inaccessibilité), de les diffuser selon une maille simple et originale, en carrés, de longueurs de côté variables. Par exemple, les estimations de densité de population du programme WorldPop sont généralement diffusées selon des grilles de carreaux de 100x100 mètres assurant une cohérence entre leurs différents travaux. Aujourd'hui, la visualisation carroyée s'est imposée comme le standard pour représenter les estimations de population à partir d'images satellites.

[4] https://www.worldpop.org/

Figure 2 : Approches descendantes et ascendantes de l'estimation de population carroyée à partir d'images satellites



Sources : Wardrop et al. 2018: « Spatially disaggregated population estimates in the absence of national population and housing census data »

Ces deux approches présentent chacune des avantages et des limites spécifiques. Par exemple, l'approche descendante n'est pas toujours adaptée à tous les contextes, en particulier lorsqu'un recensement récent n'est pas disponible. En effet, en l'absence de données actualisées, il est nécessaire de procéder à des projections démographiques, dont la fiabilité diminue avec le temps (Keilman, 2008). Cette incertitude est d'autant plus marquée lorsque les projections reposent uniquement sur des taux d'accroissement, sans intégrer des facteurs clés tels que la natalité, la mortalité, les migrations ou la mobilité spatiale.

Quant à l'approche ascendante, bien qu'elle soit innovante et essentielle dans de nombreux cas – comme pour les régions difficiles d'accès en raison de conflits ou les pays n'ayant pas réalisé de recensement depuis longtemps –, son développement reste limité. Cela s'explique en grande partie par la quasi-nécessité de réaliser un micro-recensement préalable, une opération qui exige des ressources importantes en termes de temps, de moyens financiers et de logistique et repose sur la possibilité d'accéder au terrain. Par ailleurs, son application demeure restreinte, car à ce jour et à ma connaissance, aucun modèle démographique basé sur cette approche ne parvient à obtenir des estimations fiables et précises dans des contextes différents de celui pour lequel il a été conçu. Il n'y a pas encore de projet capable de généraliser un modèle ascendant à l'échelle d'un continent par exemple.

Questions de recherche

Les travaux existants sur l'estimation de la population se concentrent principalement sur des approches carroyées, qui présentent des limites en termes de précision et de granularité. De plus, ces méthodes dépendent de données de recensement, ce qui les rend inopérantes dans les contextes où ces données sont obsolètes ou indisponibles. Face à ces défis, cette thèse propose d'explorer de nouvelles méthodes pour estimer la population résidentielle directement à l'intérieur des zones bâties, en s'appuyant sur les récents progrès en télédétection et en analyse de données spatiales. La question centrale est donc : dans quelle mesure est-il possible d'estimer la population résidentielle directement dans le bâti, de manière précise et à toute date, en s'appuyant d'abord sur des données de recensement par une approche descendante, puis en proposant des approches alternatives fonctionnant sans recensement préalable ?

Cette question de recherche soulève un double enjeu que nous expliciterons au fur et à mesure des chapitres de cette thèse. Premièrement, dans le cas où un recensement est disponible, comment exploiter ces données de manière optimale pour obtenir des résultats précis et utiles à une échelle fine ? En effet, les données de recensement sont souvent disponibles à une résolution spatiale trop grossière (par exemple, au niveau des districts ou des communes, qui peuvent couvrir plusieurs centaines de km²) et ne sont pas adaptées aux besoins de planification locale (Metzger et al., 2022). Se pose donc la question de savoir comment rendre ces données accessibles et exploitables à des échelles spatiales beaucoup plus fines. Pour y répondre, nous reprendrons et adapterons les méthodes de désagrégation spatiale, en les appliquant spécifiquement à des unités spatiales détaillées, telles que les bâtiments ou groupes de bâtiments.

Ensuite, comment suivre l'évolution de la population de manière précise en dehors des périodes de recensement ? Plus spécifiquement, il est essentiel de pouvoir continuer à estimer les populations dans les années suivant le recensement, surtout si un nouveau recensement n'est finalement pas réalisé dans les dix années qui suivent. Cela permettrait une mise à jour continue des données démographiques, garantissant ainsi leur pertinence et leur fiabilité à tout moment. Cette problématique peut s'étendre aux pays qui, comme la République Démocratique du Congo, n'ont pas pu réaliser de recensement depuis trop longtemps, et pour lesquels des estimations de population fiables et précises seraient très utiles dans l'élaboration de politiques publiques, comme base d'échantillonnage pour la préparation d'enquêtes spécifiques

ou dans le but de prévoir les moyens nécessaires au déploiement du prochain recensement.

Afin de répondre à ces questions et de comprendre les facteurs qui influencent la précision des estimations de population, une connaissance approfondie des méthodes et des données (démographiques et satellitaires) est nécessaire. Les données sur lesquelles reposent ce travail sont ici brièvement présentées afin de poser les bases nécessaires à la compréhension de la suite de la démarche de recherche.

L'identification du bâti à partir d'images satellites

Selon l'Agence spatiale européenne (ESA), la télédétection est un ensemble de techniques permettant d'obtenir des informations sur des objets en recueillant et en analysant des données sans contact direct entre l'instrument utilisé et l'objet étudié. Dans cette thèse, l'objet d'étude est le continent africain, et l'objectif est d'analyser plusieurs régions à partir d'images satellites pour identifier les zones bâties et estimer la population qui y réside. Avant de poursuivre cette analyse, il est essentiel de définir certains concepts clés.

Les images satellites

Une image satellite est une prise de vue transmise par un satellite en orbite autour de la terre. Chaque point élémentaire de l'image est un carré d'une taille variable que l'on définit comme un pixel ; assembler tous ces pixels permet de réunir toute l'information d'une image satellite. Une image satellite est définie par cinq caractéristiques ; en premier lieu la résolution, qui permet de comprendre à quelle distance un pixel de l'image est détecté. Cette distance varie de quelques mètres à quelques kilomètres et les objets qu'il est possible de discerner dépendent de la résolution spatiale utilisée. Vient ensuite la fréquence, qui permet de savoir au bout de combien de temps le satellite repasse au-dessus de la même zone géographique. Puis, les longueurs d'ondes et la source de l'émission déterminent respectivement le type de données collectées (bandes spectrales visibles, infrarouges ou radios) et la technologie utilisée (optique, radar et lidar) pour capturer ces longueurs d'ondes. Pour finir, la couverture géographique est la zone que le satellite couvre.

Les images satellites sont utilisées dans de nombreux domaines tels que la surveillance de l'environnement, l'agriculture, la cartographie ou encore la météorologie. Un grand nombre de satellites est nécessaire afin de répondre à ces

besoins. D'après la base de données satellites UCS[5] il y aurait en ce moment plus de 7500 satellites en orbite dans le ciel.

Ces satellites proviennent de différents organismes fournisseurs dont les plus connus sont : Maxar, Landsat et Copernicus. Parmi eux, seuls Landsat et Copernicus offrent des images actuelles et gratuites. Landsat, lancé en 1972, fournit des images multispectrales d'une résolution allant jusqu'à 30 m avec un taux de rafraîchissement élevé (Artadji, Lipovac, Hasina Andriamanantena, Rousse, 2024).

Copernicus est le programme d'observation de la Terre et de l'Union européenne qui fournit des services d'informations basés sur l'observation de la Terre par satellite, et ce sur, 6 thèmes primordiaux : (1) La surveillance de l'atmosphère, (2) la surveillance du milieu marin, (3) la surveillance des terres, (4) le changement climatique, (5) sécurité / services liés à la sécurité, (6) urgences / la gestion des urgences. Les services d'informations de Copernicus sont basés sur une constellation de 6 familles de satellites connus sous le nom de Sentinel et dont le premier nommé « Sentinel-1A » a été lancé en 2014. Les « Sentinels » ont pour objectif d'assurer une source de données de haute qualité pour les services de Copernicus mais également de remplacer les anciennes missions d'observation de la terre comme ERS et Envisat afin d'assurer une continuité. Depuis 2016 et l'association avec Sentinel-1B, Copernicus délivre une image tous les 6 jours avec une résolution de 10 mètres (et depuis 2022, Sentinel-1B est hors service donc les données radar fournies ne sont plus disponibles que tous les 12 jours[6]). Il existe six missions Sentinel, mais seules les deux premières concernent la télédétection :

Sentinel-1 (figure 3) est composé de deux satellites (d'un satellite depuis 2022) en orbite polaire fonctionnant jour et nuit et réalise des images radar dont les ondes permettent de capturer des images de jour et de nuit et ce peu importe la météo ou la couverture nuageuse.

[5] Source : UCS Satellite Database : https://www.ucsusa.org/resources/satellite-database.
[6] https://www.agrotic.org/veille/sentinel-1b-prend-sa-retraite-plus-tot-que-prevu/

Figure 3 : Satellite Sentinel-1

Source : ESA - Introducing Sentinel-1

Sentinel-2 (figure 4), lancé en 2015, a quant à lui pour objectif la surveillance terrestre, la surveillance de la végétation, des sols et des zones côtières en fournissant des images optiques à haute résolution. Comme l'optique utilise la lumière solaire réfléchie par la surface de la Terre, elle est dépendante de la présence de lumière solaire et n'est pas efficace de nuit ou lors d'une forte couverture nuageuse. C'est pour cela que les images optiques sont souvent couplées aux système radar.

Figure 4 : Satellite Sentinel-2

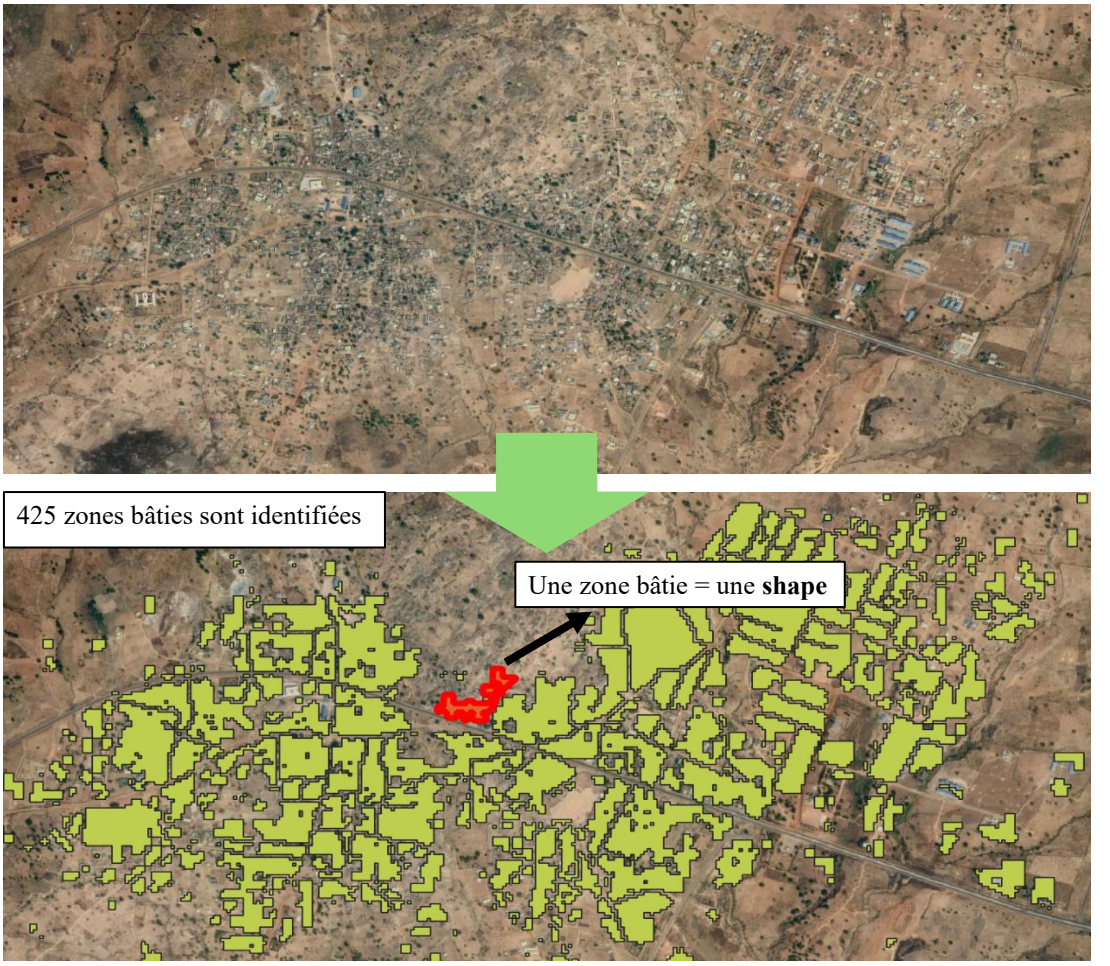
Source : ESA - Sentinel-2

Identifier les zones bâties

Le programme TeleCense[7] créé par l'entreprise Diginove, vise à fournir des estimations de population localisées, à une échelle biannuelle, à partir du traitement d'images satellites européennes en accès libre et des données démographiques disponibles, de manière à accompagner le développement et la planification, en milieu urbain comme en milieu rural. Pour cela, TeleCense utilise la combinaison d'images des satellites Sentinel 1&2, afin d'identifier, puis de caractériser des zones bâties en ce qu'on appelle communément dans ce programme des « shapes » (chacune des formes vertes – figure 5).

Ces zones bâties sont le point de départ de nos travaux, dans le sens où c'est la population de ces shapes que nous souhaitons estimer.

Figure 5 : Illustration des algorithmes de détection TeleCense – identification de zones bâties dans la région d'Abuja, Nigéria - 2018.



Source : TeleCense – Image issue du logiciel QGIS

[7] https://diginove.com/index.php/fr/2018/03/08/telecense-2/

Comme l'objectif de la thèse est d'estimer la population directement à l'intérieur du bâti et non plus de manière maillée, nous devons faire l'hypothèse que le projet TeleCense est capable de détecter correctement les zones bâties dans la région étudiée. Il faut néanmoins rappeler que les images mobilisées par le projet TeleCense ont une résolution de 10 mètres et que la détection n'est de fait, pas infaillible. Il peut donc y avoir des erreurs de faux-positifs et de faux-négatifs[8], c'est-à-dire des erreurs de classification dans le processus de détection des zones bâties. Le fait de travailler avec les images Sentinel est un choix assumé et réfléchi du fait de leur avantage économique (elles sont en accès libre), temporel (elles permettent de suivre l'évolution du bâti depuis 2015, année de lancement du programme) et de temps de traitement (images beaucoup moins lourdes à traiter et à stocker). Cependant, l'utilisation de ces images à dix mètres de résolution présente un défi potentiel en termes de précision, notamment par rapport à d'autres fournisseurs. Ainsi, nous émettons l'hypothèse que cette résolution pourrait être suffisante pour notre objectif, à savoir de correctement répartir la population sur le territoire, mais qu'elle risque cependant d'être insuffisante pour produire des estimations de population à l'échelle des bâtiments. Il est nécessaire pour cela d'utiliser des images de résolution plus fine, qui captent mieux la diversité et les caractéristiques du bâti.

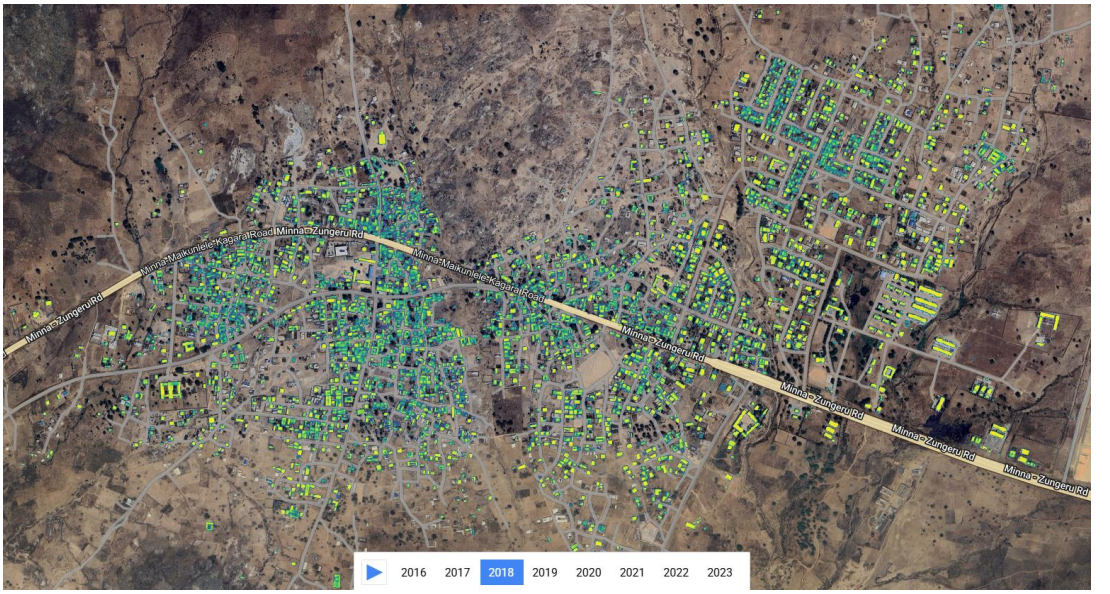
Les données du programme « Open Buildings 2.5D Temporal Dataset[9] » publiées en septembre 2024 par Google Research, fournissent une détection de bâti plus fine grâce à l'association d'images Sentinel et d'autres images très haute résolution (de 0.5 à 1 mètre). L'identification des bâtiments ainsi que leur hauteur sont donnés avec une résolution spatiale de 4 mètres pour les années 2016 à 2023. La figure 6 permet de comparer les détections des deux programmes dans la même ville du nord-ouest d'Abuja[10] et illustre effectivement la précision accrue des données Google 2.5D, avec une détection quasiment bâtiment par bâtiment (la teinte de couleur variant du jaune au bleu est l'indice de confiance liée au fait d'identifier le bâtiment).

[8] Faux positifs : Ce sont des zones identifiées à tort comme étant bâties alors qu'elles ne le sont pas. Par exemple, une forêt dense, une zone agricole ou des formations géologiques spécifiques peuvent être confondues avec des structures artificielles (bâtiments) en raison de leur texture ou de leur forme sur les images satellitaires. Un exemple courant est de confondre un rocher avec un bâtiment ; erreur que l'on peut contrecarrer en utilisant des images en saison sèche et en saison humide car le rocher n'aura plus la même caractérisation entre les deux saisons et peut être définitivement écarté des zones bâties identifiées.
Faux négatifs : Ce sont des zones effectivement bâties qui ne sont pas détectées comme telles.

[9] Open Buildings 2.5D Temporal Dataset

[10] Visualiser la détection des zones bâties en 2018 par Earth Engine Apps : https://mmeka-ee.projects.earthengine.app/view/open-buildings-temporal-dataset#longitude=6.476453237214019;latitude=9.683504071293072;zoom=15;year=2018;band=0

Figure 6 : Identification des zones bâties Google 2.5D dans la région d'Abuja, Nigéria - 2018



Source : Identification du bâti issue du programme Open Buildings 2.5D Temporal Dataset
Image issue de l'application Earth Engine Apps

Intérêt de la visualisation par shapes plutôt que par format carroyé

Le format carroyé s'est imposé comme standard dans la cartographie de la population grâce à plusieurs avantages : il est simple à produire et, sous forme raster (c'est-à-dire une grille de cellules régulières attribuant une valeur de population à chaque carreau), il facilite la superposition avec d'autres informations géographiques et l'agrégation des données de population à n'importe quel niveau souhaité. Dans le cadre du projet WorldPop, initialement développé pour des applications de santé, ce format était particulièrement adapté car il permettait d'évaluer rapidement la population affectée par un désastre naturel (inondation, cyclone) en croisant directement la grille avec la zone touchée.

Les projets historiques comme le Gridded Population of the World (GPW) ou le Global Rural Urban Mapping Project (GRUMP) utilisaient des carreaux d'1 km de côté. Grâce aux avancées technologiques, les résolutions se sont progressivement affinées : 250 m avec le GHSL (Global Human Settlement Layer), 100 m avec WorldPop, voire 30 m avec le HRSL (High Resolution Settlement Layer).

Aujourd'hui, grâce aux images Sentinel ou Google, il est possible d'identifier les bâtiments ou groupes de bâtiments à une résolution égale ou inférieure à 10 m. Cette évolution ouvre la voie à une approche vectorielle, où les unités d'analyse ne

sont plus des carreaux arbitraires mais des shapes correspondant aux zones bâties identifiées. Cette représentation reste compatible avec des traitements rapides, comme la sélection d'une zone d'intérêt ou l'agrégation de la population, tout en offrant une meilleure pertinence spatiale pour la modélisation de l'activité humaine, puisqu'elle évite d'affecter artificiellement des habitants à des zones non résidentielles telles que les cimetières, les casernes, les usines ou les bâtiments administratifs. Les shapes sont des entités vivantes qui peuvent évoluer dans le temps, conservant ainsi une cohérence géographique. De nouvelles variables peuvent leur être associées, telles que la surface de la shape, la densité de bâti ou encore la hauteur des bâtiments, récemment appréhendée par le projet Google Research. Ces caractéristiques sont mises à jour à chaque nouvelle détection des constructions, permettant un suivi précis de l'évolution du bâti et, par conséquent, de la population.

Cette nouvelle visualisation permet également des applications plus fines, qu'il s'agisse d'urbanisme – par exemple pour analyser les îlots de chaleur, planifier des réseaux de transport ou évaluer la couverture mobile –, ou encore d'évaluation de la qualité de l'air et du risque sanitaire, où une précision de l'ordre de quelques dizaines de mètres est déterminante (la qualité de l'air pouvant varier fortement à 100 mètres près). Enfin, dans le cas de la montée du niveau de la mer ou de l'érosion côtière, il est bien plus pertinent de connaître la population à l'échelle des bâtiments afin d'évaluer le risque en fonction du type de bâti et du nombre d'habitants potentiellement exposés.

Les données démographiques

Les données issues des recensements, ainsi que les limites administratives correspondant aux zones géographiques étudiées, constituent un préalable indispensable à ce travail de recherche. Les décomptes de population sont issus d'un travail de collecte de la part des gouvernements et des instituts nationaux de la statistique. Lorsqu'ils sont disponibles sous forme de bases de données structurées, leur exploitation est directe. Toutefois, il est fréquent que ces informations ne soient accessibles que via des rapports publiés à l'issue des recensements, souvent en format PDF. Dans ce cas, un travail de reprise et de mise en forme est nécessaire pour les intégrer dans des formats adaptés à l'analyse statistique et spatiale.

En parallèle, il est nécessaire de disposer des limites administratives correspondant aux zones géographiques étudiées afin de relier les décomptes de population à leur unité spatiale d'origine. Ces découpages peuvent être obtenus auprès des instituts nationaux eux-mêmes, mais proviennent le plus souvent de bases de

données de référence utilisées dans les projets de cartographie démographique, telles que GADM, geoBoundaries ou HDX (Humanitarian Data Exchange). Les frontières administratives sont organisées selon une hiérarchie standardisée : le niveau 0 correspond à l'ensemble national, suivi du niveau 1 (régions, provinces ou départements), puis des niveaux 2 et 3 et voir plus comment on pourrait dire ça (districts, communes, etc.), selon les spécificités administratives propres à chaque pays.

L'articulation entre décomptes de population et limites administratives est fondamentale car une fois ces deux types de données reliées, elles sont utilisées dans l'approche descendante afin de répartir la population des zones administratives dans les zones bâties identifiées à partir d'images satellites. Elles sont également mobilisées dans le cadre des approches ascendantes pour évaluer les estimations produites.

Positionnement de la thèse

La majeure partie de ce travail de thèse repose donc sur l'identification du bâti TeleCense. Travailler avec l'identification du bâti d'une entreprise privée n'est pas anodin. Je suis effectivement en contrat doctoral avec l'entreprise Diginove, porteur du projet TeleCense. De fait, le projet doctoral a été pensé d'une part, pour répondre à ma question de recherche et d'autre part, pour répondre à un besoin de l'entreprise qui, avec ce projet de cartographie de la population, souhaitait mettre au point une méthode fiable pour estimer la population de manière précise au sein des zones géographiques étudiées, qu'elles soient urbaines ou rurales.

Cette thèse est la dernière étape d'un parcours universitaire débuté par une licence de mathématiques et informatique, poursuivi du master mathématiques appliquées et sciences sociales – analyse des populations (MASS POP) de Aix-Marseille Université où dès la première année, j'ai effectué mon mémoire dans l'entreprise Diginove. Durant ces trois mois de stage, j'ai pu me familiariser à la télédétection, aux méthodes d'observation de la terre à partir d'images satellites et commencer un travail d'estimation de population des agglomérations urbaines de la région d'Abuja au Nigéria (Lipovac, 2020). Ce champ de recherche nouveau pour moi, et ces techniques particulières m'ont poussé à continuer à travailler dans ce domaine en master-2 à travers une alternance, toujours dans l'entreprise Diginove. Dans mon mémoire de master-2, j'ai pu explorer les méthodes d'analyses spatiales, techniques très peu utilisées dans ce domaine, pour venir compléter la modélisation

linéaire classique de la population des agglomérations, toujours à Abuja (Lipovac, 2021).

Dans la continuité de ces deux premiers travaux, cette thèse tire parti de ces recherches initiales pour effectuer ce travail d'estimation à l'échelle nationale, ce qui permet une meilleure adaptabilité à d'autres contextes nationaux. Nous avions initialement sélectionné le Kenya, qui, en raison de sa diversité agro-climatique, de la qualité de son dernier recensement (2019), et des connaissances apportées par ma directrice de thèse (Golaz, 2009), semblait être une option idéale. Cependant, en raison de l'inaccessibilité d'une donnée primordiale (les découpages administratifs du Kenya), nous avons dû changer de pays. Finalement, le travail a pris forme à Madagascar, un choix tout aussi pertinent et même rassurant, compte tenu du fait que de nombreux chercheurs des équipes Demosud et du LPED travaillent à Madagascar et que nous avons de nombreux contacts à l'Institut national de la statistique (INSTAT).

Par ailleurs, les objectifs et le projet final de cette thèse ont régulièrement évolué au cours de ces trois années. En effet, ce travail de doctorat a été influencé par les activités de l'entreprise Diginove et par certaines échéances liées à leurs projets en cours. C'est en partie pour ces raisons que j'ai mené un cas d'application sur le Bénin et un exercice de validation dans le département de Mayotte. J'ai dû m'adapter aux situations et aux opportunités qu'elles m'ont offertes pour continuer à travailler et tenter de produire un travail de qualité, tant pour l'entreprise et ses divers besoins en estimation de population que pour le manuscrit de thèse.

C'est dans ce sens que, la sortie du jeu de données « Open Buildings 2.5D Temporal Dataset » en septembre 2024 – soit à quelques mois de la fin de mon contrat doctoral – a ouvert de nouvelles perspectives qui ont fait évoluer les orientations du travail final . En effet, le dernier chapitre de la thèse, consacré à l'évaluation de ces nouvelles données n'était initialement pas prévu. C'était néanmoins une opportunité que je ne pouvais pas manquer pour conclure cette recherche, qui m'a permis de comparer les résultats avec ceux obtenus via le programme TeleCense.

Démarche et structure de la thèse

La première partie de cette thèse porte sur les données et les méthodes. Après être revenu dans le chapitre 1 sur les différentes méthodes d'estimation de la population et avoir présenté les porteurs et les projets de cartographie de population existants, nous détaillons dans le chapitre 2 le programme TeleCense ainsi que les méthodes utilisées dans les chapitres suivants afin de mettre en œuvre les approches

de modélisation de population. Les différentes variables mobilisées seront présentées ainsi que certaines spécificités de la caractérisation et de l'identification des shapes pouvant affecter la précision de l'estimation finale de la population.

La deuxième partie de la thèse (chapitres 3 et 4) est consacrée à l'élaboration d'une méthode descendante, une approche qui fonctionne dès lors que l'on dispose d'un minimum de données démographiques assez récentes. La mise en œuvre de cette désagrégation se fera sur Madagascar, dans six régions aux caractéristiques environnementales, climatiques et de bâti variées. Ce choix permet de capturer une diversité maximale tout en limitant les coûts en temps et en stockage. L'enjeu est de comprendre les relations entre la population, la structure du bâti et l'environnement, afin de répartir avec le plus de précision possible la population. La principale différence avec les autres projets de désagrégation, qui estiment la population de manière carroyée, est que notre approche vise à estimer directement la population au niveau des zones bâties.

Par analogie avec les travaux de WorldPop et des autres projets de cartographie de la population, notre méthode descendante repose également sur des données de recensement. C'est pourquoi le chapitre 3 commence par la présentation d'une méthode permettant de lier les décomptes de population du 3ème Recensement général de la population et de l'habitation (RGPH3) aux territoires des communes de Madagascar. C'est une étape indispensable qui peut cependant s'avérer très compliquée si une des deux informations n'est pas disponible ou pas complètement à jour. Elle souligne aussi l'importance du travail de terrain pour recenser la population et mettre à jour le découpage administratif, un besoin incontournable, même pour les approches basées sur l'imagerie satellitaire.

Dans la suite du chapitre, les différentes méthodes de répartition de la population – celles présentées dans le chapitre 2 – sont testées dans l'objectif de comprendre si les méthodes adaptées peuvent être ajustées pour offrir des résultats aussi précis, voire meilleurs, lorsqu'elles sont appliquées aux shapes de TeleCense. Nous cherchons à comprendre si ces méthodes peuvent s'adapter à cette nouvelle unité d'analyse. Nous faisons notamment l'hypothèse qu'à partir du moment où nous disposons de décomptes de population associés à des limites administratives relativement précises, la désagrégation au niveau des shapes sera tout aussi précise. Il semble en effet pertinent de penser que notre approche, privilégiant d'utiliser le bâti comme donnée d'entrée directe, plutôt qu'en tant que variable dérivée (par ex. distance ou présence) comme dans le projet d'estimation de WorldPop (Stevens, Gaughan, Linard, Tatem, 2015) puisse donner des résultats plus précis, a mesure que

l'on estime directement la population à l'endroit où les habitants vivent et non pas une densité de population estimée au niveau de carreau de grille.

Dans la continuité, une validation supplémentaire est menée au Bénin dans le chapitre 4, où les résultats du RGPH-4 de 2013 doivent être projetés sur plusieurs années pour permettre la désagrégation. La cartographie du bâti sur plusieurs années consécutives soulève des questions sur nos hypothèses liées au bâti et invite à réfléchir à la cohérence d'utiliser la désagrégation lorsque le recensement est trop lointain.

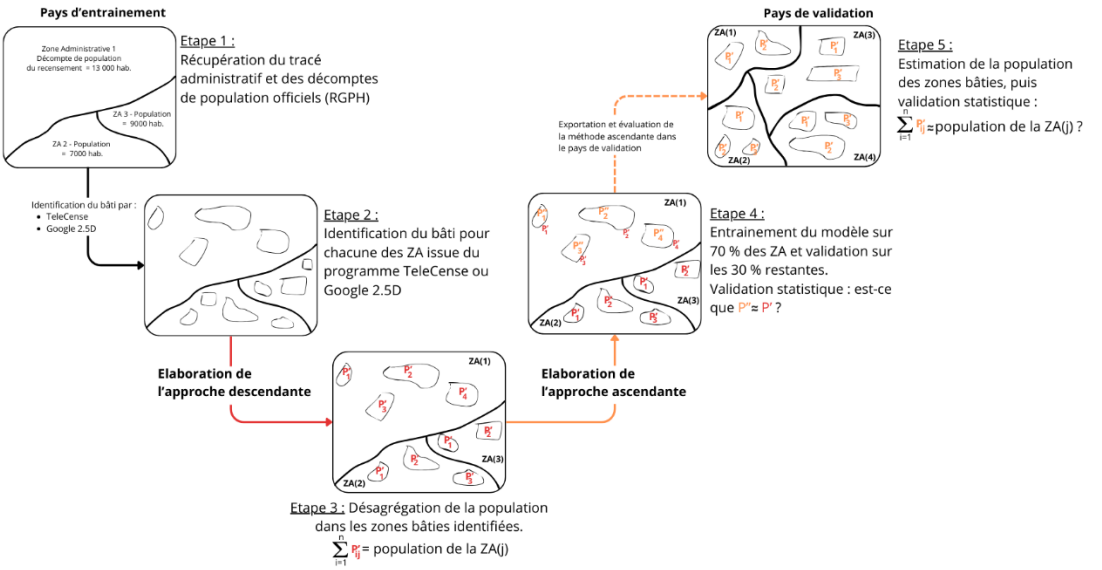
Cette partie de la thèse à travers l'approche descendante, apporte une première réponse essentielle pour la distribution et la répartition de la population lorsqu'un recensement est disponible. Dans la troisième et dernière partie, les méthodes ascendantes sont explorées pour produire une estimation de la population sans données de recensement en entrée. Les chapitres 5 et 6 visent à répondre à une question clé : est-il possible de concevoir un modèle démographique totalement indépendant d'un recensement, y compris sans micro-recensement ou enquête ? L'objectif est de créer un modèle détaché de toute donnée de population en entrée, ce qui le rendrait plus généralisable et exportable au-delà de la zone géographique ou du pays où il a été initialement développé. Cette approche est d'autant plus importante que l'une des principales limites des méthodes ascendantes actuelles réside dans leur difficulté à être transposées à d'autres contextes géographiques.

L'idée centrale de cette thèse est là, il s'agit de créer un modèle démographique innovant capable d'estimer la population des zones bâties d'une région en l'absence de données de recensement. Pour y parvenir, il est nécessaire de disposer, dans les zones bâties détectées, d'une donnée de référence sur laquelle entraîner le modèle. C'est en entrainant le modèle sur des données existantes que le modèle est capable de s'exporter et d'estimer la population de zones bâties similaires mais d'un autre contexte géographique. Ce modèle, appelé modèle ascendant ou "bottom-up", s'oppose au modèle précédent de désagrégation, qui est dit descendant.

Pour ce faire, nous proposons d'utiliser les résultats obtenus avec la méthode descendante afin de développer la méthode ascendante. Concrètement, nous commençons par répartir la population dans les zones bâties (shapes) en utilisant la méthode offrant la plus grande précision (étape 3 de la figure 7). Ensuite, nous utilisons cette population répartie comme donnée de référence pour entraîner un modèle sur de nombreuses zones bâties, permettant ainsi une généralisation à d'autres contextes (étape 4 de la figure 7). Le chapitre 5 détaillera dans un premier temps les étapes d'élaboration de la méthode ascendante sur Madagascar, puis dans un second temps les étapes de vérification et de validation sur des régions extérieures à

Madagascar. Nous commencerons par évaluer le modèle ascendant à Mayotte, département français proche de Madagascar et relativement similaire en termes de structures de bâti, où l'INSEE a fourni à Diginove des données de recensement très précises, puis dans la sous-préfecture d'Abidjan, région urbaine relativement dense, pour l'Afrique sub-saharienne.

Figure 7 : Développement étapes par étapes des approches descendantes et ascendantes de la thèse



L'une des questions les plus importantes est de savoir si les relations observées dans la région test (les six régions de Madagascar) peuvent être généralisées à d'autres contextes géographiques. L'hypothèse est que dans des environnements bâtis similaires, cela pourrait être possible, mais dans d'autres, cela pourrait s'avérer plus difficile. L'objectif est donc d'évaluer dans quelle mesure les relations établies dans le pays test peuvent être appliquées à d'autres pays ou régions.

Plus largement, il s'agit d'examiner la validité scientifique de l'approche adoptée : peut-on véritablement partir des résultats de l'approche descendante pour construire l'approche ascendante ? En théorie, la méthode ascendante est indépendante du recensement, puisqu'elle permet d'estimer la population sans données de recensement en entrée. Toutefois, ici, son développement repose initialement sur une phase d'apprentissage utilisant des résultats issus du recensement (désagrégation). Un point clé est donc de savoir dans quelle mesure cette dépendance initiale pourrait introduire un biais. Le modèle repose sur l'hypothèse implicite que la

désagrégation de la population au niveau des shapes (et donc des bâtiments) est précise et fiable. Si cette hypothèse ne tient pas, il devient difficile d'espérer des estimations précises en sortie du modèle ascendant.

Enfin, une réflexion globale s'impose sur la méthode de désagrégation la plus adaptée : vaut-il mieux utiliser une répartition simple, avec peu de variables, et laisser le modèle ascendant identifier d'autres relations pertinentes ? Ou, au contraire, une désagrégation plus complexe, intégrant davantage de variables, est-elle préférable, au risque d'entraîner un surapprentissage qui limiterait la capacité du modèle à s'adapter à d'autres contextes ?

Un autre enjeu clé concerne la capacité à estimer la population au-delà de 2020, un défi face auquel de nombreux projets existants rencontrent encore des difficultés. Le modèle ascendant, construit sur une région test dont le recensement date de 2018, est-il encore pertinent pour estimer la population des dates différentes, notamment au-delà de 2020 ? Dans ce cas, faudrait-il ajuster les estimations nationales à l'aide des données des Nations Unies, ou bien l'observation de l'évolution du bâti suffira-t-elle à refléter correctement l'évolution de la population ? Nous faisons l'hypothèse que si la détection du bâti reste suffisamment précise sur plusieurs années, et que le modèle intègre d'autres variables socio-démographiques liées à la croissance démographique, les estimations de population devraient rester pertinentes. L'enjeu sera de savoir si ce modèle peut capturer fidèlement les dynamiques démographiques au fil des années. Par exemple, dans une zone où le bâti stagne, il est possible que la population continue à croître en raison de facteurs démographiques classiques, comme un taux de natalité élevé ou des migrations internes. Dans ce cas, l'évolution du bâti ne suffirait pas à refléter l'augmentation de la population, et il serait nécessaire d'ajouter des variables supplémentaires pour affiner l'estimation.

Ces hypothèses soulèvent des questions méthodologiques cruciales que nous chercherons à résoudre au fil de notre étude. Les réponses à ces questions détermineront non seulement la validité de nos approches, mais aussi leur capacité à être généralisées à d'autres contextes géographiques.

Enfin, le dernier chapitre de la thèse viendra réaliser la même méthode globale (étape 3 à 5 de la figure 7) mais avec la détection du bâti issue du programme Google 2.5D. Cette identification plus précise permettant d'obtenir des zones bâties de plus petites surfaces est utilisée afin de comparer les résultats de l'approche descendante et ascendante à Madagascar et à Abidjan selon deux identifications du bâti différente et de statuer si une résolution plus fine couplée à des données de hauteur de bâtiment est bénéfique ou non pour l'estimation de population à partir d'images satellites.

PARTIE 1 :
Etat de l'art et approche méthodologique pour l'estimation de population à partir d'images satellites

Chapitre 1 – De l'estimation démographique aux projets de cartographie : un état des lieux

Ce chapitre propose un état des lieux des méthodes et des initiatives clés en matière d'estimation démographique et de cartographie de la population. Il s'agit de revenir en détail sur les approches méthodologiques brièvement évoquées en introduction, en commençant par les méthodes descendantes, puis par les méthodes ascendantes. Nous examinerons ensuite les principaux projets de cartographie de la population ainsi que les fournisseurs de données satellitaires qui les soutiennent. Enfin, nous présenterons le projet Africapolis, un précurseur essentiel pour le projet TeleCense de cartographie du bâti, qui constitue le socle de notre travail de thèse. À travers cette revue, ce chapitre vise à offrir une vision globale des pratiques actuelles et à éclairer les choix méthodologiques que nous avons adoptés dans la suite de cette recherche.

1. Méthodes descendantes

Le principe de la cartographie de la population descendante est de partir d'unités administratives pour lesquelles on dispose de données de population (recensement ou projections) puis de les désagréger dans les unités plus petites qui les composent (Wardrop et al., 2018).

Les premières utilisations d'images satellites pour estimer la population remontent à la fin des années 1990 (LO, 1995; Sutton, Roberts, Elvidge, Baugh, 2001). Ces travaux étaient toutefois limités à des zones géographiques spécifiques, principalement des environnements urbains.

Le véritable tournant dans la cartographie de population s'opère au milieu des années 2000 avec les projets AfriPop et AsiaPop, qui ont été fusionnés en 2013 avec AmeriPop sous le nom plus connu de WorldPop (WorldPop, 2024). En standardisant les méthodes de désagrégation de la population à l'échelle mondiale, WorldPop a non seulement amélioré la précision des estimations démographiques dans des zones variées, mais a également facilité l'intégration de nouvelles sources de données, telles que les images satellites à haute résolution, les enquêtes démographiques et les données géographiques. Ces avancées motivent le fait de commencer l'analyse des méthodes descendantes sous l'angle des développements du projet WorldPop.

Le projet AfriPop a entre 2005 et 2013 développé des méthodes permettant la répartition de la population à partir des classes d'occupation du sol et des délimitations urbaines et rurales (Linard, Gilbert, Snow, Noor, Tatem, 2012; Tatem, Noor, Hagen, Gregorio, Hay, 2007). La première itération du projet a permis de montrer que la combinaison de données de recensement à haute résolution avec des informations sur l'habitat et la couverture du sol permettait la création de cartes de population précises et à moindre coût. Plus précisément, ce sont des densités de population par classes d'occupation du sol qui ont été calculées grâce à des données précises provenant des zones d'énumération du recensement du Kenya en 1999. Une fois les densités connues, elles pouvaient être utilisées comme poids pour répartir de manière dasymétrique la population. La cartographie dasymétrique est une méthode d'interpolation spatiale qui consiste à désagréger des données surfaciques vers des unités plus petites en s'appuyant sur des informations auxiliaires pour mieux localiser les phénomènes étudiés (Mennis, 2003; Vignes, Rimbourg, 2013). Cette approche est couramment utilisée pour la répartition spatiale de la population ou d'autres variables démographiques (Hallot, Grippa, Stephenne, Wolff, 2019).

Le projet AfriPop donne lieu à une deuxième itération en 2012 en utilisant cette fois-ci les données de la base de données mondiale d'occupation des sols nommée GlobCover[11] (appelée aujourd'hui WorldCover[12]). GlobCover proposait une classification des types d'occupation des sols allant de la forêt aux terres agricoles jusqu'aux zones urbaines, et ce à une résolution de 300 mètres. Dans le même temps, le projet AsiaPop utilise une approche similaire en s'appuyant cependant sur davantage de données existantes, notamment des informations sur les zones climatiques, et produit le même type de cartes sur le continent asiatique (Gaughan, Stevens, Linard, Jia, Tatem, 2013).

Les distributions dasymétriques de la population dans ces projets, relativement précises pour l'époque et le nombre de variables utilisées, vont connaître une évolution significative avec la fusion d'AfriPop et AsiaPop dans le projet global WorldPop. En 2015, l'équipe de WorldPop propose une nouvelle approche pour désagréger la population. La méthode demeure fondée sur une redistribution dasymétrique et utilise toujours les décomptes de population du dernier recensement en date, à partir des zones administratives les plus précises possibles. Au lieu de se concentrer sur des liens par classe d'occupation des sols, ce qui change, c'est le schéma de pondération, qui intègre désormais un ensemble de variables socio-économiques et géographiques, relatives au bâti et à l'environnement. Pour ce faire,

[11] https://due.esrin.esa.int/page_globcover.php
[12] https://esa-worldcover.org/en

un modèle de forêt aléatoire[13] est créé, utilisant un grand nombre de covariables pour générer la couche de pondération qui estime le nombre de personnes par pixel de 100 x 100 mètres (Stevens et al., 2015). Ce modèle repose sur les zones administratives, et la variable de réponse étudiée est le logarithme de la densité de population. Ce renouvellement méthodologique ouvre la voie à une amélioration de la précision des estimations de population

En utilisant l'approche dasymétrique pour répartir la population, il faut, indépendamment de la couche de pondération retenue, supposer que les relations établies au niveau de l'unité administrative puissent être transposées à l'échelle du pixel de 100 x 100 mètres. Étant donné que les variables propres aux unités administratives sont calculées à partir des pixels (par exemple, pour les jeux de données continues tels que l'intensité de la lumière la nuit, on calcule la moyenne zonale qui est la moyenne des valeurs de tous les pixels dans une zone administrative) cette méthode se justifie.

Concernant les variables utilisées, les recherches précédentes (AfriPop et AsiaPop) ont montré que la répartition de la population était souvent corrélée avec les types de couverture terrestre (Stevens et al., 2015). En se basant sur ces constatations, WorldPop intègre non seulement ces variables, mais également des éléments désormais classiques dans le domaine de la télédétection, tels que les données climatiques (précipitations et température), l'intensité lumineuse observée la nuit (qui renseigne sur l'utilisation de l'électricité et l'urbanisation), ainsi que les données d'altitude (élévation et pente). De plus, la Productivité Primaire Nette (NPP), qui mesure la quantité d'énergie disponible sous forme de biomasse dans un écosystème, est également prise en compte. En plus de ces variables, les chercheurs ajoutent de nombreuses autres variables potentiellement corrélées avec la présence humaine, telles que les réseaux de routes et de voies navigables, les grands plans d'eau, les zones de peuplement (au sens d'OpenStreetMap, c'est-à-dire des zones géographiques habités, comme des villes, villages, hameaux ou autre localités), la délimitation des aires protégées, ainsi que la présence de bâtiments comme des hôpitaux ou des écoles.

Afin de créer une couche de pondération, il est essentiel de disposer des mêmes variables à la fois au niveau de l'unité administrative et au niveau du carreau de 100x100 mètres. Le fait d'utiliser des données carroyées est un avantage car étant des données spatiales harmonisées il est facile de créer la variable au niveau administratif, par exemple par moyenne zonale comme mentionné précédemment ou en comptant le nombre de pixels d'une catégorie pour une variable binaire. Par ailleurs, l'utilisation

[13] https://github.com/wpgp/wpgpRFPMS

de grilles uniformes présente un autre avantage majeur : elle facilite la comparaison entre différents travaux de recherche. En effet, grâce à cette homogénéité, les données sont directement comparables (OCHA, 2020), ce qui permet d'éviter les biais liés à la variabilité des unités spatiales. Cette approche résout également les problèmes posés par le MAUP (Modifiable Areal Unit Problem), un enjeu central en analyse spatiale (Hallot et al., 2019). En effet, le MAUP souligne que les résultats des études peuvent varier en fonction de la taille et de la forme des unités spatiales utilisées, ce qui peut fausser les analyses. En adoptant des grilles standardisées, on minimise ces variations.

Une fois le modèle de forêt aléatoire établi, après une sélection rigoureuse des variables, il permet d'estimer la relation entre le logarithme de la densité de population et les variables explicatives, à l'échelle de la zone de dénombrement. Cette relation est ensuite utilisée pour attribuer des poids de redistribution à des unités plus fines (100 mètres par pixel), permettant ainsi une désagrégation dasymétrique. Idéalement, ces poids reflètent les mécanismes sous-jacents à la répartition inégale de la population sur le territoire (Stevens et al., 2015) .

En conclusion, les cartes de population créées avec cette méthode affichent des valeurs d'erreur plus faibles, attestant d'une meilleure correspondance avec les comptages réels du recensement. Comparées aux cartes produites par les projets AfriPop, AsiaPop ou d'autres approches antérieures comme GPW et GRUMP (détaillés dans la partie 3 de ce chapitre), elles s'avèrent donc, en moyenne, plus précises. Cette méthode, basée sur des apprentissages statistiques par forêts aléatoires, a été confirmée comme la référence dans le domaine et a été utilisée pour créer de nombreux jeux de données, notamment le jeu de données intitulé « Individual countries 2000-2020 UN adjusted » [14] (Lloyd et al., 2019), où les unités représentent le nombre de personnes par pixel, les totaux des pays étant ajustés pour correspondre aux estimations officielles de la population fournies par les Nations Unies. De plus, cette méthode a été appliquée dans des recherches récentes (Darin et al., 2022), confirmant ainsi son efficacité et son importance dans le domaine de l'estimation de la population.Depuis, de nombreux autres projets de cartographie de la population ont vu le jour ; nous en dressons l'inventaire dans la partie suivante. Ces projets utilisent des méthodes similaires et produisent des données carroyées, avec une résolution de 100x100 mètres au mieux, sinon moindre.

Une conclusion fréquemment observée dans les recherches sur la redistribution dasymétrique de la population est que l'un des défis majeurs réside dans la fiabilité des données démographiques d'entrée. Wardrop et al. (2018) soulignent

[14] https://hub.worldpop.org/geodata/listing?id=69

que, au-delà de la sélection des données auxiliaires nécessaires à la construction du modèle, le véritable obstacle est d'assurer l'exactitude et la fiabilité des données démographiques utilisées, particulièrement dans les pays à faibles et moyens revenus. Cela est d'autant plus complexe dans le contexte d'une croissance démographique rapide, où maintenir une continuité des données démographiques entre les recensements est un défi considérable. Ils mettent en évidence que les estimations de population maillées issues d'une désagrégation descendante sont intrinsèquement liées à la qualité des données de recensement sur lesquelles elles s'appuient.

Cela met également en lumière une autre limite majeure : l'utilisation de ces approches lors des périodes intercensitaires. En effet, en l'absence de recensements disponibles, il devient nécessaire de projeter la population. Bien que les projections de population, lorsqu'elles prennent en compte avec rigueur les facteurs de natalité, de mortalité et de migration, puissent fournir des estimations globalement fiables à court terme (Lutz, Sanderson, Scherbov, 2001), nous l'avons vu en introduction, leur fiabilité tend à diminuer avec le temps et au fur et à mesure des années. Il y a aussi la question de la mobilité spatiale, difficile à prendre en compte à l'échelle des unités spatiales les plus fines disponibles, telles que les communes, les villages ou les quartiers, et qui pourtant conditionne fortement les effectifs de population. De plus, la majorité des projets et des méthodes rencontrent des contraintes de temps et d'accès aux données, ce qui rend difficile l'obtention d'informations au niveau administratif le plus fin disponible, rendant ainsi ces projections précises quasi impossibles. Dans ces situations, il devient nécessaire de projeter la population en se basant sur des taux d'accroissement (Stevens et al., 2015), mais cette approche risque de masquer les dynamiques démographiques non uniformes qui existent entre les zones administratives.

Enfin, l'étape de validation des estimations constitue sans doute l'une des principales limites des méthodes descendantes, et plus largement, de l'estimation de population à partir d'images satellites. Dans un premier temps, il est possible de comparer les cartes de population à d'autres jeux de données et projets démographiques, comme le font Stevens et al. (2015). Cependant, ces comparaisons ne servent généralement qu'à établir qu'une méthode est meilleure qu'une autre, ou à évaluer l'efficacité relative de la valeur prédictive de chaque méthodologie à l'aide de critères tels que le RMSE[15] (Root Mean Squared Error). Elles ne permettent pas

[15] Le RMSE est l'erreur quadratique moyenne, qui est une mesure statistique utilisée pour évaluer la précision des modèles de prédiction. Il s'agit de la racine carrée de la moyenne des carrés des écarts entre les valeurs prédites et les valeurs réelles. Plus un RMSE est faible, plus les erreurs entre les prédictions et les valeurs réelles sont plus faibles en moyenne ce qui indique une meilleure précision du modèle.

d'affirmer la validité des estimations à une échelle très précise, celle des pixels de 100 x 100 mètres. Une autre méthode classique consiste à désagréger la population à partir d'un niveau administratif moins détaillé. Par exemple, au Kenya, plutôt que d'utiliser les *sub-locations* (niveau 5), les chercheurs se basent sur les *locations* (niveau 4). Ensuite, ces estimations sont agrégées au niveau des *sub-locations*, et une comparaison est effectuée pour calculer le pourcentage d'erreur relative afin de déterminer si la population de la zone administrative a été sous-estimée ou surestimée. Bien que cette approche permette d'évaluer si la population est correctement répartie, elle ne valide pas véritablement les estimations à un niveau aussi fin que souhaité. En d'autres termes, la validation est limitée par l'absence de décomptes de la population à une échelle de 100x100 mètres. Ainsi, il n'existe pas de jeux de données de validation précis, ce qui complique l'estimation du pourcentage d'erreur réel des estimations produites par les méthodes descendantes.

En conclusion, les méthodes descendantes ont considérablement amélioré la cartographie de la population, offrant une solution efficace et peu coûteuse pour produire des cartes couvrant un grand nombre de pays. Cependant, ces approches ne sont pas toujours adaptées à tous les contextes, en particulier lorsqu'un recensement récent n'est pas disponible. Face à ces défis, les années 2010 ont marqué le début d'une littérature émergente, encore limitée mais prometteuse, consacrée aux méthodes ascendantes.

2. Méthodes ascendantes

Les méthodes ascendantes, plus souvent appelées « bottom-up » à l'international, s'appuient sur un modèle prédictif, établi par exemple à partir de micro-recensements locaux, capable d'estimer la population dans des zones où aucun recensement n'a été réalisé (Neal et al., 2022). Elles répondent à un besoin concret lorsque les recensements nationaux ne sont pas disponibles ou sont trop anciens, comme en République Démocratique du Congo, où le dernier recensement remonte à 1984, ou à Madagascar, où jusqu'en 2018, le dernier recensement remontait à 1993. Ces méthodes sont qualifiées d'ascendantes car elles partent de données fines (micro-recensements, caractéristiques du terrain) pour remonter à des estimations plus larges. Même lorsque le résultat final est une grille uniforme – comme celles de 100x100mètres utilisée par WorldPop –, le processus reste ascendant : il consiste à transformer les caractéristiques locales en estimations à plus grande échelle. En agrégeant les estimations des petites zones, ces méthodes permettent également de fournir des résultats à l'échelle des unités administratives ou au niveau national (Wardrop et al., 2018).

Afin de mettre en œuvre les approches ascendantes, il est nécessaire d'accéder à des décomptes de population à un niveau beaucoup plus détaillé que celui des unités administratives, souvent très larges, utilisées lors de la désagrégation. Il s'agit ici de données provenant d'enquêtes et de recensements, et la majorité du temps, il s'agit d'enquêtes de micro-recensement réalisées dans des zones définies à l'avance. Ces zones représentent une portion infime d'un recensement national (Wardrop et al., 2018), mais elles sont sélectionnées pour représenter la diversité du territoire, notamment en termes géographiques, socio-économiques et démographiques.

L'objectif est ensuite d'associer des données de population extrêmement détaillées, typiquement issues de micro-recensements, avec un large ensemble de variables (similaires à celles mentionnées dans la partie précédente), afin d'identifier des liens pertinents entre ces variables et la population. Les avancées récentes en systèmes d'information géographique (SIG) ont facilité la création et l'harmonisation de nombreuses variables à l'échelle des cellules de grille, permettant ainsi une meilleure compréhension des disparités socio-économiques, démographiques et géographiques (Darin et al., 2022). Ces innovations permettent de distinguer des zones bâties qui, à première vue, paraissent similaires, mais qui peuvent en réalité présenter des caractéristiques démographiques distinctes, comme un nombre moyen de personnes par ménage différent (OCHA, 2020). Une fois ces relations établies, elles peuvent être appliquées pour estimer la population dans d'autres zones qui n'ont pas été couvertes par des micro-recensements.

En comparaison avec les méthodes descendantes, seules quelques méthodes ascendantes ont été développées, malgré une accélération notable ces dernières années. Ces méthodes ont principalement été conçues pour cartographier la population dans des contextes où les données sont limitées. Par exemple, Checchi, Stewart, Palmer et Grundy (2013) ont combiné des estimations de densité provenant de différentes sources (littérature, internet) avec des comptages manuels du nombre de structures résidentielle visibles à partir d'images satellites pour évaluer le nombre de personnes déplacées dans divers sites en Asie et en Afrique. De même, Hillson et al. (2014) ont mené des enquêtes à Bo, en Sierra Leone, pour recueillir des données démographiques. En parallèle, ils ont réalisé des décomptes de bâtiments par interprétation manuelle des images. Ils ont constaté qu'un modèle basé sur l'occupation (personnes par bâtiment) était plus précis qu'un modèle fondé sur la superficie des toits. Stewart et al. (2016) ont estimé l’occupation des bâtiments par les habitants, en utilisant des images satellites et des images au sol pour identifier les structures.

Plus récemment, deux études au Nigéria (Leasure, Jochem, Weber, Seaman, Tatem, 2020; Weber et al., 2018) ont réussi à estimer la population indépendamment des données de recensement, notamment grâce à un modèle bayésien hiérarchique qui permet d'obtenir des estimations fiables de la population et de l'incertitude à n'importe quelle échelle spatiale. Cette méthode est maintenant utilisée par WorldPop lorsqu'il s'agit d'appliquer une approche ascendante à une région ou un pays.

Encore plus récemment, Metzger et al. (2022) ont développé une méthode appelée POMELO, qui permet non seulement de désagréger la population, mais également de la prédire dans des pays où les chiffres de recensement ne sont pas disponibles. Les estimations sans contraintes, c'est-à-dire sans données de recensement pour le pays cible, sont réalisées à une résolution spatiale de 100 mètres. Autrement dit, comme le programme WorldPop, le modèle prédit la population pour chaque cellule d'une grille de 100 mètres par 100 mètres couvrant la région d'intérêt. Bien que les résultats finaux ne soient pas aussi précis que ceux obtenus par désagrégation, cette approche constitue un cas rare où une méthode ascendante peut être appliquée sans nécessiter de données de micro-recensement préexistantes, ce qui représente un atout majeur. Pour ce faire, ils ont utilisé des données de sept pays différents, puis entraîné tour à tour par validation croisée le modèle sur six de ces pays afin de tester sa capacité à prédire la population dans le septième pays. Ensuite, ils ont comparé les résultats à d'autres méthodes et ont obtenu des résultats supérieurs dans les trois pays testés. Cependant, les performances sont moins précises que lorsque des données de recensement sont disponibles pour le pays cible. Bien que les résultats soient prometteurs et satisfaisants, les valeurs du coefficient de détermination R^2 (mesure statistique permettant d'évaluer la performance d'un modèle) sont encore bien inférieures à celles de la désagrégation (de 48-69 % pour l'estimation sans contrainte à 85-89 % pour la désagrégation), montrant malgré l'abondance des données d'entrainement qu'il reste des améliorations à apporter à ce type d'approche.

Le développement relativement limité des méthodes ascendantes s'explique en grande partie par la nécessité de réaliser un micro-recensement en amont, ce qui demande souvent des ressources considérables en termes de temps et de financement. De plus, pour que ces approches soient efficaces à l'échelle nationale, il est crucial que l'échantillon soit suffisamment représentatif ; ce qui implique de capturer les caractéristiques socio-économiques et environnementales dans leur diversité.

Comme pour les méthodes descendantes, la validation reste l'étape la plus délicate et la moins fréquente. En l'absence de données aussi détaillées que celles nécessaires pour estimer la population, il est difficile de valider les résultats de manière rigoureuse. Ce qui est généralement pratiqué, c'est une validation statistique.

Par exemple, Wardrop et al. (2018) et plus récemment Darin et al. (2022) ont réalisé une validation croisée en utilisant 70 % des zones de dénombrement pour l'entraînement du modèle et 30 % pour le test. Bien que cette approche permette de tester la robustesse du modèle et de généraliser les prédictions à d'autres zones, elle ne constitue pas une validation directe de qualité des estimations car on ne compare pas à des valeurs réelles de population. Cette limitation est l'un des obstacles majeurs à l'estimation de population à partir d'images satellites, que ce soit via des méthodes descendantes ou ascendantes.

Une limite notable des techniques ascendantes est leur application restreinte. À l'exception du projet POMELO, récent mais imparfait comme nous l'avons constaté, aucun autre projet n'a permis une estimation fiable de la population à l'échelle des pixels de 100 x 100 mètres, en particulier dans les pays africains. Les micro-recensements étant souvent limités à des zones précises, ces méthodes restent difficilement généralisables à une majorité de pays. Il manque encore un projet capable de généraliser un modèle ascendant à divers contextes nationaux.

Afin de conclure, je souhaite citer Neal et al. (2022) qui énoncent que « les approches dépendantes et indépendantes du recensement ont chacune leurs avantages et inconvénients ». L'estimation de population basée sur les recensements présente l'avantage d'être plus économique, car elle exploite des données déjà disponibles. De plus, elle peut être fait de manière quasiment automatique dans un large éventail de pays. Cependant, elle peut entraîner des erreurs si les projections intercensitaires s'écartent de la réalité, ce qui se produit inévitablement à mesure que l'on s'éloigne de la date du recensement initial. À l'inverse, les approches indépendantes des recensements reposent sur des micro-recensements, qui permettent de recueillir des données détaillées à une échelle plus fine que les zones administratives classiques. Bien que la collecte de ces données à grande échelle soit plus coûteuse, elles offrent une information de référence précise qui n'est pas disponible avec les méthodes dépendantes des recensements.

Dans la suite, nous passons en revue les différents projets de cartographie existants et les différents fournisseurs de cartographie de bâti par images satellitaires.

3. Révolution cartographique : les projets de mise à disposition de données de population géoréférencées

Selon POPGRID[16], une initiative collaborative visant à regrouper et discuter les données géoréférencées en accès libre sur la population, il existe six principaux ensembles de données de population maillée à l'échelle mondiale :

- Gridded Population of the World (GPW)[17],
- Global Rural Urban Mapping Project (GRUMP)[18],
- Global Human Settlement Layer (GHSL)[19],
- LandScan[20],
- WorldPop[21],
- High Resolution Settlement Layer (HRSL).

Ces fichiers offrent une couverture mondiale et ont la particularité de tous utiliser des méthodes descendantes pour répartir la population dans des cellules de grille (TReNDS, 2020). Chaque projet se distingue par son accessibilité, sa résolution ou sa couverture temporelle.

Le GPW et le GRUMP utilisent les recensements nationaux de la population. Le GPWv4, développé par le CIESIN (Center for International Earth Science Information Network) est basé sur les données de recensements réalisés entre 2005 et 2014, qui sont projetées pour produire des estimations de population pour les années 2000, 2005, 2010, 2015 et 2020. Le GRUMP, bien qu'ayant une résolution plus grossière, a été pionnier dans la modélisation des populations urbaines et rurales. Il a apporté des innovations en utilisant des fichiers auxiliaires sur le logement et les implantations humaines, mais a été dépassé par des outils plus récents. Bien que comme WorldPop, le GPW propose des estimations avec une caractérisation en âge et en sexe, le GPW et le GRUMP ont désormais des résolutions trop faibles (1km^2) et sont donc moins avantageuses que les suivantes.

Les projets GHS-POP et HRSL utilisent la cartographie des surfaces bâties, détectées par observation satellitaire, pour estimer la répartition de la population à

[16] https://www.popgrid.org/
[17] https://sedac.ciesin.columbia.edu/data/collection/gpw-v4
[18] https://sedac.ciesin.columbia.edu/data/collection/grump-v1
[19] https://human-settlement.emergency.copernicus.eu/
[20] https://landscan.ornl.gov/
[21] https://www.worldpop.org/

partir des données de recensement. Le GHS-POP[22], développé par le *Joint Research Centre* (JRC) de la Commission européenne, met en avant une approche systématique qui permet des analyses temporelles cohérentes à travers les années. Plus précisément, la base de données fournit des estimations à des intervalles de 5 ans entre 1975 et 2020, avec des projections jusqu'en 2030. La meilleure résolution est de 100 mètres et utilise la détection de bâti de leur autre projet nommé GHS-BUILT-H[23] avec notamment une proposition d'estimation de la hauteur des bâtiments (Pesaresi, Corbane, Ren, Edward, 2021).

High Resolution Settlement Layer (HRSL) était le fruit d'une initiative conjointe de Facebook, du CIESIN et de la Banque mondiale sortie en 2017. Depuis Facebook est devenu Meta et a créé, toujours en collaboration avec le CIESIN, des cartes de population[24] afin d'estimer la densité de population avec une résolution de 30 mètres[25] avec comme unique année disponible l'année 2020.

LandScan, de son côté, intègre diverses données géospatiales pour fournir des estimations démographiques prenant en compte la répartition des populations selon les moments de la journée, ce qui en fait un outil adapté aux analyses de risques et aux réponses en cas de catastrophe. LandScan donne des estimations jusqu'en 2022 mais lui aussi à une résolution de $1km^2$, ce qui en fait une solution moins efficace que le programme WorldPop, dont la résolution est dix fois plus précise.

Enfin, WorldPop se distingue par l'utilisation d'algorithmes d'apprentissage automatique pour redistribuer les populations ainsi que par leur caractérisation en âge et en sexe. La transparence et l'accessibilité des codes en plus de la disponibilité des estimations pour les années 2000 à 2020 en font une ressource adaptée autant pour des études longitudinales que pour une année précise.

Tous ces projets utilisent des techniques descendantes pour estimer et répartir la population, dont la limite principale est la nécessité de projeter la population dans le temps pour estimer les années pour lesquelles il n'y a pas de recensement. De plus, hormis LandScan donnant des estimations pour 2021 et 2022, la majorité des projets s'arrêtent à 2020, comme c'est le cas de WorldPop et GHS-POP. Comme entrevu précédemment, il n'existe pas actuellement de base de données ouverte permettant l'estimation de population par méthode ascendante. C'est pour le moment le fruit de méthodes indépendantes sur certains pays, ou des estimations ascendantes produites

[22] https://human-settlement.emergency.copernicus.eu/ghs_pop2023.php
[23] https://human-settlement.emergency.copernicus.eu/ghs_buH2023.php
[24] https://dataforgood.facebook.com/dfg/tools/high-resolution-population-density-maps
[25] https://dataforgood.facebook.com/dfg/docs/methodology-high-resolution-population-density-maps

par WorldPop et Grid3[26] mais qui ne concernent qu'une dizaine de pays (Thomson, Rhoda, Tatem, Castro, 2020) ou encore du projet récent POMELO.

Par ailleurs, un des moteurs permettant d'améliorer l'estimation de la population est l'utilisation croissante et l'accessibilité accrue des données satellitaires à un degré de précision élevé, car cela permet une cartographie de plus en plus précise des bâtiments. La disponibilité de ces données a considérablement transformé les méthodes d'estimations démographique, notamment au niveau des méthodes ascendantes. Plusieurs fournisseurs d'imagerie satellitaire offrent des solutions variées que nous détaillons dans le paragraphe suivant.

4. Fournisseurs de données sur le bâti : diversité et spécificités

De nombreux fournisseurs produisent des données sur le bâti, chacun se distinguant par la fréquence de publication et la résolution des données fournies.

- Open Street Map[27] :

OpenStreetMap (OSM) est un projet collaboratif qui vise à constituer une base de données géographiques et permettre de créer des cartes libres. En tant que projet collaboratif, la précision des données varie considérablement d'une région à l'autre (Haklay, 2010) et certaines informations extraites d'images à très haute résolution (Very High Resolution soit VHR, paragraphe suivant) sont déjà intégrées dans OSM. Les données géographiques fournies par les bénévoles (ou volunteered geographic information soit VGI) sont précieuses car elles permettent d'obtenir des informations sur les empreintes des bâtiments, les emplacements des routes ou des installations (Munoz, Srivastava, Tuia, Falcao, 2021; Wardrop et al., 2018). Elles peuvent également renseigner sur la nature du bâtiment, à savoir si ce dernier est un hôpital, une ambassade ou un cimetière par exemple, données particulièrement précieuses lorsque l'on souhaite estimer la population. En revanche, les informations OSM reposant sur des images satellites non datées, de périodes variables, elles ne peuvent être datées, l'idée étant simplement d'avoir les données les plus récentes possibles à partir des images satellites de haute résolution libres d'accès.

[26] https://data.grid3.org/

[27] https://www.openstreetmap.org/#map=8/-13.689/48.994

- Google[28], Microsoft[29] :

Google et Microsoft fournissent des données très précises dans lesquelles les bâtiments sont identifiés grâce à des images VHR d'une résolution de 1m (voire de 0.5m) mais ces données ne peuvent pas être datées en raison de la couverture nuageuse dans les régions tropicales et équatoriales (certains pixels d'entrée ont été acquis il y a plus de 5 ans) et du petit nombre d'images d'entrée disponibles en Afrique. Ainsi, bien que la base de données ait été publiée en 2022, elle correspond à une mosaïque d'images datant des années précédentes, sans qu'il soit possible de dater précisément chacune des images. Le projet Google Open Buildings[30] permet de télécharger la cartographie du bâti dans un format shapefile adapté aux systèmes d'informations géographiques.

En septembre 2024, Google a développé un nouveau projet « Open Buildings 2.5D Temporal Dataset[31] », qui vise à résoudre le problème de mosaïque de données en utilisant des images Sentinel à moindre résolution, mais permettant de correctement dater la cartographie du bâti. La base de données couvre les années 2016 à 2023, pour la majorité des pays africains avec une résolution effective de 4 mètres.

- Ecopia[32]

Ecopia utilise également des images satellites à très haute résolution (VHR) produites par Maxar[33] (30 à 50 cm de résolution) pour les années 2005 à 2020 afin de produire des données de bâti principalement pour les pays d'Afrique subsaharienne. Les données sont sous une licence commerciale personnalisée qui restreint principalement leur utilisation à des applications humanitaires (Chamberlain et al., 2024).

- DLR[34] (Agence Spatiale Allemande) :

Les images proviennent de satellites radar SAR et couvrent l'ensemble du globe à une résolution de 10 mètres. Les dernières données ont été produites en 2019 et aucune annonce de mise à jour n'a été faite.

- GHSL[35] (Global Human Settlement Layer) :

[28] https://sites.research.google/open-buildings/
[29] https://planetarycomputer.microsoft.com/dataset/ms-buildings
[30] https://sites.research.google/gr/open-buildings/
[31] https://sites.research.google/gr/open-buildings/temporal
[32] https://www.ecopiatech.com/products/buildingbased-geocoding
[33] https://www.maxar.com/maxar-intelligence/products/satellite-imagery
[34] https://www.dlr.de/en/eoc/research-transfer/projects-missions/world-settlement-footprint-wsf-r
[35] https://human-settlement.emergency.copernicus.eu/

Développé par le *Joint Research Centre* (JRC), le GHSL utilise des images Sentinel-2 pour identifier les établissements humains grâce à un modèle de réseau neuronal artificiel entraîné par zone UTM. Les dernières données disponibles remontent à 2018 et 2022.

- *Meta*[36] *(anciennement Facebook)*

Meta a développé un projet qui vise à créer une carte détaillée des bâtiments à l'échelle mondiale à partir de données satellitaires à haute résolution. Meta a notamment collaboré avec le CIESIN pour utiliser ces données dans des contextes humanitaires et de développement.

Le tableau 1 récapitule tous ces projets de cartographie, et l'article de Chamberlain et al. (2024) qui compare les données sur les empreintes de bâtiments pour les pays d'Afrique apporte des informations supplémentaires.

Tableau 1 : Récapitulatif des projets de cartographie de bâti

Fournisseur	Résolution	Couverture terrestre	Plage temporelle
OSM		Mondiale et dépend de la communauté	Dépend de la communauté et des images sources mais mis à jour
Google	Entre 0.5 et 1 mètre	Afrique Asie du sud	2022
Microsoft	Entre 0.5 et 1 mètre	Principalement Afrique subsaharienne	2014-2022
Ecopia	Entre 0.3 et 0.5 mètres	Principalement Afrique subsaharienne	2005-2020
DLR	10 mètres	Mondiale	2019
GHSL	10 mètres	Mondiale	2018 2022
Meta	30 mètres	Amérique du sud et Afrique	2020

Ces divers fournisseurs de jeux de données illustrent la possibilité de cartographier le bâti directement, y compris à des résolutions très fines, comme le démontre la base de données Google Open Buildings. Cependant, il existe peu de recherches explorant l'utilisation de la détection de bâti à très haute résolution, et encore moins d'études cherchant à estimer la population à l'intérieur des zones bâties détectées. Dans ce qui suit, nous présentons un exemple pertinent de cette démarche : Africapolis.

[36] https://dataforgood.facebook.com/dfg/tools/buildings

5. Africapolis et les agglomérations : cartographier l'urbanisation africaine

En Afrique, l'Organisation de coopération et de développement économiques (OCDE), en partenariat avec le Club du Sahel et de l'Afrique de l'Ouest (CSAO) et en collaboration avec E-Geopolis[37], soutient la transition urbaine grâce à une base de données unique nommée Africapolis. Cette initiative vise à garantir la comparabilité entre les agglomérations à l'échelle africaine, en se concentrant sur les villes de plus de 10 000 habitants. En s'appuyant sur des recensements nationaux et des images satellites, le programme Africapolis a constitué, pour l'année 2015, une base de données complète et homogène sur les villes et les dynamiques d'urbanisation en Afrique. Ce programme met l'accent sur l'urbain en cartographiant toutes les villes dépassant le seuil de 10 000 habitants.

Le programme Africapolis[38] se distingue parmi les rares initiatives qui proposent une approche alternative à la cartographie de la population, différente des méthodes maillées précédemment présentées. Ce programme se concentre sur la cartographie des agglomérations, en définissant des zones bâties continues. Africapolis utilise la même définition physique de l'agglomération que l'Insee, « les unités urbaines sont définies en France métropolitaine et dans les DOM comme une commune ou un ensemble de communes présentant une zone de bâti continu (sans coupure de plus de 200 mètres entre deux constructions) comptant au moins 2 000 habitants ». Ces seuils, tant pour la continuité du bâti que pour la population, reposent sur des recommandations adoptées au niveau international (Insee, 2015).

Pour estimer la population des agglomérations, le programme Africapolis s'appuie sur les données de recensement de chaque pays – des données qui peuvent varier considérablement d'un pays à l'autre – ainsi que sur des « registres électoraux et d'autres sources officielles ». Les images satellites viennent ensuite enrichir ces documents en identifiant les « zones bâties et la localisation précise des foyers de peuplement ».

Africapolis postule qu'il n'existe pas de définition universelle ou largement acceptée de la ville ou de l'urbain, tout en soulignant l'importance cruciale d'une définition harmonisée de l'urbain. Cela faciliterait l'élaboration de politiques publiques de développement adaptées à la réalité du terrain. Selon (OCDE, Club du

[37] https://e-geopolis.org/
[38] https://africapolis.org/fr

Sahel et de l'Afrique de l'Ouest, 2020), les définitions communément reconnues du phénomène urbain peuvent être classées en trois catégories : villes, agglomérations et régions. Les limites d'une ville sont souvent déterminées par l'État, tandis qu'une agglomération se définit comme un ensemble de constructions denses, dont la densité peut être mesurée par le nombre d'habitants, par unité de surface, ou par la distance maximale séparant les constructions (OCDE, Club du Sahel et de l'Afrique de l'Ouest, 2020). En Afrique, la taille minimale d'une agglomération urbaine est souvent fixée de manière variable et arbitraire selon les pays. Africapolis illustre ce point avec l'exemple du Kenya, dont la définition officielle de la zone urbaine a changé entre deux recensements, compliquant ainsi la comparaison. En effet, certaines localités peuvent être considérées comme urbaines lors d'une enquête démographique, puis redevenir rurales lors de la suivante.

En conclusion, le programme Africapolis plaide en faveur d'une homogénéisation des concepts urbains et des zones urbaines en général, proposant une définition harmonisée. Pour permettre des analyses comparatives à long terme des dynamiques d'urbanisation en Afrique, Africapolis a introduit un concept d'agglomération basé sur deux critères. Le premier est physique : deux zones bâties sont considérées comme agglomérées si la distance les séparant est inférieure à 200 mètres. Le second est démographique : toutes les zones bâties regroupées sont considérées comme une agglomération urbaine si leur population totale dépasse 10 000 habitants. Selon Africapolis, cette visualisation présente l'avantage de montrer que « contrairement aux villes en général, dont les limites sont fixes, les agglomérations, telles que définies ici, sont des unités dont la forme, le contenu et les limites évoluent dans le temps en fonction de l'évolution et de l'environnement bâti ».

Bien que le programme Africapolis présente de nombreux atouts, certaines limites méritent d'être soulignées. Tout d'abord, il se concentre exclusivement sur l'urbain, négligeant complètement les zones rurales. Cette approche est compréhensible, étant donné que la population tend à se concentrer majoritairement dans les zones urbaines avec près de 2 milliards et demi d'Africains prévus en 2050, dont la majorité dans les villes (Magrin et al., 2022), mais elle laisse néanmoins de vastes segments de la population non cartographiés. Au même titre que WorldPop, la dernière mise à jour a eu lieu pour tous les pays en 2020 et a lieu tous les cinq ans. Enfin, le programme ne propose pas de modélisation pour estimer la population, et son fonctionnement d'attribution de population manque parfois de transparence. Cependant, Africapolis compense ces lacunes par une définition harmonisée de l'urbain et la mise à disposition de nombreux rapports et fiches par pays, fournissant ainsi des informations actualisées et fiables sur 51 pays d'Afrique, notamment concernant les découpages territoriaux et divers indicateurs socio-économiques.

6. Conclusion

Ce chapitre a exploré les méthodes et initiatives clés en matière d'estimation démographique géoréférencée et de cartographie de la population, en mettant en lumière les approches méthodologiques descendantes et ascendantes. Les méthodes descendantes, bien que largement développées, présentent des limites, notamment dans les régions où les données de recensement sont obsolètes ou inexistantes. Les méthodes ascendantes, bien que prometteuses, sont encore en développement et nécessitent souvent des ressources considérables pour la collecte de données de micro-recensement.

Les avancées récentes en imagerie satellitaire et en traitement des données géospatiales ont permis d'atteindre des résolutions de plus en plus élevées, facilitant ainsi la cartographie précise des bâtiments. Cependant, malgré ces progrès, aucun programme n'a encore pleinement exploité ces capacités pour estimer la population directement à l'intérieur de bâtiments ou de groupements de bâtiments.

C'est dans ce contexte que le programme TeleCense se positionne, en s'inspirant d'initiatives comme Africapolis pour la cartographie du bâti. A partir des données Sentinel et en mobilisant les approches descendantes et ascendantes nous souhaitons développer une nouvelle méthode d'estimation de la population, directement à l'intérieur des groupements de bâtiments de TeleCense. Ces zones bâties sont nommées des shapes et sont présentées dans le chapitre suivant.

Chapitre 2 – Présentation du programme TeleCense et enjeux méthodologiques

Le programme TeleCense est développé par l'entreprise Diginove, créée à la fin de l'année 2016 et spécialisée dans le traitement d'images. TeleCense vise à détecter et caractériser les bâtiments tous les six mois, depuis 2017 jusqu'à aujourd'hui, sur différentes zones géographiques, principalement situées en Afrique. Provenant de l'observation de la Terre depuis les satellites Sentinel, les données générées sont précises et datées. Ces années d'étude permettent de suivre l'étalement urbain et l'accroissement géolocalisé de la population, et peuvent servir de base à des projections. Contrairement à la plupart des projets existants, TeleCense cherche à estimer la population directement à l'intérieur des bâtiments, fournissant ainsi des estimations de population résidentielle plutôt que des densités de population basées sur des cellules de grille.

L'objectif de ce chapitre est de présenter le fonctionnement du programme TeleCense, en expliquant comment, à partir de simples images satellites, nous parvenons à identifier les zones bâties et à leur associer de nombreuses caractéristiques. Ensuite, nous présenterons les méthodes d'estimation de la population appliquées à cette identification, qui seront développées dans les chapitres 3 à 6.

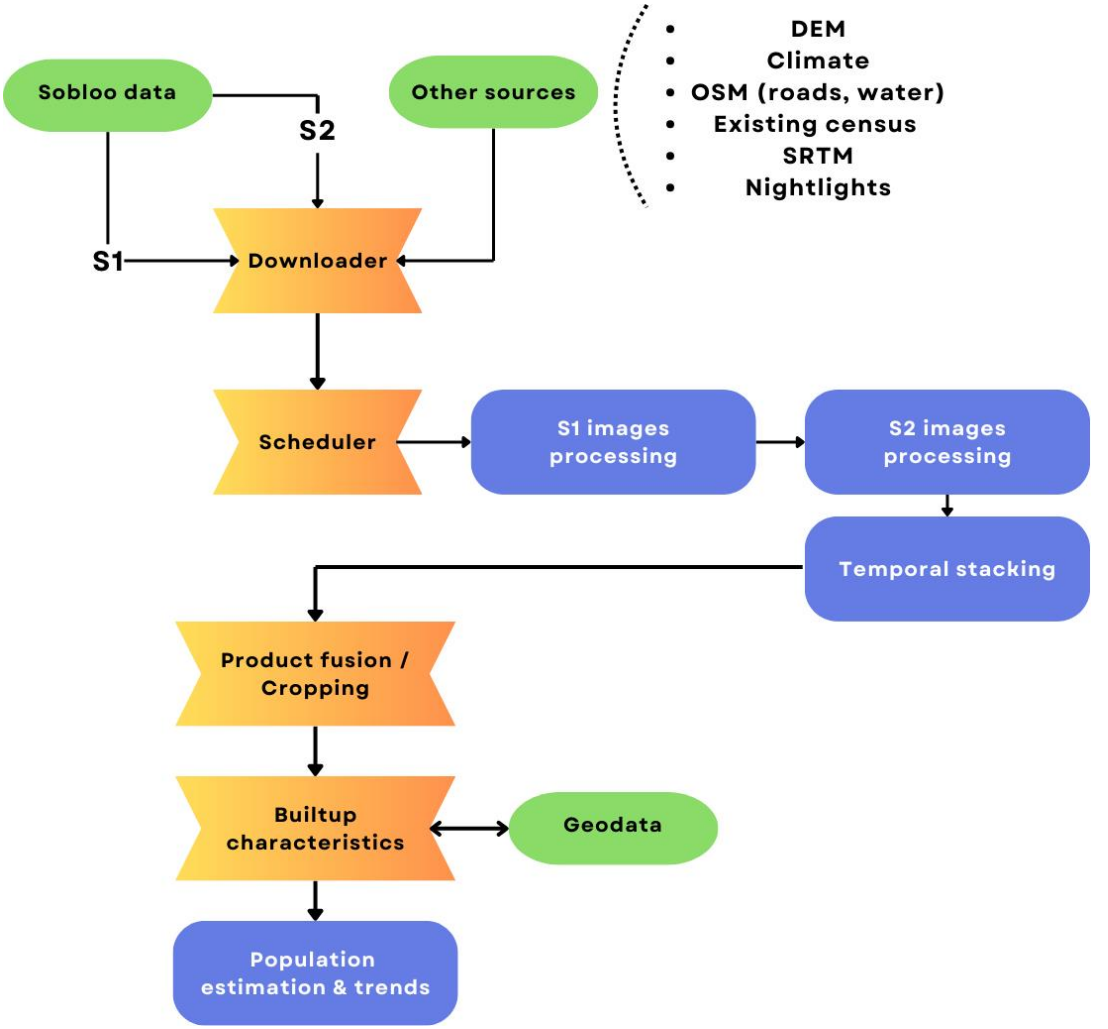
1. Identification du bâti : première étape du programme TeleCense

Le programme TeleCense repose sur de nombreux sous-systèmes, dont l'architecture est synthétisée dans la figure 8. Il fonctionne dans l'ordre suivant :

1. Le « Downloader » est chargé d'accéder aux différents sites permettant de télécharger les images et les données brutes nécessaires au processus. Il s'agit dans un premier temps de récupérer les images Sentinel 1&2 (radar et optique) de la plateforme de données Sobloo du projet Copernicus. Puis, celles du modèle numérique de terrain (Shuttle Radar Topography Mission, SRTM), de lumières de nuit (VIIRS), certaines données d'Open Street Map (OSM) et l'empreinte du bâti d'autres fournisseurs (Google, Microsoft, DLR etc…)
2. Le « Scheduler » gère la chaîne des traitements et leur parallélisation, pour une utilisation optimale de la plateforme.

3. Le processus S1 (radar) vise à rendre les images radar plus exploitables pour extraire l'emprise et la densité des établissements. Il consiste en la radiométrie et l'interférométrie, en combinant les résultats en une image multibande qui correspond à la tuile S2 (origine, taille, projection). Le fichier précis de l'orbite est d'abord appliqué, suivi de la calibration, du filtrage du speckle (qui est le processus de réduction du bruit granulaire causé par les motifs d'interférence dans les images radar), de l'aplatissement du terrain, de l'empilement temporel et de l'enregistrement, de la correction du terrain et du recadrage. Enfin, la divergence du speckle est estimée pour identifier les établissements.
4. Le processus S2 (optique) extrait des images radar et optiques traitées les empreintes des établissements, de la végétation, de l'eau et du sol nu. Il combine les résultats du processus S1 et l'analyse des bandes multispectrales de S2. Les images sont d'abord corrigées des perturbations atmosphériques et normalisées en fonction de leur localisation. Des indices radiométriques sont calculés et combinés pour extraire la végétation, l'eau et le sol nu. La divergence est calculée pour chaque pixel et les images radiométriques, les indices et la divergence sont combinés pour identifier les établissements.
5. L'empilement temporel et la combinaison d'images en saison sèche et humide permet de se débarrasser du bruit autour des rivières et des routes. Pour cela au moins 6 mois sont utilisés.
6. La fusion des données permet de créer des fichiers de bâti utilisables.
7. Les modèles démographiques sont utilisés pour répartir ou estimer la population, ce qui constitue le travail présenté dans la thèse.

Figure 8 : Architecture du programme TeleCense



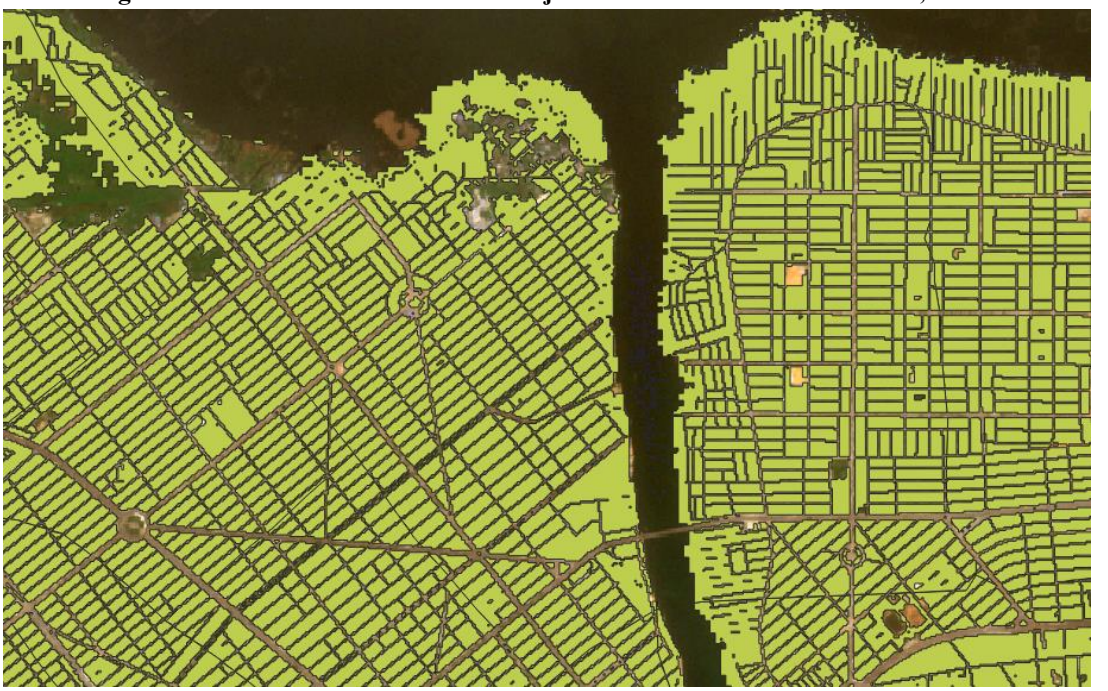
Sources : TeleCense

Lors de l'étape numéro 6, le résultat est vectorisé et comparé à d'autres bases de données disponibles (celles présentées dans la section 3 précédente) afin d'améliorer la précision de la détection. Les figures 9 et 10 illustrent le résultat final de ce processus. La première image présente le nord de Cotonou vu par une image Sentinel-2 alors que la deuxième image montre le résultat du passage des algorithme TeleCense qui permettent d'identifier le bâti.

Figure 9 : Nord de Cotonou (Bénin) vu de l'espace par Sentinel 2 – 2023

Sources : TeleCense, image issue du logiciel QGIS

Figure 10 : Identification du bâti au 1er janvier 2023 au nord de Cotonou, Bénin



Sources : TeleCense, image issue du logiciel QGIS

A noter que contrairement à l'image fournie en introduction générale (la figure 5), la figure 9 peut sembler de moins bonne qualité, ou du moins les bâtiments sont moins visibles à l'œil nu. En effet, c'est le cas ! Les images utilisées ici proviennent du projet Copernicus, et ce sont ces images que le programme TeleCense utilise pour cartographier les bâtiments. Elles ont une résolution de 10 mètres, c'est pour cela qu'il est difficile à l'œil nu de correctement identifier tous les bâtiments.

Une fois le bâti détecté il peut être caractérisée selon deux échelles :

1) Au niveau de l'empreinte du bâti :

C'est ce qu'on appelle une « shape », et qui n'est pas tout à fait équivalent à un bâtiment. En effet, comme nous utilisons des images à une résolution de 10 mètres, une shape est en fait un groupement de bâtiments, de plus ou moins grande taille. Un travail important est par ailleurs fourni pour avoir la plus grande homogénéité dans la taille des shapes, ce qui facilite les travaux d'adaptation à divers contextes nationaux.

2) Au niveau de l'agglomération :

Une agglomération est composée de nombreuses shapes situées à moins de 200 mètres d'écart les unes des autres, reprenant la définition du programme Africapolis. Cela permet de voir l'agglomération comme une ville vivante, qui évolue et se développe au-delà des limites des zones administratives. C'est donc le même concept que celui d'Africapolis sauf que nous n'avons pas imposé de critère démographique, ce qui signifie qu'une agglomération peut aussi bien être formée d'une seule shape que de plusieurs milliers (en l'absence d'autres shapes dans un rayon de 200 mètres, une shape unique est ainsi considérée comme une agglomération à part entière, une simplification qui facilite le traitement statistique).

Dans le projet TeleCense, cela revêt une moindre importance, car notre objectif principal est d'utiliser le niveau d'identification des shapes pour estimer la population. L'agglomération devient ainsi un outil de visualisation supplémentaire – au même titre que les zones administratives par exemple. Toutefois, ce niveau d'identification est important, car il joue un rôle clé dans le calcul de certaines variables, nous y reviendrons dans la partie suivante.

Enfin, notons que si une surface n'est pas identifiée comme bâtie, elle est classée soit comme sol nu, végétation, ou zone d'eau, à condition que cette caractérisation soit cohérente sur une période de six mois consécutifs. Par exemple, une prairie est classée comme végétation si, pendant six mois, la surface est constamment identifiée comme telle.

2. Caractérisation du bâti et caractéristiques propre aux shapes

La base de la détection de bâti donnée par TeleCense est donc la donnée appelée « shape ». En milieu urbain, les zones bâties sont généralement délimitées par les rues. Lorsque le réseau routier est récent et bien cartographié, il est possible d'obtenir une découpe claire et précise des zones bâties. En revanche, en milieu rural, les délimitations routières sont souvent moins précises ou difficiles d'accès, ce qui peut entraîner des shapes moins bien définies, englobant ainsi plusieurs habitations proches les unes des autres. Cette disparité explique la variabilité de taille, parfois très grande, entre les shapes, et pose parfois un problème d'homogénéité (un encart important sera dédié à cette question dans la section 2 du chapitre 4 sur le Bénin). Pour autant, c'est la granularité de détection la plus fine et elles représentent ainsi l'unité statistique principale. Tout comme les programmes d'estimation de la population qui s'appuient sur diverses variables au niveau des cellules de grille, nous appliquons une utilisons des variables définies au niveau des shapes pour modéliser la population à partir des zones bâties. C'est donc à partir de la détection des shapes que différentes caractéristiques sont produites, tout comme d'autres sont calculées au niveau des agglomérations, puis des zones administratives. Toutes ces variables sont récapitulées dans le tableau 2.

2.1 Variables liées aux shapes

a) Surface (1)

La variable area correspond à la surface totale de chaque shape, exprimée en mètres carrés (m^2). Chaque shape est un regroupement de pixels de 10 mètres de résolution (donc de 100m² de surface), la surface peut également être calculée par la formule suivante :

$$area = pixels * 100$$

b) Intensité de la lumière la nuit (nightlights)

Les lumières nocturnes proviennent de la mission satellitaire VIIRS (Visible infrared Imaging Radiometer Suite) à une résolution de 500 mètres. Nous calculons une moyenne des données qui représentent une synthèse mensuelle afin d'éviter le bruit et datent de 2018.

c) Altitude

L'élévation du sol est calculée à partir de la mission SRTM avec une résolution de 30 mètres.

d) Données climatiques

Le programme WorldClim / BioClim propose plusieurs résolutions et déclinaisons de données climatiques au format Raster, ce qui signifie qu'elles sont représentées sous forme de grille de pixels, chaque pixel contenant une valeur spécifique de température ou de précipitations pour une zone géographique donnée. Nous avons décidé de retenir celles du jeu « Historical climate data » à une résolution de 1 km² (Fick, Hijmans, 2017) pour les années 1970 à 2000. Plus précisément, nous utilisons les variables BIO1 et BIO12, qui représentent respectivement la température annuelle moyenne et les précipitations annuelles moyennes. Ces valeurs moyennes sont calculées sur la période de 30 ans, de 1970 à 2000, en utilisant les données mensuelles fournies par WorldClim. L'objectif est d'observer la tendance et l'hétérogénéité des zones étudiées, ce qui fait des données historiques une base fiable pour analyser ces variations climatiques.

e) Distances

Plusieurs variables correspondent à des distances entre la shape et certaines entités. Cela inclut les distances à la route la plus proche, à la zone d'eau la plus proche, au centre de santé le plus proche, ainsi qu'à la grande agglomération la plus proche. Les tracés routiers et les centres de santé[39] proviennent de la base de données OpenStreetMap (OSM). Les zones d'eau sont identifiées à partir de la classification TeleCense qui inclut largement les littoraux, fleuves, lacs et rivières. Une grande agglomération est définie comme une agglomération dont la surface totale des shapes dépasse 1 km². Ce seuil a été choisi pour garantir l'inclusion des grandes villes dans les zones géographiques étudiées, tout en évitant d'intégrer un nombre excessif d'agglomérations moins significatives.

Il est important de noter que, en raison de la faible résolution spatiale des variables d'intensité de la lumière nocturne et des conditions climatiques, celles-ci sont d'abord calculées au niveau des agglomérations, puis attribuées aux différentes shapes. Concrètement, chaque shape est associée à une agglomération via un index, ce qui permet de lui assigner les valeurs calculées pour cette agglomération. Étant donné la nature de ces données à résolution spatiale limitée, il est tout à fait cohérent que les shapes d'une même agglomération partagent les mêmes valeurs pour ces variables,

[39] https://data.humdata.org/dataset/hotosm_zaf_health_facilities?

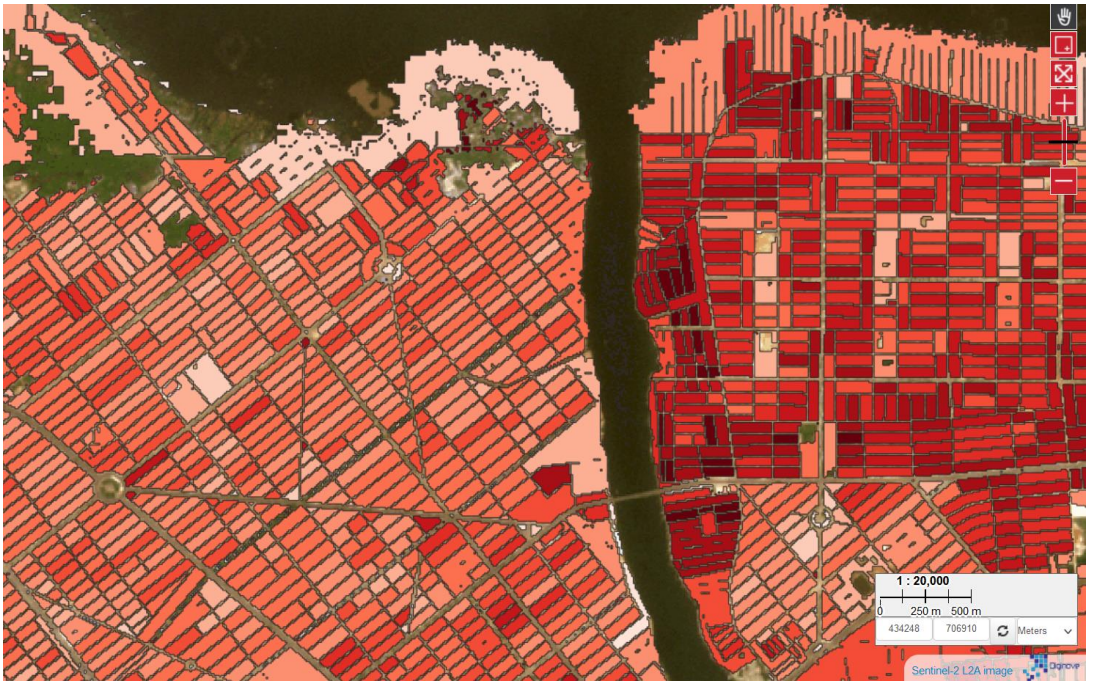
car elles dépendent de caractéristiques spatiales communes. De plus, cela réduit considérablement le temps de calcul, ce qui n'est pas négligeable puisqu'il y a généralement cinq fois plus de shapes que d'agglomérations.

La même méthode est appliquée pour les variables de distances. Par exemple, une agglomération urbaine a toujours une distance de 0 mètre par rapport à une route. Il semble donc logique que toutes les shapes formant cette agglomération héritent des mêmes valeurs concernant les distances, garantissant ainsi une cohérence spatiale et méthodologique.

f) Densité de bâti

La variable de densité de bâti est une combinaison de densité horizontale et verticale, calculée à partir des données radar Sentinel-1. Cette mesure adimensionnelle prend des valeurs discrètes comprises entre 0 et 10 (présentées par une variation de rouge dans la figure 11, toujours au nord de Cotonou, Bénin), dérivées principalement de l'intensité du signal radar, tout en intégrant légèrement un score indiquant si d'autres programmes de détection du bâti ont également identifié une présence bâtie (voir section 4 du chapitre 1 pour ces programmes). Concrètement, plus la valeur est élevée, plus l'intensité du signal radar est forte, ce qui indique généralement une concentration plus importante de bâtiments et une occupation du sol plus dense. En d'autres termes, cette variable tente de caractériser la structure du bâti : par exemple, les shapes en milieu urbain auront une valeur de densité de bâti moyenne plus élevée que celles en milieu rural.

Figure 11 : Variation de la densité de bâti des shapes



Sources : TeleCense, image issue du logiciel QGIS

g) Surfaces (2)

Quelques caractéristiques concernant la surface sont ajoutées :

a. Surface dite habitable

Afin de ne pas estimer la population de surfaces urbanisées mais non habitées (zones d'activités économiques, infrastructures de transport, hôpitaux, cimetières, etc.), nous utilisons la surface dite « habitable » qui repose sur le type de bâti identifié par les algorithmes de TeleCense et les données d'OSM. Dans la figure 12, nous reprenons un détail de l'image précédente avec un quartier situé au nord de Cotonou. L'image du haut montre toujours l'identification TeleCense, celle du bas montre l'identification OSM à la même date. Dans le premier cercle rouge, on voit qu'une partie de la shape est caractérisée comme un hôpital (hôpital BETHESDA) ; de même pour le cercle du bas, une très grande partie de la shape est caractérisée comme un cinéma (le Canal Olympia Wologuèdè), un autre bâtiment est la mairie de Cotonou, a priori bâtiment non résidentiel, et plus haut une partie correspond à un marché.

Figure 12 : Identification de l'utilisation du sol via OSM pour calculer la surface habitable des shapes

Sources : TeleCense, OSM, image issue du logiciel QGIS

Comme nous n'avons pas une précision suffisante pour identifier chaque bâtiment individuellement (ce qui permettrait de simplement enlever les bâtiments non résidentiels de nos estimations de population), nous décidons d'exclure un certain nombre de pixels de la shape totale selon le calcul suivant :

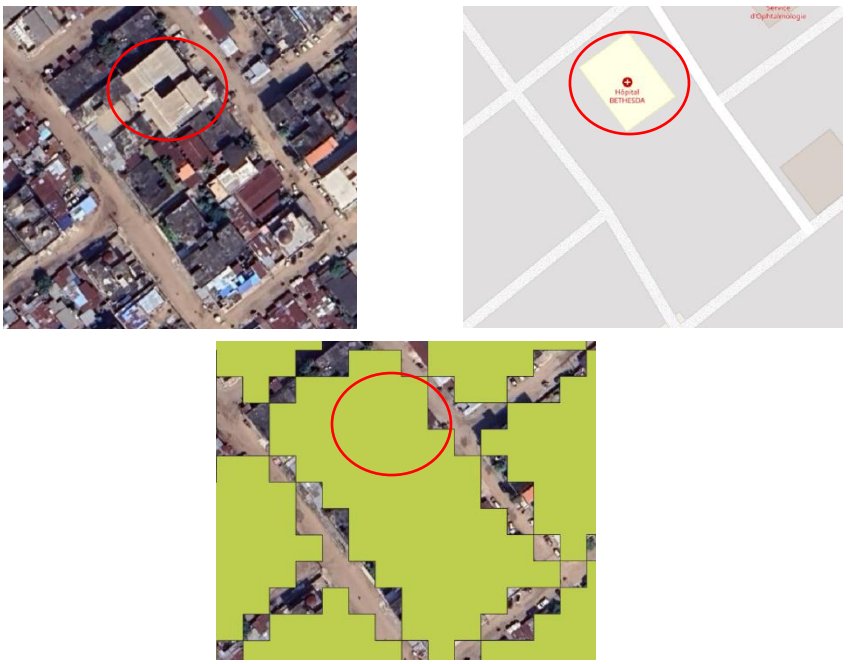
$$proportion_{habitee} = \frac{pixels - pixels\_empty}{pixels}$$

Qui permet ensuite de calculer la surface habitable :

$$surface_{habitable} = area * proportion_{habitee}$$

Ainsi, pour la shape à l'intérieur du premier cercle rouge, qui a un nombre de pixels égal à 61 (équivalent à une surface de 6100 m^2) et un nombre de pixels vides égal à 12 (représentant les pixels de l'hôpital), cette shape à une proportion habitée de 80 % qui donne une surface habitable finale égale à 4880 m^2. Au vu de la figure 13 suivante, il semble en effet pertinent de n'enlever qu'une partie de la shape et garder le reste des bâtiments résidentiels. De manière similaire, la surface habitable de la shape contenant le cinéma, le marché et la mairie est bien inférieure à la surface initiale.

Figure 13 : Occupation de l'hôpital dans la shape totale



Sources : OSM, TeleCense, images issues du logiciel QGIS

En général, pour identifier le bâti via OSM, nous utilisons les informations sur l'utilisation du sol (« land use ») et retenons de nombreuses étiquettes que nous considérons comme non résidentielles. La liste de ces étiquettes est disponible en annexe 1 où est présentée la requête OSM complète.
La principale limite de cette méthode réside dans le fait que ces données géographiques sont collectées et fournies par des bénévoles, ce qui rend leur disponibilité très hétérogène. Cela pose un défi, car nous nous appuyons fortement sur l'identification du bâti pour éviter d'estimer la population dans des zones non résidentielles.

b. Surface pondérée par la détection Google :

Nous utilisons les détections réalisées par Google Open Buildings à une résolution de 0.5 et 1 mètre, que nous adaptons ensuite à une résolution de 10 mètres pour les aligner avec les données des shapes TeleCense. Cette adaptation permet de calculer le nombre de pixels identifiés par Google à l'intérieur de chaque shape de TeleCense. La surface de la shape pondérée par la détection Google est alors calculée de la manière suivante:

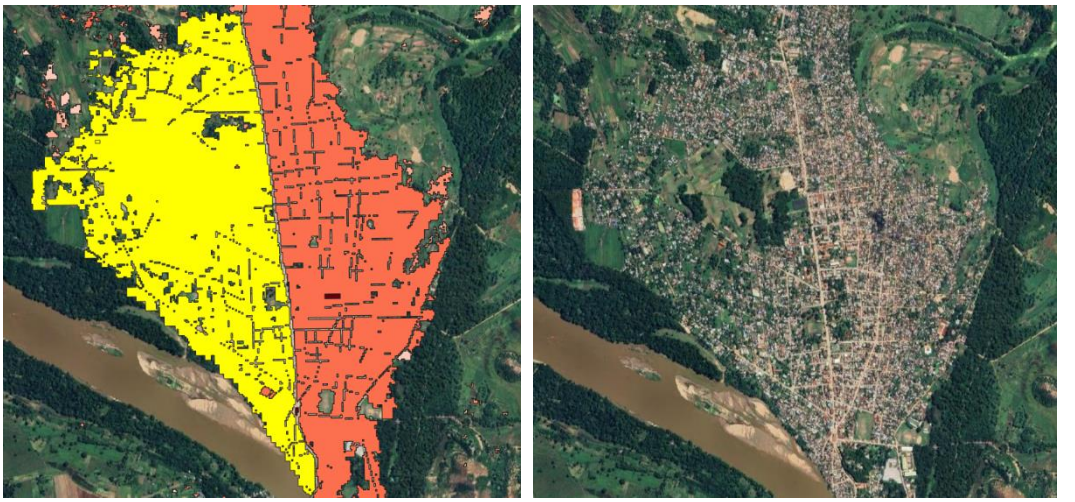
$$surface_{google} = pixels_{google} * 100$$

Puis nous calculons la surface dite habitable pondérée par la détection Google Open Buildings :

$$surface_{habitable\_google} = surface_{google} * proportion_{habitee}$$

La variable de surface Google est intéressante car elle peut venir raffiner certains calculs de surface erronés des shapes du programme TeleCense. La différence est souvent minime : par exemple si nous reprenons l'exemple de la shape avec l'hôpital, la surface habitable google est de 4580 m^2 au lieu 4880 précédemment. Mais elle peut aussi être importante : en effet, dès lors que le réseau routier n'est pas assez précis ou actualisé, il arrive que certaines shapes soient moins bien définies et qu'on identifie des faux-positifs. Bien que faisant figure d'exception – car on y voit une très grande erreur – la figure 14 illustre parfaitement ce propos :

Figure 14 : Problème d'identification pour certaines grandes shapes – Région de Diana – Madagascar - 2018



Sources : TeleCense, images issues du logiciel QGIS

Sur l'image de gauche, on voit une très grande shape identifiée (sélectionnée en jaune) et sur l'image de droite, la réalité du terrain à travers l'image satellite. Au centre de cette shape, on constate l'absence de bâtiments, ce qui indique clairement une erreur d'identification. L'utilisation de la surface Google peut corriger ces erreurs : dans ce cas, on passe d'une surface de 2,2 km² à une surface estimée par Google de 0,88 km², soit une réduction de 60 %. Cette différence est significative, et bien que l'exemple précédent soit une exception, il suffit d'une dizaine de Shapes similaires pour entraîner des erreurs importantes dans les estimations de population.

Dans cet exemple, la détection TeleCense date de 2018, tandis que celle de Google, bien que publiée en 2022, repose sur des images dont la date varie en fonction de leur disponibilité : certaines sont récentes (2022), tandis que d'autres proviennent d'années antérieures. Est-il donc raisonnable d'utiliser les données de Google dans ce cas ? En réalité, il est fort probable que les bâtiments détectés en 2018 soient encore présents en 2022, à moins qu'ils n'aient été démolis, ce qui reste relativement rare. De plus, lorsqu'une destruction a lieu, elle est souvent suivie d'une reconstruction, ce qui permet de supposer l'utilisation des données de Google de 2022 pour affiner les surfaces détectées en 2018 peut être justifiée, surtout si ces données apportent des corrections aux erreurs de détection initiales. En revanche, il est important de noter que les bâtiments construits après 2018 et détectés par Google ne seront pas reflétés dans les données TeleCense de 2018, car ils n'existaient pas encore à cette date.

Cette approche permet donc d'améliorer la précision des surfaces pour les bâtiments existants tout en tenant compte des limitations liées à la période de détection. Cependant, pourquoi ne pas utiliser directement les données Google ? Tout

simplement parce que ces données datent de 2022 et, comme nous l'avons déjà mentionné dans le chapitre précédent, elles constituent une mosaïque issue de plusieurs années. Par conséquent, elles ne sont pas adaptées pour une estimation de la population à une date précise, ce qui est essentiel pour notre analyse.

2.2 Variables liées aux Zones Administratives

a) Décomptes de population :

Nous récupérons les effectifs de population aux recensements principalement via les plateformes des instituts nationaux de statistiques ou des plateformes comme The Humanitarian Data Exchange[40] (HDX). La disponibilité et la qualité de ces données peuvent varier considérablement d'un pays à l'autre.

b) Découpage administratif :

Les découpages des zones administratives sont généralement accessibles en libre accès à l'échelle mondiale, via différentes sources comme geoBoundaries[41], le bureau des Nations Unies pour la coordination des affaires humanitaires[42] (OCHA) ou encore GADM[43]. Ces différents fournisseurs sont souvent regroupés sur la plateforme de The Humanitarian Data Exchange (HDX). Cependant, la précision et la disponibilité des données varient considérablement d'un pays à l'autre. Par exemple, au Kenya, la précision peut atteindre le niveau 5 avec les sub-locations, mais il reste très difficile d'obtenir des fichiers actualisés et librement accessibles pour ces tracés administratifs. Dans les faits, les tracés sont souvent accessibles mais ne sont souvent pas complètement actualisés afin de correspondre aux limites géographiques du dernier recensement. En effet, comme le soulignent Runfola et al. (2020), l'intérêt pour les limites administratives n'a pas été accompagné d'un effort significatif de collecte et il en résulte que les tracés administratifs restent inégalement accessibles et actualisés, notamment dans les pays africains. Dans ce contexte, nous sélectionnons les tracés administratifs les plus précis et les plus récents disponibles pour chaque zone d'étude.

Les décomptes de population et le découpage administratif sont des données essentielles et sont les plus facilement accessibles. Bien qu'il soit possible d'ajouter d'autres caractéristiques, cela nécessite une recherche approfondie, un temps souvent limité pour des projets comme TeleCense, qui vise à automatiser autant que possible ses processus. Toutefois, il est souvent possible d'obtenir le nombre moyen de personnes par ménage, c'est donc une variable qui est ajoutée par défaut.

[40] https://data.humdata.org/
[41] https://www.geoboundaries.org/
[42] https://cod.unocha.org/
[43] https://gadm.org/

On calcule également certaines variables, telles que la densité de population par zone administrative, ainsi que la proportion de bâti détecté à l'intérieur de ces mêmes zones administratives. Toutes les variables sont récapitulées dans le tableau 2 suivant.

Tableau 2 : récapitulatif des variables

Nom de la variable	Description de la variable
1 – Données du projet TeleCense	
fid	Index statistique de la Shape (feature identification)
city_fid	Index indiquant à quelle agglomération la Shape appartient
pixels	Nombre de pixels de 10*10 mètres composant la Shape
area	Surface (m^2)
elevation	Elévation au sol (calculée à partir de la mission SRTM)
slope	Pente (%)
density	Densité de bâti de la Shape, valeur comprise entre [0,1,…,10]
distanceToWater	Distance au premier point d'eau calculée à partir de la classification TeleCense (m)
distanceToCity	Distance à la première grande agglomération (def. TeleCense)
population_TeleCense	Population que l'on a répartie / estimée dans la shape
2 – Données additionnelles ajoutées ou calculées à partir de sources extérieures	
nightlights	Moyenne de l'intensité de lumière la nuit (nW/cm^2/sr)
distanceToRoad	Distance minimale à la route (m)
distanceToHealth	Distance minimale au premier point de santé (m)
preci	Précipitations moyennes annuelles (WorldClim)
temp	Température moyenne annuelle (WorldClim)
pixels_ggle	Nombre de pixels identifiés par Google dans la Shape
area_ggle	Surface pondérée par identification Google (m^2)
pixels_xx	Nombre de pixels identifiés par DLR, OSM, GHS et Microsoft
pixels_empty	Nombre de pixels supposément non habités
surface_habitable	Surface supposément habitée (m^2)
surface_habitable_ggle	Surface supposément habitée pondérée par Google (m^2)
3 – Région d'intérêt (ROI)	
ADM0_EN	Zone administrative niveau 0 – correspond au pays
ADM1_EN	Zone administrative niveau 1
ADM2_EN	Zone administrative niveau 2
ADM3_EN	Zone administrative niveau 3 (si existant)
ADM4_EN	Zone administrative niveau 4 (si existant)
ADM5_EN	Zone administrative niveau 5 (si existant)
population_RGPH	Décomptes de population venant du dernier RGPH pour chacune des zones administratives disponibles
pop_density	Densité de population pour chacune des zones administratives disponibles
nb_moyen_menage	Nombre moyen de personnes par ménage
proportion_bati	Proportion de bâti pour chacune des zones administratives disponibles

Sources : Variables issues du projet TeleCense

3. Méthodologie permettant l'estimation de population à l'intérieur du bâti

L'objectif central de la thèse est d'adapter les méthodes d'estimation de la population connues, généralement conçues pour produire des données carroyées, à un ensemble de zones bâties que nous appelons shapes dans ce projet. Pour ce faire, nous développons une méthode descendante classique visant à répartir la population de manière précise dans les shapes. Ensuite, nous utilisons ces estimations au niveau des shapes pour créer un modèle ascendant théoriquement capable d'estimer la population dans d'autres zones géographiques que la zone d'entraînement. La vision globale est donc d'avoir une méthode permettant d'estimer la population sans recensement préalable (voir le schéma récapitulatif en fin d'introduction, figure 7). Dans cette partie, nous détaillons le fonctionnement de ces méthodes, leur utilisation et leur validation.

3.1 Approche descendante – La désagrégation

L'objectif de la méthode descendante, est de répartir la population connue au niveau de zones dites sources vers des zones dites cibles. Plus concrètement, les caractéristiques de la population connues au niveau de la zone administrative (zone source, par exemple la commune à Madagascar), comme par exemple son effectif, sont distribuées entre plusieurs shapes (zones cibles) incluses et constituant une partition de la zone administrative. Cette démarche nommée désagrégation, ou encore cartographie dasymétrique peut être exprimée par la formule suivante (Eicher, Brewer, 2001) :

$$\hat{P}_S = P_{ZA} * W_{ZA,S}$$

Avec $\hat{P}_S$ la population estimée dans la shape S, P_{ZA} la population connue de la zone administrative ZA et $W_{ZA,S}$ la proportion de la population source (la zone administrative ZA) alloué à la cible S.

Lorsque les shapes constituent une partition de la zone administrative ZA :

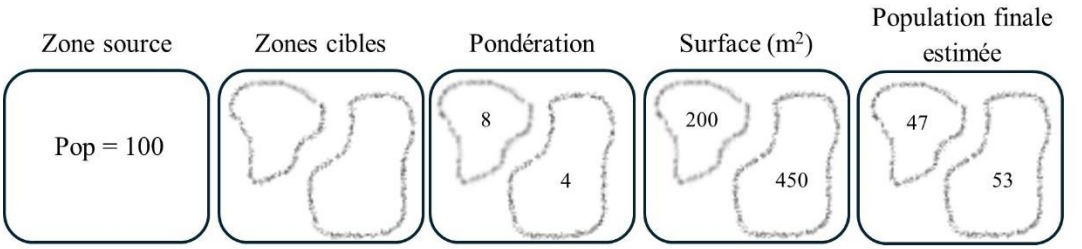
$$W_{ZA,S} = \frac{P_{ZA,S}}{P_{ZA,}}$$

Estimé par :

$$\frac{A_S * W_S}{\sum s(A_S * W_S)}$$

Avec A_S la surface habitable connue de la shape et W_S la pondération retenue ou calculée (expliquée dans la partie suivante) permettant de théoriquement répartir la population dans les shapes en tenant compte de leurs caractéristiques corrélées à la densité de population, comme par exemple la hauteur des bâtiments, la couverture du sol ou même un ensemble de variables physiques et environnementales ayant un lien avec la densité de population. La figure 15 suggère l'importance du choix de la pondération avec un exemple d'une zone source de 100 habitants à répartir dans deux shapes. Malgré la différence de surface des deux zones cibles, elles ont une population finale estimée très proche car la pondération associée à la première zone cible est le double de celle associée à la seconde.

Figure 15 : Exemple schématique de désagrégation avec pondération



Le calcul pour la première shape d'une surface de 200 m^2 et d'une pondération de 8 est le suivant :

$$\hat{P}_{S1} = 100 * \frac{(200 * 8)}{(200 * 8 + 450 * 4)} = 47{,}06$$

Avec A1 = 200, A2 = 450 ; W1 = 8 et W2 = 4.

3.1.1 Mise en œuvre de l'estimation de population par désagrégation

Dans la thèse, nous commencerons par appliquer la désagrégation à Madagascar, qui servira de pays de test et de base pour les analyses ultérieures. Ensuite, nous testerons l'approche au Bénin, qui nous permettra d'évaluer la précision de la désagrégation au fil des années. Quelle que soit la région ou le pays étudié, les prérequis à la désagrégation sont clairs et se définissent comme suit :

1) Récupérer les décomptes de population suivant les zones sources

Pour cela, il faut se fier aux résultats du dernier recensement national disponible afin d'obtenir la répartition de la population selon les différents niveaux administratifs disponibles.

2) Récupérer les tracés géographiques puis leur associer les décomptes de population du recensement

Une fois les décomptes de population selon les différents niveaux administratifs obtenus, il faut les lier géographiquement à leur niveau administratif respectif. Pour ce faire, il est nécessaire de disposer des tracés géographiques des limites administratives de la région étudiée. Cependant, comme mentionné précédemment, il existe de nombreuses situations où les découpages administratifs relatifs à un recensement récent et les limites administratives géoréférencées accessibles en ligne ne correspondent pas. Cela peut être dû au fait que les tracés administratifs ne soient pas actualisés, ou qu'ils ne soient pas disponibles ou simplement pas rendus publics. Dans de tels cas, la question se pose de savoir comment mettre en correspondance ces sources de données.

Parfois, c'est très simple, comme au Bénin (chapitre 4) où il suffit d'ajuster quelques noms de zones administratives pour garantir l'appariement. D'autres fois, c'est plus compliqué, comme à Madagascar, où nous avons dû développer une méthode qui modifie légèrement la base de données du recensement ainsi que le fichier géoréférencé des limites administratives des communes afin de lier les deux bases de données (chapitre 3).

3) Cartographie du bâti :

Une fois la zone d'étude délimitée, nous mettons en place les algorithmes de TeleCense afin d'identifier et caractériser les shapes. Parfois, certaines shapes sont à cheval sur deux zones administratives. Dans ce cas, nous décidons de découper la shape en deux pour que chacune des zones bâties n'appartiennent bien qu'à une seule zone administrative. Cela assure de ne pas faire d'erreur lors de la répartition de la population à partir du niveau administratif.

4) Choix ou calcul de la pondération W_S.

Dans cette thèse, la pondération W_S est retenue ou calculée selon quatre approches différentes. L'objectif est de déterminer quelle méthode permet la répartition de la population au sein des shapes avec le moins d'erreurs et donc avec la plus grande

précision (l'évaluation des méthodes est détaillée dans la section 5 suivante) : d'une part, pour assurer une transition cohérente d'un niveau administratif à un autre plus détaillé, et d'autre part, pour obtenir une distribution aussi représentative que possible dans les shapes. Cela nous fournira une base de population pertinente pour entraîner un modèle ascendant dans la suite de cette thèse.

Ici, ces quatre définitions des pondérations sont présentées de la plus simple à la plus technique :

a) Allocation proportionnelle à la surface / Interpolation surfacique

Le principe de cette méthode est de répartir la population des zones administratives considérées au sein des shapes proportionnellement à la surface de ces dernières. C'est la méthode de répartition la plus simple. Dans ce cas :

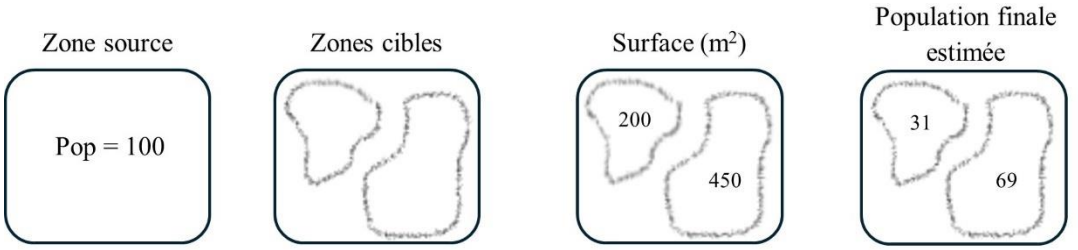
$$W_{ZA,S} = \frac{A_S}{A_Z}$$

et :

$$\hat{P}_S = P_{ZA} * \frac{A_S}{\sum s \in A_S}$$

La population de la shape est donc égale à la population de la ZA multipliée par le quotient de la surface de la shape sur la somme totale des surfaces des shapes dans la ZA. Avec cette méthode, la pondération est égale à 1 pour toutes les shapes et c'est comme s'il n'y avait pas de pondération, comme illustré dans l'exemple en figure 16.

Figure 16 : Exemple schématique de la désagrégation par interpolation surfacique



b) Allocation proportionnelle à la variable de densité de bâti

Pour cette deuxième méthode, la pondération utilisée est l'une des variables disponibles : la densité de bâti, qui prend des valeurs allant de 0 à 10. Pour cette deuxième approche, aucun calcul préalable n'est nécessaire mais cela suppose

néanmoins que les valeurs correspondent à une mesure, c'est-à-dire une échelle quantitative. La formule devient alors la suivante :

$$P_S = P_{ZA} * \frac{A_S * W_S}{\sum S \in ZA\ (A_S * W_S)}$$

Avec $W_S = density_S = \{0,1,2,\dots,10\}$.

Cette méthode de répartition s'apparente à une représentation en volume, où la surface est multipliée par la variable de densité pour créer une sorte de « volume bâti ». Ainsi, la répartition de la population tient compte non seulement de l'étendue de la zone cible (shape), mais aussi de sa densité de construction, offrant une estimation prenant en compte l'occupation de l'espace (voir figure 16 précédente).

Cette approche peut être étendue en prenant en compte plusieurs variables (x_1,…,x_p) explicatives de la densité dont une fonction f(x1,…,xp) est proportionnelle à la densité de la population. Il faut bien comprendre que la relation ne peut pas être établie au niveau des shapes d'intérêt puisque cela nécessiterait de connaitre leur densité (et donc leurs effectifs de population et dans ce cas là, la question de désagrégation serait déjà résolue). Nous allons donc chercher à estimer la fonction f à un niveau supérieur aux shapes (par exemple le niveau communal) puis nous l'appliquerons aux shapes pour prédire la densité. Cela suppose donc que les variables puissent être observées pour chacune de ces deux échelles d'observation. C'est le cas des variables présentées dans le tableau 2 précédemment. Pour cette approche étendue, nous utilisons deux méthodes : la régression linéaire multivariée et les forêts aléatoires.

c) <u>Modélisation de la densité de population des ZA pour estimer une variable de pondération au niveau des shapes</u>

i. <u>Modélisation linéaire de la densité</u>

Cela consiste à modéliser la densité communale par un modèle linéaire multivarié. Soit Yi la densité de la commune i de ZA. Alors :

$$Y_i = f(X_1, X_2 \dots X_k) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_k x_{ki} + \varepsilon_i$$

Où i est l'indice variant de 1 à n (n étant le nombre de communes), les X_k sont les facteurs susceptibles d'expliquer la variable Y de densité de population, les coefficients β_k sont les paramètres du modèle à estimer pour chacune des variables et ε_i est le terme résiduel à estimer. Son estimation correspond à $Y_i - \widehat{Y_i}$, l'écart entre la

densité de la population modélisée et sa prédiction. Le modèle est estimé par la méthode des moindres carrés, les inférences étant rendues possibles par la supposition que les termes ε_i sont indépendants et distribués selon une distribution $N(0, \sigma^2)$.

La régression linéaire suppose néanmoins une relation linéaire entre la variable dépendante (densité de population) et les variables indépendantes, ce qui peut donner quelques limites à son utilisation. En effet, dans de nombreux cas, les relations entre ces variables peuvent être non linéaires, ce qui limite la capacité du modèle linéaire à capturer la complexité des données et peut avoir comme conséquence une capacité prédictive limitée.

Dans ce contexte nous testons également une modélisation par forêts aléatoires qui permet de capturer des relations non linéaires, ce qui pourrait être plus adapté à un contexte d'estimation de population dans des contextes avec des données complexes et hétérogènes.

ii. Par forêt aléatoire :

Pour l'algorithme de forêt aléatoire, nous reprenons l'approche de Stevens et al. (2015), avec comme objectif d'estimer la population dans les shapes et non dans des pixels de 100x100 mètres. Ainsi, nous utilisons un ensemble d'arbres de décision pour estimer la densité de population :

1. Construction du modèle :

Les forêts aléatoires sont appliquées à l'ensemble d'entraînement. Chaque arbre est construit en utilisant des échantillons aléatoires des données et des sous-ensembles des variables explicatives. Pour chaque arbre f_i la prédiction est :

$$\widehat{Y_i} = f_i(X_1, \dots, X\_p)$$

Où $X_1, \dots X\_p$ sont les variables explicatives.

2. Validation croisée :

La validation croisée avec cinq plis est utilisée pour optimiser le modèle. Le modèle est entraîné cinq fois, chaque fois avec quatre plis pour l'entraînement et un pli différent pour la validation. Les prédictions de tous les arbres sont combinées pour obtenir une estimation finale :

$$\hat{Y} = \frac{1}{N} \sum_{i=1}^{N} \widehat{Y_i}$$

Où N est le nombre total d'arbres dans la forêt.

iii. Évaluation des modélisations

Pour les méthodes de modélisation linéaire et de forêt aléatoire, le processus d'estimation de la densité de population est similaire en termes de préparation des données ; il s'agit de séparer les données en deux ensembles : un ensemble d'entraînement (70 % des ZA) et un ensemble de test (30 % des ZA). Cette séparation permet de former le modèle sur un sous-ensemble des données et de tester sa performance sur un sous-ensemble distinct.

Les performances des modèles sont ensuite évaluées et comparées en utilisant des métriques telles que l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2) sur l'ensemble de test. L'utilisation de forêts aléatoires permet d'estimer la densité de population en combinant les prédictions de plusieurs arbres de décision, théoriquement cela devrait offrir une estimation sensée être plus robuste et moins sensible aux variations des données par rapport aux méthodes de régression linéaire.

1) Evaluation des méthodes de désagrégation :

Une fois les différentes approches de désagrégation effectuées, nous avons quatre répartitions de la population différentes à comparer. Comme il est impossible de comparer la précision de l'estimation directement au niveau de la shape (faute de données terrain de validation à ce niveau unique), il s'agit généralement de désagréger à partir d'un niveau administratif plus grossier (moins fin) jusqu'au niveau le plus fin disponible. Autrement dit, si le niveau administratif le plus fin du pays étudié est le niveau 3 (ADM3), nous déciderons afin de comparer les résultats de désagréger à partir du niveau 2 (ADM2).

Cela permet de comparer les prédictions additionnées des populations estimées à partir d'une désagrégation depuis un niveau administratif moins précis aux décomptes officiels de population des zones administratives les plus précises. Il s'agit ensuite de calculer le pourcentage d'erreur relative pour chacune des communes selon cette formule :

$$Pct.Erreur_{relative} = \frac{Decomptes_{RGPH} - Predictions}{Decomptes_{RGPH}} * 100$$

Cela permet une visualisation cartographique, montrant les communes sous- ou surestimées et permettant de faire des comparaisons entre les méthodes de désagrégation. L'approche donnant l'erreur moyenne la plus faible sera retenue

comme celle permettant la meilleure répartition de la population dans les zones bâties. Elle sera donc utilisée pour prédire la population dans les shapes à partir du niveau communal, pour ensuite entrainer le modèle ascendant.

3.2 Approche ascendante

L'objectif de la méthode ascendante est d'estimer la population des zones bâties sans s'appuyer sur des données de recensement, mais uniquement sur les caractéristiques des shapes. Pour entraîner un modèle suivant cette approche, nous proposons d'exploiter les résultats de la désagrégation afin que la population estimée par l'approche descendante serve de référence pour les futurs entraînements et ajustements du modèle. Ces modèles seront développés à l'aide d'une modélisation linéaire et d'un algorithme de forêt aléatoire. Les méthodes employées seront détaillées plus précisément dans le chapitre 5 dédié.

4. Conclusion

Ce chapitre a présenté le programme TeleCense, un outil innovant pour la détection et la caractérisation des bâtiments à partir d'images satellites. En utilisant les données des satellites Sentinel, TeleCense permet de suivre l'évolution rurale et urbaine et a pour but de fournir des estimations précises de la population résidentielle à l'intérieur des groupements de bâtiments identifiés.

Pour chaque shape identifiée, des caractéristiques spécifiques sont attribuées ou calculées, permettant ainsi de mettre en œuvre les méthodes développées en fin de chapitre. Il s'agit des modèles descendants et des modèles ascendants qui visent d'une part à prédire la répartition spatiale de la population de manière précise et d'autre part à estimer la population dans d'autres zones géographiques où par exemple les données de recensement sont obsolètes ou inexistantes.

La suite de cette thèse examine les approches descendantes, avec l'élaboration et le test de ces méthodes sur six régions de Madagascar. Pour cela, on utilise les données du recensement de 2018 (RGPH-3) couplées à la cartographie du bâti de la même année. Une validation complémentaire est menée au Bénin, où il sera nécessaire de projeter la population du recensement de 2013 (RGPH-4) sur plusieurs années, permettant d'évaluer l'efficacité de la méthode même avec des données de recensement plus anciennes et, simultanément, d'examiner la précision de la détection du bâti et l'estimation de la population au fil du temps.

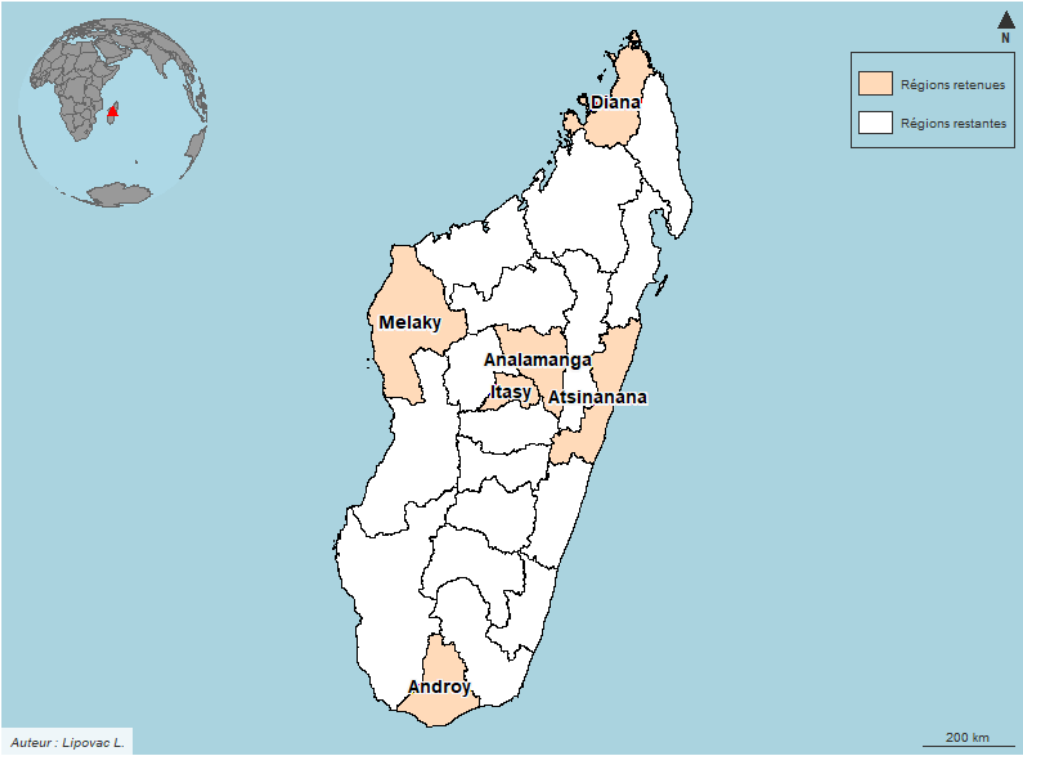
PARTIE 2 : Estimation de population par méthode descendante

Chapitre 3 – Désagrégation de la population à Madagascar

En 2018, le troisième Recensement Général de la Population et de l'Habitat (RGPH-3) a été réalisé à Madagascar. Il s'inscrit dans une série de recensements effectués dans le pays, les précédents ayant eu lieu en 1975 et 1993. Ce recensement, particulièrement attendu, fournit des données actualisées essentielles à la planification et aux politiques publiques, notamment des informations sur les caractéristiques démographiques et socio-économiques de la population, ainsi que sur les conditions de logement et le bien-être des ménages, jusqu'au niveau géographique le plus fin (INSTAT, 2021a).

Madagascar est subdivisé en 22 régions et comptait, selon le RGPH-3, 25 674 196 habitants en 2018. La population est majoritairement rurale (80,7 %), avec une densité moyenne de 43,4 habitants/km². L'unité administrative de base est le fokontany, souvent assimilé à un village ou un quartier urbain. Toutefois, les résultats du recensement sont agrégés et disponibles au niveau des communes, ce qui nous conduit à travailler à cette échelle pour notre étude. Cette analyse se concentre sur six régions spécifiques, représentées et localisées dans la figure 17.

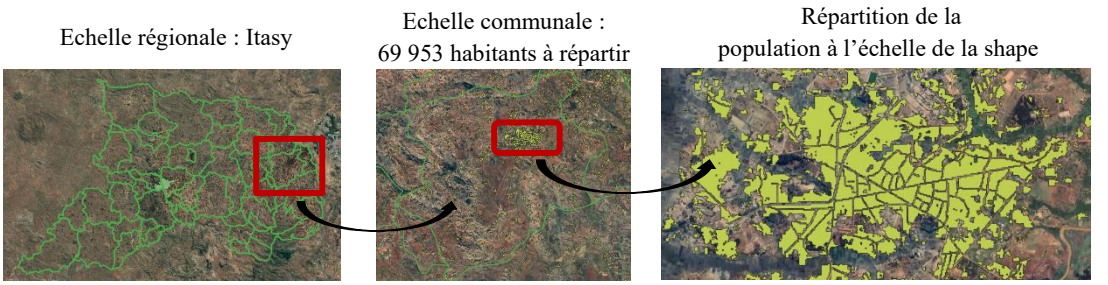
Figure 17 : Carte de situation et présentation des six régions retenues à Madagascar



Sources : Tracés des limites administratives fournis par OCHA

Ce chapitre met en œuvre la méthodologie d'approche descendante décrite dans le chapitre 2 en s'appuyant sur les résultats du recensement pour redistribuer la population des communes malgaches au sein des shapes identifiées par TeleCense. Pour cela, les contours géographiques des communes sont d'abord récupérés et associés aux décomptes de population officiels du RGPH-3. Cette démarche est illustrée dans la figure 18 à travers l'exemple de la région d'Itasy, qui regroupe 51 communes. Parmi elles, la commune d'Imerintsiatosika compte 69 953 habitants, que l'on cherche à redistribuer au sein des shapes qui lui sont rattachées (image de droite). C'est cette approche qui sera appliquée à l'ensemble des communes des six régions sélectionnées à Madagascar.

Figure 18 : Illustration de la désagrégation à partir d'une commune de la région d'Itasy



Sources : Tracés des limites administratives fournis par OCHA et identification du bâti par TeleCense
Images issues du logiciel QGIS

1. Préparation des zones sources : les communes malgaches

1.1 Les six régions du cas d'étude

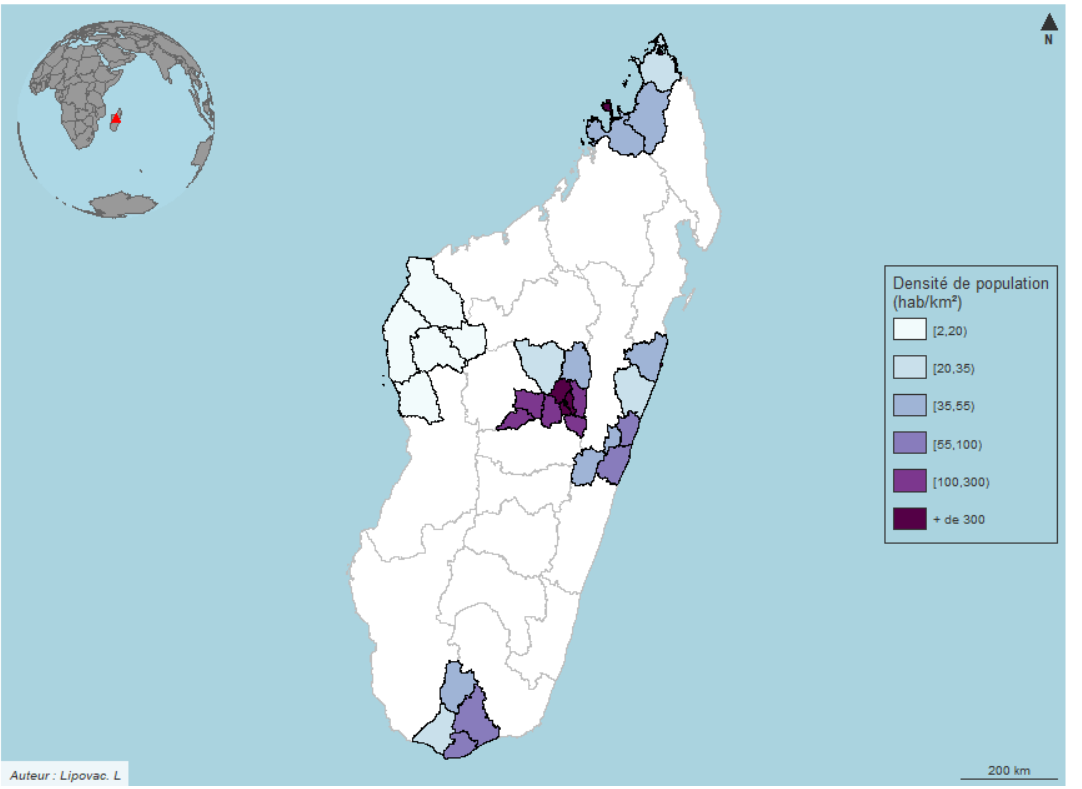
Bien que cartographier le bâti de tout Madagascar ne présente pas de difficultés techniques particulières pour le projet TeleCense, cela représente un coût en temps, en gestion de données et en stockage très important, notamment du fait que le pays est légèrement plus grand que la France, avec une superficie égale à 587 041 km^2. Pour ces raisons, nous décidons de restreindre notre étude à 6 régions de Madagascar, choisies pour leur diversité environnementale, climatique et de bâti et présentées et localisées par les figures 17 et 19.

Le choix de ces six régions avait comme double enjeu de pouvoir mobiliser un échantillon représentant la diversité des régions de Madagascar et, dans la mesure du possible, de l'Afrique en général, afin d'être en accord avec l'objectif global de la thèse visant à exporter les relations trouvées ici, en dehors de Madagascar.

La figure 19 illustre la distribution des densités de population au sein des régions sélectionnées. À l'est de Madagascar, la région d'Atsinanana, située près du littoral, présente une densité de population relativement élevée de 67,1 hab./km². Au centre du pays, les densités de population atteignent leur maximum avec 208,9

hab./km² pour Analamanga et 136,6 hab./km² pour Itasy. Analamanga, en tant que région capitale abritant Antananarivo, concentre près de la moitié de la population urbaine du pays (Africapolis, 2024). Son inclusion dans l'étude était donc indispensable. Itasy, adjacente à Analamanga, fait partie des Hautes Terres, une région caractérisée par de nombreux massifs et lacs, ainsi qu'une densité de population historiquement élevée, favorisée par des conditions sanitaires et politiques avantageuses (Magrin et al., 2022; Raison, 1974). En progressant vers l'ouest, les densités de population décroissent significativement, atteignant un minimum de 7,6 hab./km² dans la région de Melaky, où les espaces sont quasi vides de population. Globalement, la partie occidentale du pays, à l'exception de quelques zones côtières, est faiblement peuplée.

Figure 19 : Densité de population par district des six régions retenues à Madagascar



Sources : Décomptes de population issus du RGPH3 (INSTAT) et tracés des limites administratives fournis par OCHA

Diana, dans la partie nord du pays, et Androy, à son opposé au sud, présentent des densités de population similaires, respectivement de 45 et 48 hab./km², mais des dynamiques distinctes. Diana, avec un taux d'urbanisation de 34 %, est l'une des régions les plus urbaines de Madagascar (moyenne nationale de 19,3 %) et bénéficie d'un climat tropical. Androy, quant à elle, est principalement rurale et aride. En raison de ses contraintes particulières, les activités agricoles y sont très limitées, et cette

région est caractérisée par une forte émigration, enregistrant le troisième plus grand solde migratoire négatif de Madagascar (INSTAT, 2021a).

1.2 Des changements administratifs fréquents et nombreux à Madagascar

Depuis son accès à l'indépendance en 1960 et la création de la 1ère république, Madagascar n'a connu que trois opérations de recensements, à savoir en 1975, en 1993 et le dernier en 2018 (INSTAT, 2021b). Pour autant, Madagascar a comme beaucoup de pays en développement modifié son cadre politique, juridique et institutionnel à de nombreuses reprises (Bidou, Droy, Fauroux, 2008). Actuellement, d'après le RGPH-3, « selon le Décret n° 2014-020, l'Etat malagasy repose sur un système de Collectivités Territoriales Décentralisées. L'organisation administrative regroupe des fokontany dans une commune, des communes dans un district, des districts dans une région et des régions dans une province » (INSTAT, 2021c). On compte ainsi en 2018, six provinces, 22 régions, 119 districts, 1 693 communes (dont 1 617 rurales et 76 urbaines) et 18 251 fokontany (INSTAT, 2021c). Avant cela, six grands changements administratifs avaient été effectués. Parmi les plus notables, Madagascar voit en 1994 lors de la 3ème république et la loi de décentralisation, la création de communes dotées de pouvoirs étendus. A cette époque, la loi 94-001 dénombrait alors 28 régions, 158 départements et 1295 communes (République de Madagascar, 1995). C'est le 17 juin 2004 d'après la loi 2004-001 que les 22 régions actuelles sont créées et seront à la fois des Collectivités Territoriales Décentralisées et des circonscriptions administratives (République de Madagascar, 2004). Finalement, la 4ème république et la constitution de 2010 à travers la loi 2014-020 (République de Madagascar, 2014), fait également des communes des Collectivités Territoriales Décentralisées et on décomptera alors 6 provinces, 22 régions et 1549 communes urbaines ou rurales. La loi n°2015-002 nous amène à une situation proche de 2018 et du RGPH-3 en redécoupant le pays en 1693 communes (République de Madagascar, 2015). Dernier signe d'une forte instabilité juridique : depuis la fin du recensement en 2018 deux nouvelles régions ont encore vu le jour ; en 2021, Vatovavy a été officiellement désignée comme la 23ème région de Madagascar[44], suivie d'Ambatosoa en 2023 portant le total à 24 régions[45].

[44] https://www.presidence.gov.mg/actualites/1405-23eme-region-de-madagascar-enieme-velirano-realise.html

[45] https://assemblee-nationale.mg/adoption-de-la-loi-instaurant-la-24eme-region-ambatosoa/

1.3 Les sources disponibles et mobilisées

1.3.1 Le recensement de la population malgache

Dans le tome 2 des résultats globaux du RGPH-3 (INSTAT, 2021d), les décomptes de population présentés sont triés par régions, districts et communes. Dans le but d'obtenir la table de données de population correspondant aux décomptes de population des communes malgaches au niveau communal (niveau 3 – le plus fin disponible), il faut exporter les résultats du tableau n°10 « Effectif de la population recensée dans les ménages ordinaires par milieu de résidence et sexe selon le District et la Commune » allant de la page 37 à la page 91 du format PDF au format Excel. En 2018, le recensement fait état de 1693 communes. Comme nous nous limitons à six régions, la table finale de données filtrée comprend 463 communes.

Certaines erreurs se sont glissées dans le nom des communes du RGPH-3, il est donc nécessaire de les corriger par l'erratum (INSTAT, 2021e) publié en parallèle des deux premiers tomes définitifs. Par exemple, certaines communes dans la région de Diana, comme celles d'Andranovondronina, de Ramena, d'Ambolobozobe et d'Antsakoabe étaient toutes renseignées sous le nom d'Antsiranana-II. Les lignes du tableau étant numérotées il suffit de modifier la commune par son nom exact comme indiqué dans l'erratum. Voici les 18 changements qu'il a été nécessaire d'effectuer pour les communes des 6 régions sélectionnées (il n'y a pas de changements à effectuer pour les régions d'Itasy et d'Analamanga) :

<u>Région de Diana :</u>

- au lieu de Antsiranana-II lire Andranovondronina[46],
- au lieu de Antsiranana-II lire Ramena,
- au lieu de Antsiranana-II lire Ambolobozobe,
- au lieu de Antsiranana-II lire Antsakoabe,
- au lieu de Ambilobe lire Antsaravibe,
- au lieu de Nosy-Be lire Hell Ville,
- au lieu de Nosy-Be lire Bemanondrobe,
- au lieu de Ambanja lire Ambalahonko.

<u>Région de Melaky :</u>

- au lieu de Ambatomain-Ty lire Ambatomainty.

<u>Région d'Atsinanana :</u>

- au lieu de Manampontsy lire Antanambao Manampontsy.

<u>Région d'Androy :</u>

- au lieu de Ambovombe-Androy lire Ambovombe,

[46] Note de lecture : Antsiranana-II a été remplacé par Andranovondronina dans la table de données de population.

- au lieu de Ambovombe-Androy lire Anjeky Ankilikira,
- au lieu de Ambovombe-Androy lire Antanimora Atsimo,
- au lieu de Ambovombe-Androy lire Andoharano Ambinagny,
- au lieu de Bekily lire Bekily-Centrale,
- au lieu de Bekily lire Maroviro,
- au lieu de Beloha lire Tranoroa
- au lieu de Tsihombe lire Ankilivalo.

D'autres différences (oubli d'un tiret, d'une majuscule, lettre différente) sont également trouvées dans la suite et empêchent la jointure des deux fichiers. Elles sont modifiées et répertoriées dans la partie b) des résultats (1.5 Mise en place de la méthode et exemples de présentation dans les régions d'Itasy et de Diana) pour chacune des régions.

1.3.2 Le tracé des limites administratives de la population malgache

Alors que les décomptes de population ont été rendus disponibles assez rapidement, ni l'institut national de la statistique malgache (INSTAT) ni aucune autre instance du gouvernement malgache n'a publié pour le moment les tracés administratifs correspondant au découpage utilisé lors du RGPH-3.

On est dans la situation relativement fréquente évoquée en introduction, où aucun fichier géoréférencé correspondant exactement aux entités administratives du recensement n'est actuellement disponible. En fait, le fichier géoréférencé le plus récent disponible date du 31 octobre 2018 et est produit par les Nations Unies, plus précisément par le bureau de la coordination des affaires humanitaires (OCHA, 2018a) et est disponible en juin 2024, sur la plateforme *The Humanitarian Data Exchange* (HDX). Bien que ce fichier couvre l'ensemble du territoire, seulement 1579 des 1693 communes décomptées en 2018 sont répertoriées. Voici en détail (tableau 3), le nombre de communes qu'il manque pour chacune des régions retenues :

Tableau 3 : Récapitulatif du nombre de communes recensées en 2018 et du nombre de communes disponibles dans le fichier géoréférencé (OCHA), pour en déduire le nombre de communes manquantes pour chaque région

Région	Nombre de communes recensées en 2018 (RGPH3)	Nombre de communes disponibles dans le fichier géoréférencé (OCHA)	Communes manquantes
Itasy	53	51	2
Diana	71	65	6
Melaky	41	37	4
Atsinanana	95	88	7
Analamanga	145	138	7[47]
Androy	58	51	7

Sources : Nombre de communes issus des résultats du RGPH3 (INSTAT) et des tracés des limites administratives (OCHA)

En ne gardant que les six régions sélectionnées, le fichier géographique comprend 430 communes alors qu'il devrait en comporter 463. Il y a donc 33 communes non répertoriées et cet état correspond à la situation administrative précédant la loi de 2015. L'objectif de lier les décomptes de population des 463 communes à leur position géographique sur la carte se révèle impossible en l'état. Il est nécessaire de légèrement modifier les deux fichiers pour rendre possible la jonction de ces informations.

La section suivante propose une démarche permettant de cartographier les résultats du RGPH-3 au niveau communal, notamment les décomptes de population, en dépit de l'absence d'une base de données géoréférencée actualisée des limites administratives de Madagascar. Comme aucune source officielle n'indique comment certaines communes ont été découpées ou regroupées pour créer les nouvelles communes, nous ne pouvons pas retravailler le tracé des limites administratives pour qu'il corresponde au découpage utilisé lors du RGPH-3. Nous mettons donc en place une méthode qui va légèrement modifier la base de données du recensement, puis le fichier géoréférencé communal en identifiant les communes filles et les communes mères afin de répartir correctement la population dans le tracé disponible.

Bien que cette méthode ne soit appliquée qu'à six régions, elle est parfaitement reproductible pour les 16 régions restantes ou encore pour d'autres pays. Les points 1.4 à 1.6 suivants abordent la méthode spécifique que nous avons mise en œuvre puis

[47] En réalité, ce fichier comprend 139 communes pour Analamanga mais comme 2 communes sont fusionnées (voir partie suivante), on passe à 138 communes.

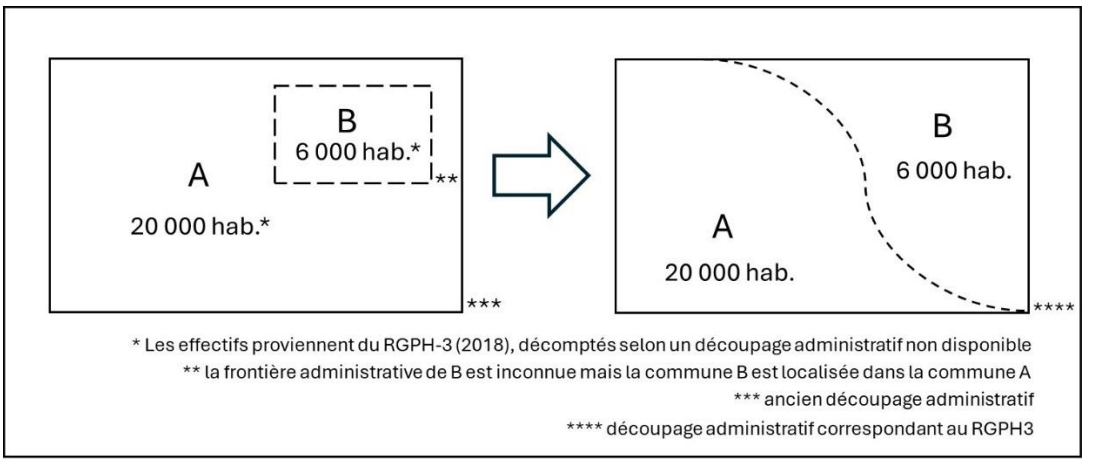
en 1.7 des indicateurs de cohérence sont présentés de manière à confirmer la validité de notre approche.

1.4 Méthodologie permettant la jointure des deux informations

Selon la loi n°2015-002 promulguée à Antananarivo, le 26 février 2015, le nombre des communes est modifié « en créant 149 nouvelles communes par l'éclatement de certaines communes mères ou par la fusion des communes existantes ou par le changement des chefs lieux, reparties dans 91 districts ». Cette loi garantit le fait qu'une nouvelle commune est créée par l'éclatement d'une commune mère ou par la fusion de communes existantes. A noter que lorsqu'une commune mère est divisée, son nom est transmis à une des deux communes filles. Au sein des 6 régions considérées ici, 33 communes se trouvent dans la première catégorie et une seule dans la seconde.

Pour la première situation, aucune source officielle n'indique comment la commune mère a été divisée pour laisser de la place à une nouvelle commune appelée « commune fille ». Il est dès lors impossible de retravailler les tracés des limites administratives (comme le suggère le scénario de la figure 20) afin de parvenir à un fichier géoréférencé contenant les 463 communes correspondant à l'actualisation des tracés pour l'année du recensement en 2018.

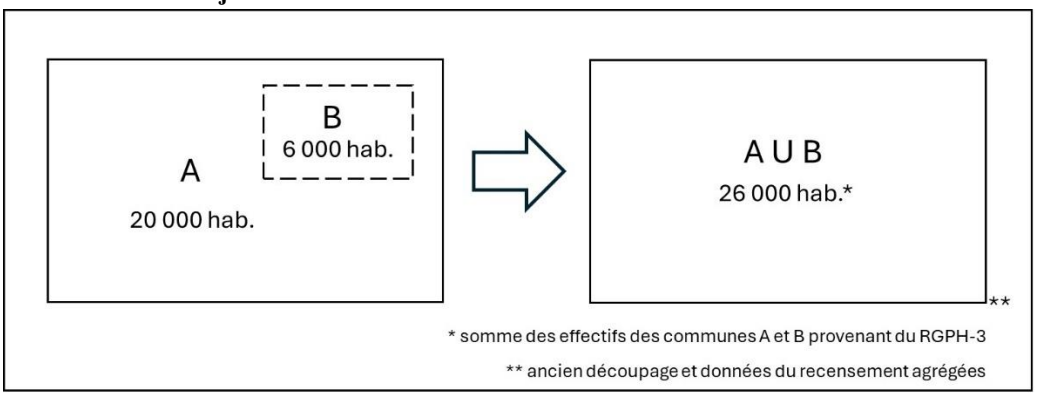
Figure 20 : Scenario idéal représentant la probable division de la commune A en communes A et B.



La solution proposée est de garder le fichier géoréférencé des 430 communes tel quel et de chercher, pour chacune des 33 nouvelles communes manquantes, leur localisation géographique qui permet de confirmer de quelle commune mère elle provient. Ensuite, il suffit d'additionner la population de la nouvelle commune et de

la commune qui a gardé le nom de la commune mère pour trouver la population de la commune mère (Figure 21).

Figure 21 : Proposition de solution : la commune B étant localisée dans la commune A, on ajoute à la commune A les 6000 habitants de la commune B.



Pour autant, aucune source officielle ne donne l'information de la localisation des communes de Madagascar et seulement deux sources notables apportent des informations : wikipedia.org d'une part et une base de données ouverte sur les minéraux, les roches et les météorites (mindat.org). Il est fréquent que les nouvelles communes n'aient pas de page Wikipédia et il faut, pour obtenir une localisation, souvent se fier à mindat.org.

Avec cette méthode, 28 des 33 communes ont pu être localisées. Les localisations de cinq communes demeurent plus complexes à déterminer et ce pour deux raisons principales : soit plusieurs communes filles avec le même nom sont trouvées dans différentes communes mères potentielles, soit aucune information ne nous permet de confirmer une localisation précise. Nous revenons dans une partie ultérieure (section 1.6) sur ces 5 communes en présentant les hypothèses effectuées pour déterminer leur localisation finale.

Par ailleurs, seule la commune d'Ivato dans la région d'Analamanga résulte d'une fusion de communes existantes et il faut dans ce cas précis, modifier le fichier géoréférencé afin de fusionner le contour des deux communes d'Ivato Aéroport et Ivato Firaisana.

Pour finir, les noms de 5 communes qui ont évolué au cours du temps sont corrigés, cette fois-ci dans le fichier fournit par OCHA afin qu'ils correspondent aux noms officiels du RGPH-3.

1.5 Mise en place de la méthode et exemples de présentation dans les régions d'Itasy et de Diana

1.5.1 Itasy

Dans la région d'Itasy, les décomptes de population de deux communes ont ainsi été modifiés (changements visibles en carte dans la figure 22) et une commune a été renommée.

a) Modifications des décomptes de population

D'après la loi n°2015-002 deux nouvelles communes ont été créées à Itasy : Talata Tsimadilo dans le district d'Arivonimamo et Ambohidanerana dans le district de Soavinandriana. Les sites wikipedia.org et mindat.org permettent d'obtenir des coordonnées géographiques et ainsi la localisation supposée de ces deux communes.

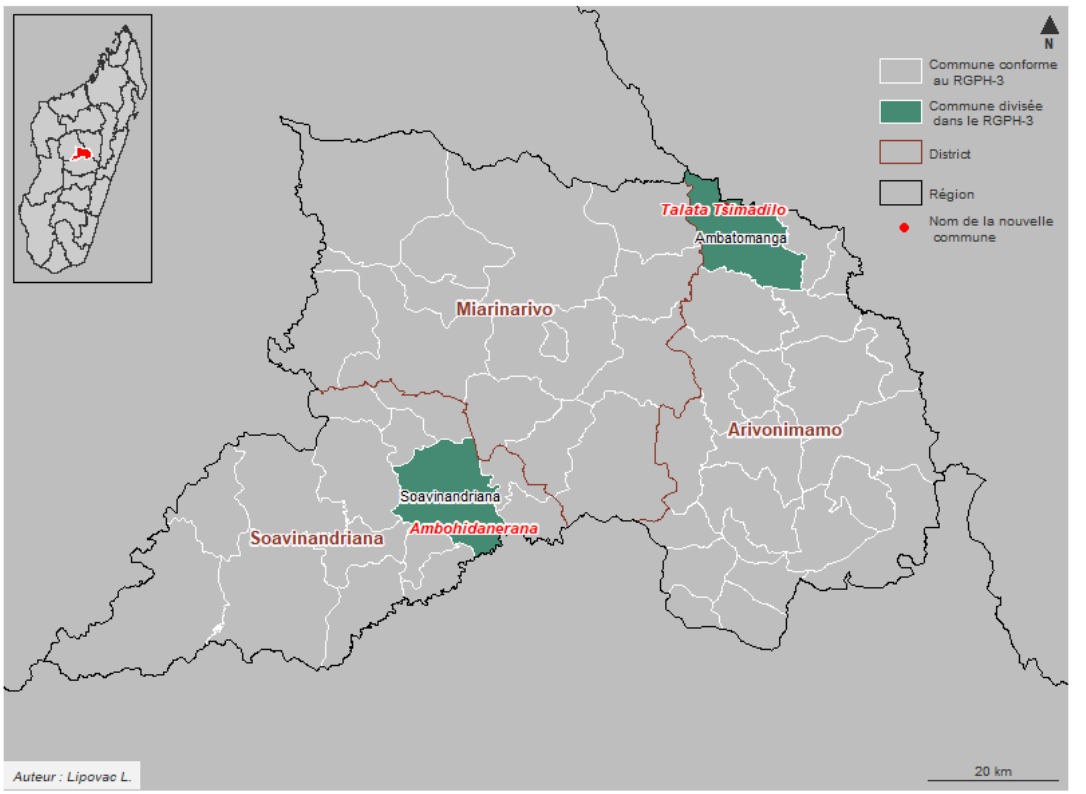
Talata Tsimadilo est localisée dans la commune mère nommée Ambatomanga (Mindat, 2024a), en haut à droite sur la figure 22, ce qui permet de considérer qu'elle provient effectivement de la division d'Ambatomanga. Dès lors, nous additionnons les 7 438 habitants décomptés à Talata Tsimadilo par le RGPH-3 aux 9 286 habitants de la commune d'Ambatomanga. La commune d'Ambatomanga (commune mère) compte désormais 16 724 habitants (9 286 +7 438).

La commune d'Ambohidanerana se trouve quant à elle dans la commune de Soavinandriana (Mindat, 2024b), qui porte le même nom que le district. De la même manière, il est décidé d'ajouter à Soavinandriana les 10 786 habitants d'Ambohidanerana pour obtenir la population de la commune mère, appelée elle aussi Soavinandriana.

b) Modifications apportées au fichier des limites administratives

La commune nommée Rambolamasoandro Andranomiely est renommée Andranomiely dans le fichier géoréférencé OCHA afin de correspondre aux résultats du RGPH-3.

Figure 22 : Localisation des nouvelles communes de la région d'Itasy



Sources : Carte réalisée par l'auteur à partir des tracés des limites administratives fournis par OCHA et localisation des nouvelles communes par Mindat

1.5.2 Diana

Dans la région de Diana, six nouvelles communes ont été créées (noms en rouge sur la figure 23). Ensuite deux communes ont été renommées dans la partie b).

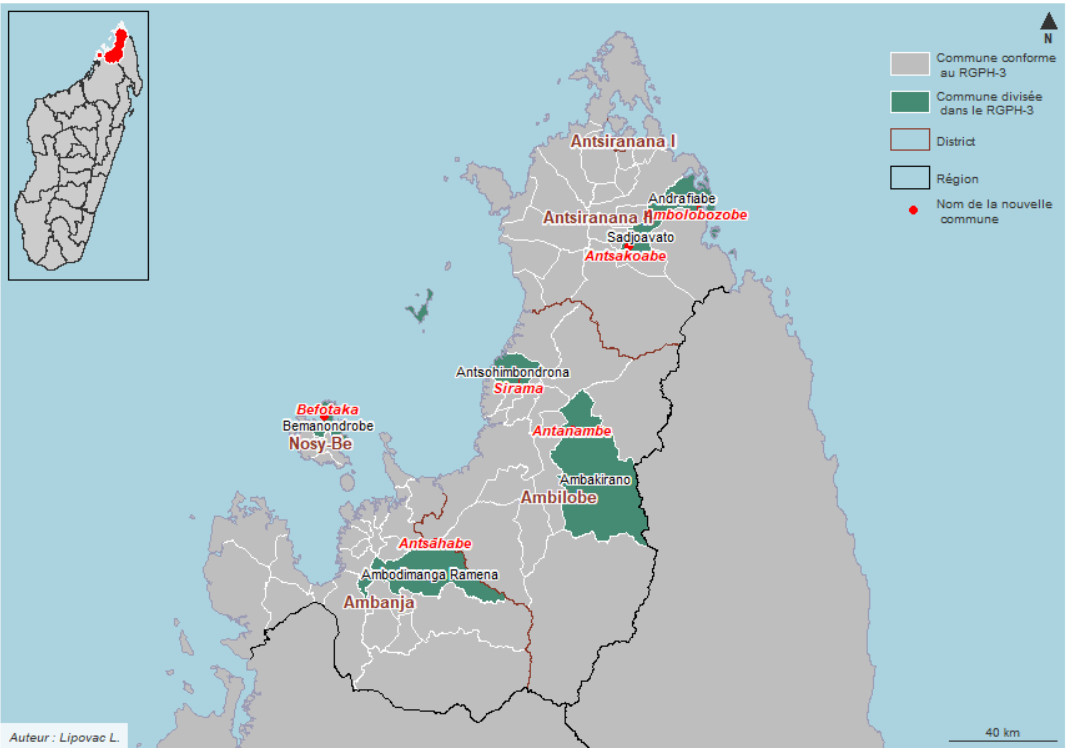
a) Modifications des décomptes de population

La commune d'Antsahabe* (district d'Ambanja) est localisée à Ambodimanga Ramena, les communes d'Antanambe et de Sirama (district d'Ambilobe) sont localisées respectivement à Ambakirano (Mindat, 2024c) et Antsohimbondrona (Wikipedia, 2023b). Les communes d'Antsakoabe et d'Ambolobozobe (district d'Antsiranana II) se trouvent respectivement à Sadjoavato (Wikipedia, 2023c) et Andrafiabe (Mindat, 2024d) et pour finir la commune de Befotaka (district de Nosy-Be) est localisée à Bemanondrobe (Mindat, 2024e).

Dès lors, nous additionnons les 6 944 habitants décomptés à Antsahabe par le RGPH-3 à la commune d'Ambodimanga Ramena, les 8 187 habitants d'Antanambe et les 22 966 habitants de Sirama respectivement aux communes d'Ambakirano et d'Antsohimbondrona, les 2 257 habitants d'Antsakoabe et les 3 328 habitants

d'Ambolobozobe respectivement aux communes de Sadjoavato et d'Andrafiabe et pour finir les 5 006 habitants de la commune de Befotaka à la commune de Bemanondrobe.
Par ailleurs, la commune Hell Ville a été corrigée en Hell-Ville dans la base de données du recensement, le tiret ayant été omis.

Figure 23 : Localisation des nouvelles communes de la région de Diana



Sources : Carte réalisée par l'auteur à partir des tracés des limites administratives fournis par OCHA et localisation des nouvelles communes par Mindat et Wikipédia

b) Modifications apportées au fichier des limites administratives

La commune nommée Diego Suarez est renommée Antsiranana I dans le fichier géoréférencé OCHA afin de correspondre aux résultats du RGPH-3.

*Antsahabe fait partie des 5 communes plus complexes à localiser, la partie suivante leur est réservée.

Les localisations des nouvelles communes dans les quatre régions restantes sont présentées en annexe.

1.6 Méthode supplémentaire pour les communes difficiles à localiser

La localisation des communes d'Antsahabe (district d'Ambanja dans la région de Diana), Bemara Atsinanana (district d'Antsalova dans la région de Melaky), Maromitety I (district de Marolambo dans la région Atsinanana), Andragnanivo (district d'Ambovombe-Androy dans la région d'Androy) et Ankilivalo (district de Tsihombe dans la région d'Androy) s'est révélée particulièrement complexe. Ces communes ne figurent pas dans les sources consultées ou, lorsqu'elles y apparaissent, elles sont mentionnées plusieurs fois en raison d'homonymies. Plusieurs tentatives ont été menées sans succès, notamment en s'appuyant sur les registres d'écoles primaires – qui portent souvent le nom de leur localité – ou en sollicitant des experts et autorités locales. Face à ces difficultés, nous avons mis en place une dernière approche combinant deux méthodes distinctes.

1.6.1 Utilisation du niveau administratif des fokontany

Il existe à Madagascar une subdivision encore plus précise qui correspond au niveau administratif quatre et qui est appelé le fokontany. Ils ont été depuis le référendum de 2007 réintroduits comme unité de base des collectivités locales (Bidou et al., 2008) et comme ils sont le premier niveau de concertation et de discussion ils peuvent jouer des rôles clés dans l'organisation de certaines communes et ainsi être facilement localisables. Si l'on trouve plusieurs des fokontany de la nouvelle commune dans une seule et même commune bien connue, on pourra définitivement la considérer comme la commune mère. Nous nous référons pour cela au document produit par le Conseil National de la Législation malgache (CNLEGIS) qui permet de connaître les fokontany de chacune des communes (CNLEGIS, 2018).

1.6.2 Comparaison des décomptes aux projections de populations

L'approche est complétée par un critère de vérification que nous chercherons ensuite à généraliser dans la partie 1.7 suivante à l'ensemble des communes. Cela consiste à comparer les décomptes officiels de population à des projections de population pour la commune mère, pour l'année 2018.

En raison de l'éclatement de certaines communes mères en 2015 nous avons, pour reconstituer leurs effectifs en 2018, décidé d'additionner les effectifs de population des communes filles qui se trouvent à l'intérieur de celles-ci. Les

communes mères ont donc un décompte de population qui est modifié, ou plus précisément, cumulé avec le décompte de la (ou des) commune(s) fille(s).

En plus du fichier géoréférencé des 1579 communes de Madagascar, OCHA[48] fournit une autre table indiquant des projections de population pour 2018 pour chacune des communes (OCHA, 2018b). Ces projections de population ont été produites par l'INSTAT et le Bureau National de Gestion des Risques et des Catastrophes[49] (BNGRC) à partir de données de population datant de 2009 (même si ce n'est pas explicitement spécifié, on peut faire l'hypothèse que ces données ont elles mêmes été calculées avec des projections à partir du recensement de 1993) puis calculées en appliquant des taux de croissance légèrement décroissants, mais uniformes dans tout le pays (2,76%, 2,75%, 2,74%, 2,72%, 2,71% et 2,7% pour les années 2015 à 2018).

Ces projections sont les seuls éléments disponibles pour valider notre démarche et nous allons donc les utiliser. Elles sont néanmoins questionnables, pour deux raisons principales. D'une part, du fait de l'espacement de 25 ans entre la date considérée (2018) et les données utilisées pour les construire (1993). En effet, plus on s'éloigne du recensement initial, plus le décalage entre valeurs projetées et valeurs réelles s'amplifie. D'autre part, les migrations internes importantes qu'a connues Madagascar entre 1993 et 2018 ne sont pas prises en compte. Or, parmi les habitants recensés en 2018, 5 millions d'habitants, soit 20,5 % de la population, ont déclaré avoir changé de district (ou de pays) depuis leur naissance (INSTAT, 2021a). Cette dynamique varie considérablement selon les régions. Dans les trois des quatre régions les plus urbanisées du pays, à savoir Analamanga, Diana et Atsinanana (INSTAT, 2021f), on observe des soldes migratoires fortement positifs, avec plus de 350 000 habitants de plus pour la première – correspondant à 9,6 % de la population résidente –, et environ 85 000 de plus pour les deux dernières représentant respectivement un solde migratoire de 9,5 % et 5,9 % de la population résidente pour Diana et Atsinanana. Ce sont des zones qualifiées de régions d'immigration, sans doute dû aux opportunités disponibles en termes d'emploi, d'éducation ou encore de santé. Itasy, de part sa proximité avec la région de la capitale à un solde migratoire de -49 000 habitants représentant 5,4 % de sa population et Androy est la 3ème région avec le plus fort solde migratoire négatif avec près de -82 000 habitants représentant 9,1 % de sa population. Cette situation s'explique par certaines difficultés environnementales et socio-économiques : Androy est une région aride, régulièrement touchée par des sécheresses qui fragilisent son économie, majoritairement basée sur l'agriculture et

[48] https://data.humdata.org/dataset/cod-ab-mdg

[49] https://bngrc.gov.mg/

l'élevage, poussant notamment les jeunes actifs à partir vers d'autres régions. Enfin, Melaky est relativement stable avec un solde de +4 000 habitants représentant 1.5 % de sa population.

De fait, ces projections de population pour 2018 demeurent une source importante renseignant sur la croissance attendue de la population de chacune des communes. Nous les utilisons afin de guider notre démarche. Par exemple, il semblera tout simplement incohérent d'ajouter de la population à une commune mère qui a d'ores et déjà largement dépassé sa projection de population. Cette méthode peut servir d'aide à la vérification et nous permet de savoir s'il semble logique de modifier le décompte d'une commune mère en ajoutant la population d'une nouvelle commune. Nous appliquons cette approche aux cinq communes précédemment citées afin de déterminer leur localisation finale.

a) Antsahabe (région Diana et district d'Ambanja)

Selon les informations trouvées sur mindat.org, la commune d'Antsahabe pourrait se situer dans deux communes du même district, à savoir Anorontsangana (Mindat, 2024f) ou Ambodimanga Ramena (Mindat, 2024g). Antsahabe a quatre fokontany mais aucun des quatre n'est localisable. Afin de décider laquelle de ces deux communes a été divisée pour accueillir Antsahabe, on compare tout d'abord le décompte de la commune mère à son estimation fournie par OCHA. Puis on compare de nouveau à l'estimation OCHA mais cette fois-ci avec le décompte cumulé, c'est-à-dire la population de la commune mère additionnée à la population d'Antsahabe, la commune fille.

Tableau 4 : Déterminer la commune mère d'Antsahabe

Commune mère potentielle	Estimation par projection (OCHA)	Décompte par recensement officiel (RGPH-3)	Différence entre décompte officiel et l'estimation par projection	Cumul de population avec la nouvelle commune à affecter (+6944 hab.)	Différence entre le cumul et l'estimation
Anorontsangana	10725	10227	-4.9 %	17171	+37.5 %
Ambodimanga Ramena	13265	5706	-132.5 %	12650	-4.9 %

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 et des projections de population calculées par l'INSTAT et le BNGRC, fournis par OCHA

D'après le tableau 4, Anorontsangana a un décompte de la population issu du RGPH-3 de 10 227 habitants (colonne 3). Ce décompte officiel est très proche de l'estimation calculée par OCHA (colonne 2) où on retrouve seulement 498 habitants d'écart (soit 5 % de différence – colonne 4). Si on décide d'ajouter les 6 944 habitants d'Antsahabe

à Anorontsangana on a une différence finale entre le cumul et l'estimation de plus de 37,5%. Cette différence montre qu'il ne serait à priori pas logique d'ajouter les habitants d'Antsahabe à Anorontsangana et donc qu'Anorontsangana ne serait pas la commune mère d'Antsahabe.

En revanche, le décompte de population d'Ambodimanga Ramena est très inférieur à la projection de 2018. Dans ce cas, si l'on décide d'ajouter 6944 habitants d'Antsahabe à la commune d'Ambodimanga Ramena, on passe d'une différence de -132.5 % (colonne 4) à une différence de -4,9 % (colonne 6). Avec ces critères, il semble bien plus logique de considérer qu'Ambodimanga Ramena est la commune mère d'Antsahabe.

b) Bemara Atsinanana (région de Melaky et district d'Antsalova)

Nous n'avons trouvé aucune information permettant de localiser Bemara Atsinanana. Un nom proche (Bemaraha Atsinanana) est situé dans la commune de Bekopaka mais ce n'est pas suffisant pour déterminer que c'est la commune mère. Parmi les 4 fokontany de Bemara Atsinanana, trois sont localisables et se situent dans la commune d'Antsalova (figure 74 – Annexe 2). De plus, comme le montre le tableau ci-dessous, il est beaucoup plus logique d'additionner la population dans la commune d'Antsalova.

Tableau 5 : Déterminer la commune mère de Bemara Atsinanana

Commune mère potentielle	Estimation par projection (OCHA)	Décompte par recensement officiel (RGPH-3)	Différence entre décompte officiel et l'estimation	Cumul de population de la nouvelle commune à affecter (+2610 hab.)	Différence entre le cumul et l'estimation
Bekopaka	12792	13223	3.3 %	15833	19.2 %
Antsalova	24457	14095	-73.5 %	16705	-46.4 %

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 et des projections de population calculées par l'INSTAT et le BNGRC, fournis par OCHA

c) Maromitety I (région Atsinanana et district de Marolambo)

D'après le site mindat.org, Maromitety I pourrait être placée dans 4 communes différentes du district de Marolambo (Marolambo, Androrangavola, Andonabe Sud, Betampona). Maromitety I a 9 fokontany mais aucun n'est localisable avec nos ressources. En fonctionnant de la même manière que précédemment et d'après le tableau 6, il semble illogique d'ajouter 10757 habitants aux communes de Marolambo, Androrangavola ou Betampona. On ajoute donc cette population à Andonabe Sud.

Tableau 6 : Déterminer la commune mère de Maromitety I

Commune mère potentielle	Estimation par projection (OCHA)	Décompte par recensement officiel (RGPH-3)	Différence entre décompte officiel et l'estimation	Cumul de population de la nouvelle commune à affecter (+10757 hab.)	Différence entre le cumul et l'estimation
Marolambo	24883	27509	9,5 %	38266	35,0 %
Androrangavola	9490	9542	0,5 %	20299	53,2 %
Andonabe Sud	23303	10503	-121,9 %	21260	-9,6 %
Betampona	15498	13466	-15,1 %	24223	36,0 %

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 et des projections de population calculées par l'INSTAT et le BNGRC, fournis par OCHA

d) Andragnanivo (région Androy et district de Ambovombe-Androy)

Andragnanivo n'est pas localisable sous ce nom ni sur mindat.org, ni sur Wikipédia, mais on peut trouver un nom similaire dans la commune de Jafaro sous l'entité « CR Andragnanivo ». Trois de ses fokontanys sont également localisés à l'intérieur de Jafaro. De plus, ajouter 3386 habitants à cette commune semble logique car ce nouvel effectif reste inférieur à la projection de 2018 (Tableau 7).

Tableau 7 : Déterminer la commune mère d'Andragnanivo

Commune mère potentielle	Estimation par projection (OCHA)	Décompte par recensement officiel (RGPH-3)	Différence entre décompte officiel et l'estimation	Cumul de population de la nouvelle commune à affecter (+3386 hab.)	Différence entre le cumul et l'estimation
Jafaro	40147	30319	-32,4 %	33705	-19,1 %

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 et des projections de population calculées par l'INSTAT et le BNGRC, fournis par OCHA

e) Ankilivalo (région Androy et district de Tsihombe)

Toujours selon le site mindat.org, Ankilivalo peut se situer soit dans la commune d'Imongy soit à Tsihombe. La majorité des fokontany se trouvent dans la commune de Tsihombe et la comparaison entre les deux cumuls semble également indiquer qu'il est plus probable que Ankilivalo se situe à Tsihombe.

Tableau 8 : Déterminer la commune mère d'Ankilivalo

Commune mère potentielle	Estimation par projection (OCHA)	Décompte par recensement officiel (RGPH-3)	Différence entre décompte officiel et l'estimation	Cumul de population de la nouvelle commune à affecter (+5714 hab.)	Différence entre le cumul et l'estimation
Tsihombe	39043	36873	-5,9 %	42587	+8,3 %
Imongy	14194	14874	+4,6 %	20588	+31,1 %

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 et des projections de population calculées par l'INSTAT et le BNGRC, fournis par OCHA

1.7 Récapitulatif et ultime vérification

Les 33 nouvelles communes créées par l'éclatement de communes mères lors de l'entrée en vigueur de la loi n°2015-002 sont récapitulées dans le tableau 9. Dans ce tableau, nous poursuivons la comparaison entre les décomptes de population officiels du RGPH-3 pour les 32 communes mères (au lieu de 33, car la commune de Tsihombe à engendré trois communes filles au lieu de deux pour toutes les autres) et les estimations de population fournies par OCHA. Nous cherchons maintenant à déterminer si cette méthode est généralisable à l'ensemble du jeu de données, et si une tendance émerge lorsque l'on compare ces résultats avec le reste des communes dont le décompte n'a pas été modifié.

Tableau 9 : Récapitulatif et vérification des localisations de chacune des nouvelles communes

Commune fille	Commune mère potentielle	Estimation par projection (OCHA)	Décompte par recensement officiel (RGPH-3)	Différence entre décompte officiel et l'estimation	Cumul de population de la nouvelle commune à affecter	Différence entre le cumul et l'estimation
Région d'Itasy						
Talata Tsimadilo	Ambatomanga	16891	9286	-81,9 %	16724	-1,0 %
Ambohidanerana	Soavinandriana	45200	40045	-12,9 %	50831	11,1 %
Région de Diana						
Antsahabe	Ambodimanga Ramena	13265	5706	-132,5 %	12650	-4,9 %
Antanambe	Ambakirano	25272	11391	-121,9 %	19578	-29,1 %
Sirama	Antsohimbondrona	30823	9168	-236,2 %	32134	4,1 %
Antsakoabe	Sadjoavato	8537	5906	-44,5 %	8163	-4,6 %
Ambolobozobe	Andrafiabe	6975	4016	-73,7 %	7344	5,0 %
Befotaka	Bemanondrobe	8122	4198	-93,5 %	9204	11,8 %
Région de Melaky						
Makaraingo	Sarodrano	10867	7057	-54,0 %	12972	16,2 %
Bemara Atsinanana	Antsalova	24457	14095	-73,5 %	16705	-46,4 %
Antsirasira	Marovoay Sud	11733	6945	-68,9 %	8780	-33,6 %
Antranokoaky	Andramy	7724	4888	-58,0 %	10124	23,7 %
Région d'Atsinanana						
Manaratsandry	Mahela	13920	8815	-57,9 %	14648	5,0 %
Antsapananа	Mahatsara	27120	11921	-127,5 %	25459	-6,5 %

Andranamb omaro	Ambodihari na	26677	13092	-103,8 %	27602	3,4 %
Maromitety I	Andonabe Sud	23303	10503	-121,9 %	21260	-9,6 %
Satrandroy	Fito	18360	8744	-110,0 %	15409	-19,2 %
Ifasina Iii	Ifasina II	7270	5101	-42,5 %	8011	9,2 %
Ambodinon oka Rangalana	Ampasimad inika	10814	4093	-164,2 %	10548	-2,5 %
Région d'Analamanga						
Anjoma Faliarivo	Alatsinainy Bakaro	36709	20370	-80,2 %	27335	-34,3 %
Moraranoso afiraisana	Tankafatra	12864	6867	-87,3 %	13385	3,9 %
Andranomis a	Amparatanj ona	9838	7061	-39,3 %	14753	33,3 %
Andranomie ly Sud	Talata Angavo	14991	9240	-62,2 %	14624	-2,5 %
Mangasoavi na	Tsaramasoa ndro	13277	9747	-36,2 %	16303	18,6 %
Ambohidrab iby	Talata Volonondry	28734	18314	-56,9 %	26187	-9,7 %
Anosy Avaratra	Sabotsy Namehana	72022	57363	-25,6 %	72606	0.8 %
Région d'Androy						
Andoharano Ambinagny	Analamary	11715	5854	-100,1 %	12913	9,3 %
Andragnani vo	Jafaro	40147	30319	-32,4 %	33705	-19,1 %
Mikaikarivo Ambatomai nty	Bekitro	27148	30776	11,8 %	35832	24,2 %
Ambatotsiv ala	Beloha	35657	33349	-6,9 %	40668	12,3 %
Maheny	Tranoroa	32693	21414	-52,7 %	29863	-9,5 %
Ankilivalo	Tsihombe	39043	28709	-36,0 %	42587	8.3 %

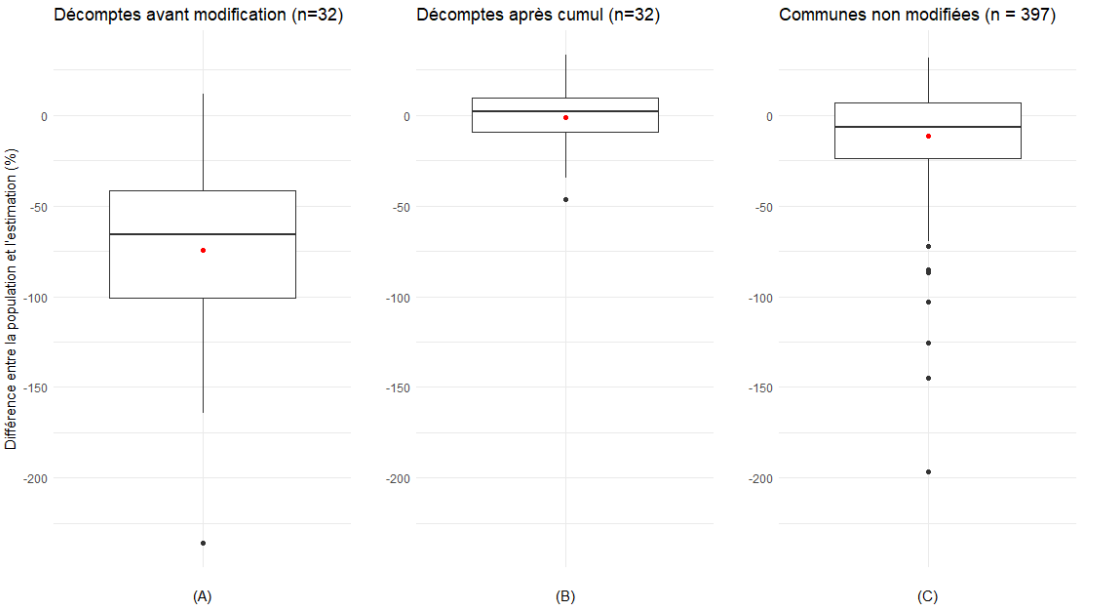
Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 et des projections de population calculées par l'INSTAT et le BNGRC, fournis par OCHA

Parmi les 32 communes mères, 31 affichent un déficit de population par rapport à l'estimation fournie par OCHA (colonne 5 du tableau 9). Ce constat laisse supposer une sous-estimation de la population dans ces communes, ce qui justifierait l'ajout d'habitants par le biais des nouvelles communes filles créées.

Afin d'appuyer cette première observation, nous avons représenté dans la figure 24 la distribution des effectifs des communes mères avant et après ajout de la population des communes filles (boites à moustaches A et B représentant respectivement la colonne 5 et 7 du tableau précédent ; figure 24) ainsi que la

distribution des effectifs de population pour les communes ne nécessitant pas de modifications (boite à moustaches C, figure 24). On observe avec les deux premières boites à moustaches que passe d'une distribution presque exclusivement négative à une distribution présentant une allure presque centrée et symétrique, et dont la variabilité est beaucoup réduite. Notamment, on passe d'une moyenne de -74.5 à une moyenne très proche de zéro après ajout de la population.

Figure 24 : Distribution des effectifs des communes avant et après modification, et des communes ne nécessitant pas de modifications



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 et des projections de population calculées par l'INSTAT et le BNGRC, fournis par OCHA

Nous confirmons finalement ces observations en analysant la dernière boîte à moustaches C, qui indique la distribution pour le reste des communes n'accueillant pas de communes filles. La distribution semble similaire à celle de la boîte à moustaches B, avec une allure elle aussi normale – bien qu'un peu influencée par les valeurs négatives –, ce que confirment la moyenne et la médiane, respectivement égales à -10 et -6. Le test de Kolmogorov-Smirnov, qui compare la forme des distributions, donne une p-valeur de 0.04473, suggérant une différence statistiquement significative entre les erreurs de projection des groupes B et C. Toutefois, cette p-valeur étant très proche du seuil de 0.05, la distinction entre ces deux groupes reste modérée. Visuellement, la distribution du groupe B est nettement améliorée par rapport à celle du groupe A et se rapproche de celle du groupe C,

confirmant ainsi que l'ajustement (l'ajout de la population d'une commune fille) a eu l'effet souhaité.

Bien qu'ils ne permettent pas nécessairement de valider les choix, ces éléments apportent crédit et logique à la méthode globale. Plus important encore, cela permet en l'absence d'autres alternatives, de créer la base de données géoréférencée finale des 430 communes pour les 6 régions étudiées et donc de pouvoir mettre en œuvre le plan initial de désagrégation présenté en introduction. La partie suivante présente la création des variables au niveau communal, dernière étape essentielle en vue de la préparation de la modélisation de la population des communes de Madagascar.

1.8 Création des variables au niveau des communes

Pour les données climatiques (précipitations et température), l'altitude, et l'intensité de la lumière nocturne, nous utilisons les mêmes rasters que ceux employés pour calculer les valeurs au niveau des shapes. Ensuite, il faut calculer des statistiques zonales. Dans QGIS, la fonction « zonal statistics » permet de calculer la moyenne des valeurs pour chaque variable dans chaque commune. Sur R, la fonction « extract » du package raster (Hijmans et al., 2023) permet d'extraire les valeurs de température et de précipitations pour chaque commune, avant d'appliquer la moyenne pour obtenir la variable finale.

Pour obtenir les variables de distance à certaines entités (routes, points d'eau, centres de santé, grandes agglomérations) au niveau communal, on calcule la moyenne des valeurs associées à chacune des shapes pour la commune. Afin d'avoir une estimation de la surface bâtie de chaque commune, on calcule la proportion de la surface bâtie en divisant la somme des surfaces bâties par la surface totale de la commune. Enfin, l'inventaire des variables est complété par le nombre moyen de personnes par ménage (INSTAT, 2021g).

En plus du niveau communal, les niveaux administratifs précédents (ADM0_EN pour le niveau national, ADM1_EN pour le niveau régional et ADM2_EN pour le niveau district) sont récupérés. Associés à leur population, ils seront utiles pour la cartographie et la vérification des résultats.

2. Cartographie des zones bâties (les zones cibles)

En 2018, année du recensement à Madagascar, les algorithmes de TeleCense ont identifié un total de 412 793 shapes, dont la répartition par région est visible le tableau 10. Sans surprise, la région de la capitale, Analamanga, compte le plus grand nombre de zones bâties, avec près de 200 000 shapes. Les régions de Diana, Androy, Itasy et Atsinanana présentent une répartition semblable, avec entre 38 000 et 66 000 shapes chacune. En revanche, la région de Melaky ne compte que 17 000 shapes, confirmant son statut de région vaste mais peu densément peuplée.

Tableau 10 : Nombre de shapes par régions

Analamanga	Diana	Androy	Itasy	Atsinanana	Melaky
194 421	38 456	46 875	50 170	65 648	17 223

Sources : TeleCense

Après avoir identifié les shapes, on leur attribue les caractéristiques détaillées dans le chapitre précédent (section 2.1 – chapitre 2). Il s'agit de caractéristiques climatiques (précipitations et température), de variables de surface (celle identifiée par TeleCense, celle pondérée par l'identification Google), d'altitude, de densité de bâti, de distance à des entités (routes, point d'eau, centre de santé, grande ville) et d'intensité de lumière la nuit.

Les calculs des variables de distance sont automatisés mais pas celle de la distance à la grande ville ; pour cela nous identifions d'abord les grandes agglomérations de Madagascar, définies comme celles dont la somme totale des shapes dépasse 1 km². Il y en a huit dans la région de Diana, deux à Melaky et à Androy, treize à Analamanga, et cinq à Atsinanana et Itasy. Ensuite, pour chaque shape, nous calculons la distance minimale à ces agglomérations. Il est à noter qu'une shape appartenant à une grande agglomération a une distance calculée de zéro mètre.

Pour finir, il faut associer géographiquement chacune des shapes à la zone administrative à laquelle elle appartient (de la même manière qu'une shape est reliée à une agglomération). Cela permet de situer chaque shape de telle sorte qu'elle appartienne désormais à une commune, un district et une région uniques. De plus, cela permet d'associer aux shapes les caractéristiques des communes à laquelle elles appartiennent, notamment les décomptes de population de la commune, ce qui nous permettra de désagréger les données par la suite. D'autres variables potentiellement utiles pour la modélisation, comme la proportion de bâti de la commune et le nombre moyen de personnes par ménage, peuvent également être obtenues. Pour ces

informations spécifiques à la commune, toutes les shapes d'une même commune auront la même valeur.

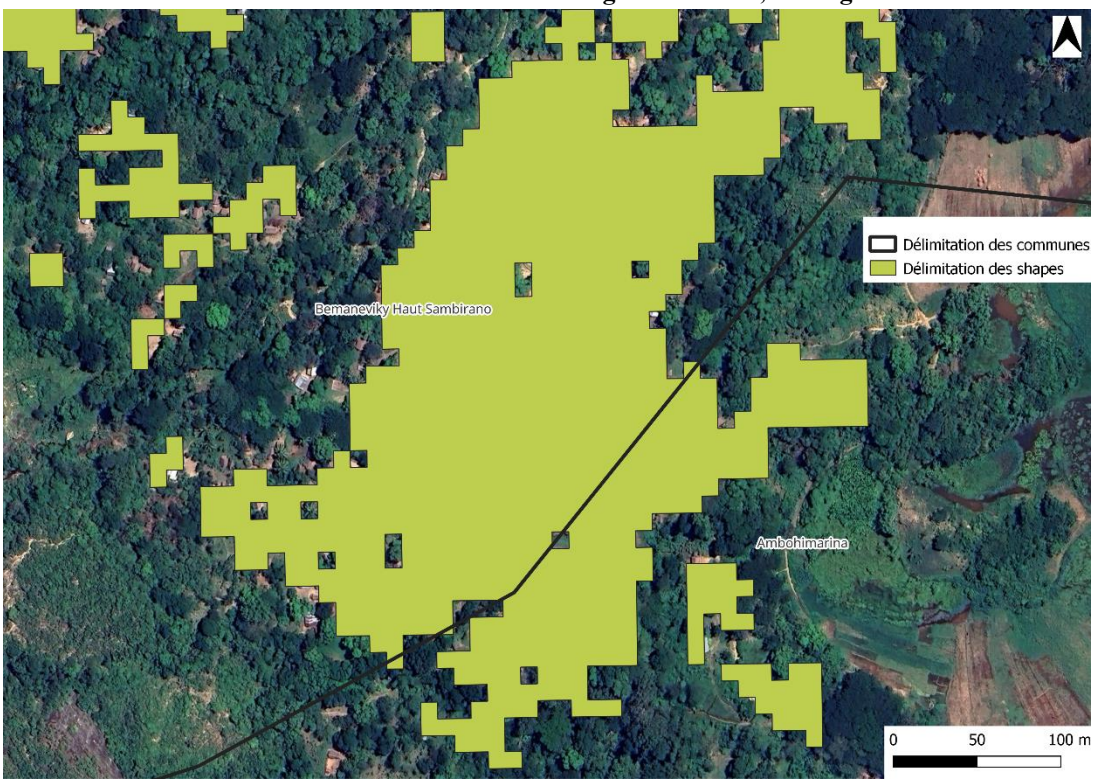
2.1 Gestion des shapes appartenant à plusieurs zones administratives

Afin de lier les shapes aux communes de Madagascar, un léger pré-traitement est cependant nécessaire. En effet, il arrive que certaines shapes soient à cheval sur deux zones administratives ou plus et cela peut se produire de deux manières :

1. Shapes de grande taille : Bien que moins fréquent dans notre jeu de données, les très grandes shapes ont plus de chances d'intersecter une ou plusieurs communes. Par exemple, dans la région de Diana, une shape intersecte les communes de Bemaneviky Haut Sambirano et d'Ambohimarina, comme illustré en figure 25.
2. Proximité des zones administratives : Dans certains endroits, comme dans le district d'Antananarivo Renivohitra de la capitale, les six communes ont une surface relativement petite et sont proches les unes des autres. Il arrive donc que même des shapes de petite taille intersectent deux communes.

Au total, 2 733 shapes intersectent deux communes, 44 en intersectent trois, et une en intersecte quatre simultanément. On remarque par ailleurs que ce phénomène est particulièrement marqué dans les villes, où les zones administratives peuvent être plus petites. En effet, 1 616 de ces shapes se situent dans la région d'Analamanga et 531 dans la région d'Atsinanana. Dès lors, comment traiter ces shapes particulières ?

Figure 25 : Exemple d'une shape intersectant les communes de Bemaneviky Haut Sambirano et Ambohimarina deux communes – Région de Diana, Madagascar



Sources : Identification du bâti par le programme TeleCense et découpage administratif issu d'OCHA – Image issue du logiciel QGIS

Dans l'ensemble, les shapes intersectant plusieurs communes sont relativement de petites surfaces, et il est assez rare de trouver des shapes de grande taille, comme celle mentionnée dans la figure précédente. Bien que cela puisse sembler marginal, nous avons décidé, afin de ne pas fausser la répartition de la population lors de la désagrégation, de diviser ces shapes pour qu'elles n'appartiennent plus qu'à une seule commune, technique conseillée parmi d'autres par Vignes et Rimbourg (2013). Pour cela, nous utilisons la commande st_intersection dans R, qui découpe la géométrie de la shape en autant de nouvelles shapes que nécessaire (par exemple, si une shape intersecte deux communes, cela crée deux nouvelles shapes). Cette commande a l'avantage de joindre directement les informations de la couche des communes. Cependant, comme le précise RStudio lors de cette opération avec le message « attribute variables are assumed to be spatially constant throughout all geometries », il est nécessaire de modifier ou de recalculer certaines variables qui ont pu changer avec la modification de la géométrie. Cela concerne notamment la variable d'identification et, plus important encore, les variables de surface. Pour la surface, il suffit d'utiliser la commande « st_area » qui va calculer la surface en fonction de la nouvelle géométrie. Il est à noter qu'avec ce type d'opération, il arrive que des shapes très petites soient créées et nous avons donc

décidé de supprimer les shapes dont la surface est inférieure à 30 m². C'étaient les seules variables à modifier, car pour les autres, on peut considérer qu'il s'agit de caractéristiques spatiales et non physiques, et donc qu'elles n'ont pas besoin d'être modifiées. Par exemple, l'altitude ne change quasiment pas, de même pour les données climatiques ou l'intensité de lumière la nuit, qui restent constantes malgré les modifications géométriques.

3. Etude descriptive des deux bases de données utilisées

Après le découpage, nous obtenons un total de 415 617 shapes réparties dans 430 communes et 6 régions. Afin de mettre en œuvre les différentes méthodes de l'approche descendante, il convient d'approfondir nos connaissances des caractéristiques des deux bases de données utilisées ici, notamment celle des communes qui est celle que nous modéliserons en premier afin de construire une variable de pondération au niveau des shapes.

3.1 La population des communes

La population étant la variable à expliquer et celle que l'on veut répartir au sein des shapes, c'est la seule variable à n'être présente que dans la couche des communes. Les décomptes de population varient de 1 170 à 303 417 habitants par commune. Les ¾ des communes ont moins de 20 000 habitants et seules huit en ont plus de 100 000 : il s'agit des six arrondissements (les noms des communes du district de la capitale sont 1er arrondissement, 2ème arrondissement etc…) de la région de la capitale et de deux autres venant des régions de Diana et d'Atsinanana. Pour des raisons pratiques et analytiques, nous choisissons de travailler avec la densité de population des 430 communes plutôt qu'avec les décomptes bruts. En effet, la densité de population permet de standardiser la mesure en tenant compte de la superficie de chaque commune, offrant ainsi une base de comparaison plus uniforme et réduisant les biais dus à la surface variable des communes. Sa distribution est résumée dans le tableau 11.

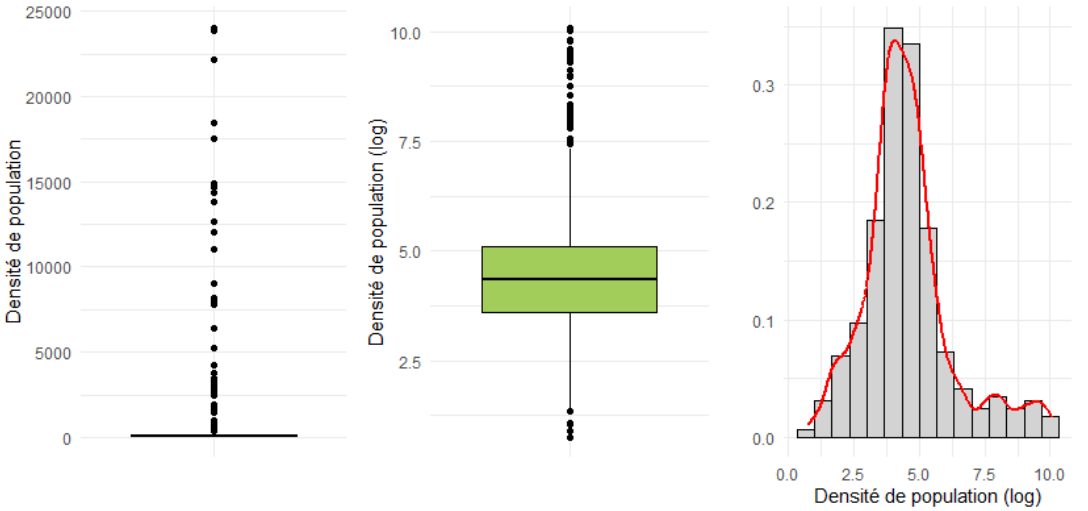
Tableau 11 : Distribution de la population des 430 communes

Variable	Minimum	1er quartile	Médiane	Moyenne	3ème quartile	Maximum
Population (RGPH-3)	1 170	6 785	11 503	18 837	19 856	303 417
Densité de population (hab./km²)	2	37	78	813	164	24 027

Sources : Décomptes de population issus de l'INSTAT et calcul par l'auteur des densités de population à partir des données de l'INSTAT et des découpages administratifs issus d'OCHA.

De plus et comme le montre la figure 26 suivante, il est nécessaire d'ajuster la densité de population à l'aide du logarithme pour améliorer la distribution de la variable. La transformation logarithmique atténue l'effet des valeurs extrêmes et rend la distribution plus symétrique. Cela permet d'avoir une distribution plus proche de la normale, de manière à conduire des analyses statistiques plus robustes et fiables.

Figure 26 : Distribution de la densité de population avant et après transformation logarithmique



Sources : Calcul par l'auteur des densités de population à partir des données de l'INSTAT et des découpages administratifs issus d'OCHA.

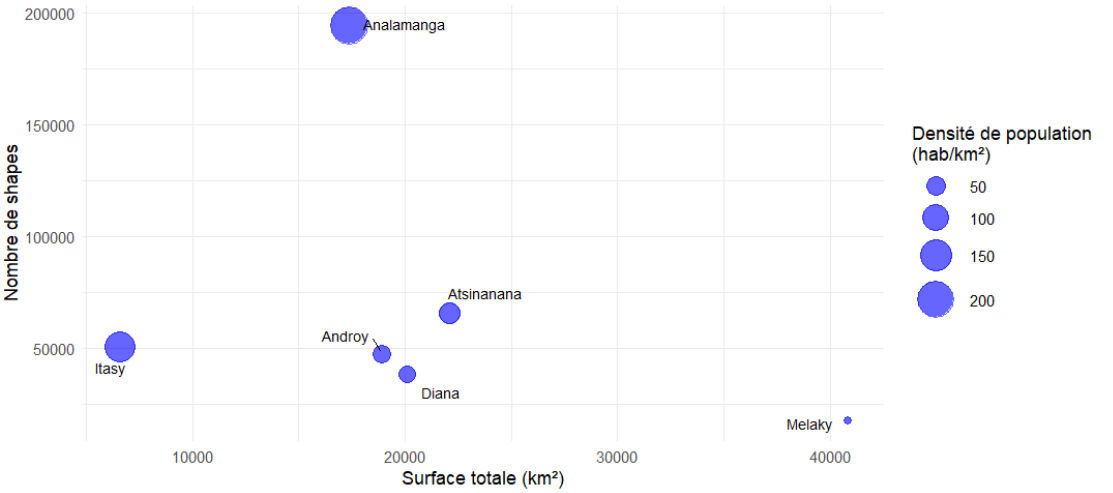
3.2 Densité d'occupation du sol par le bâti

L'INSTAT, à travers le rapport thématique numéro 5 « Habitation et cadre de vie de la population » introduit le concept de densité d'occupation du sol qui se réfère à la « concentration des habitations par rapport à la superficie sur laquelle les logements sont bâtis » (INSTAT, 2021f). Ce rapport révèle que chaque kilomètre carré de Madagascar contient en moyenne 10,3 habitations, avec des valeurs respectives de 66,5 en milieu urbain et 8,4 en milieu rural. À une échelle plus régionale, Analamanga (51,7 habitations/km²) et Itasy (31,2 habitations/km²) figurent parmi les zones les plus densément construites, tandis que Melaky affiche la plus faible densité avec seulement 1,7 habitation/km². Les trois autres régions étudiées se situent dans une fourchette intermédiaire, entre 10 et 20 habitations/km².

Bien qu'il soit compliqué de faire une analogie directe avec les shapes – qui sont de tailles assez hétérogènes et surtout qui comprennent souvent plusieurs bâtiments –, la figure 27 confirme une dynamique similaire entrevue par les résultats issus de l'INSTAT. En effet, on trouve en moyenne 11 et 7,5 shapes par km² pour

Analamanga et Itasy, suivies par Atsinanana, Androy, et Diana avec en moyenne entre 2 à 3 shapes par km². Une très faible densité de bâti est confirmée à Melaky avec seulement 0,4 shape par km².

Figure 27 : Relation entre la surface totale des régions, le nombre de shapes et la densité de population par région



Sources : Nombre de shapes issu du programme TeleCense, décomptes de population issus de l'INSTAT et calcul par l'auteur des densités de population à partir des données de l'INSTAT et des découpages administratifs issus d'OCHA.

Ces résultats confirment la pertinence du choix des six régions retenues pour cette étude. En effet, la capitale, Analamanga, était une sélection évidente compte tenu de son caractère fortement urbanisé et du fait que, comparativement à d'autres pays, Madagascar reste un territoire majoritairement rural et un des pays africains les moins urbanisés avec en 2018 un taux d'urbanisation de 19.3 % (INSTAT, 2021c). Parmi les autres régions, Itasy, qui a un nombre de shapes relativement similaire aux autres régions intermédiaires, se distingue par une densité de population élevée, qui pourrait s'expliquer par sa proximité avec la capitale et son appartenance aux Hautes Terres. Concernant Diana et Androy, bien que leurs densités de bâti soient similaires, leurs contextes environnementaux et climatiques diffèrent fortement. Diana, située au nord, est caractérisée par un climat tropical humide, tandis qu'Androy, au sud, est marqué par un climat aride. Enfin, Atsinanana, avec son positionnement côtier à l'est et son rôle économique stratégique, notamment via le port Toamasina favorisant les échanges avec l'océan Indien, peut aussi présenter une dynamique spécifique qu'il serait pertinent d'explorer plus en détail. À l'opposé, Melaky, région la moins densément peuplée, constitue un cas d'étude intéressant pour comprendre comment se structure l'habitat dans les zones à très faibles densités et où se concentre la population dans un tel contexte.

3.3 Récapitulatif et distribution des variables pour les deux couches de données

Avant d'aborder les calculs liés aux différentes méthodes de désagrégation, un récapitulatif des variables utilisées est proposé, en mettant en avant la distribution de chacune d'elles selon les niveaux communaux et shapes (Tableau 12).

Tableau 12 : Récapitulatif et distribution des variables pour les deux couches de données

Variable	Niveau	Minimum	1Q	Médiane	Moyenne	3Q	Maximum
Population (RGPH-3)	commune	1 170	6 785	11 503	18 837	19 856	303 417
	shape						
Densité de population après transformation logarithmique	Commune	0.75	3.6	4.36	4.54	5.1	10.1
	shape						
Température (°C)	commune	14,2	18,9	21,9	21,6	23,8	27,1
	shape	11,7	18,5	19,4	20,8	23,6	27,5
Précipitations (mm/an)	commune	436	1 259	1 394	1 549	1 853	2 957
	shape	398	1265	1364	1507	1586	3043
Altitude (m)	commune	0	104	500	685	1281	1921
	shape	0	197	1229	862	1341	2668
Intensité de la lumière la nuit	commune	7	14	15	50	18	1201
	shape	0	12	16	40	22	7028
Distance à la route (m)	commune	0	90	1 202	3 446	5 077	20 001
	shape	0	0	32	2 650	2 231	20 001
Distance au point d'eau (m)	commune	0	355	902	1 661	1 758	20 001
	shape	0	0	277	1 351	1 507	20 001
Distance au point de santé (m)	commune	0	2 058	3 890	4 312	5 841	20 001
	shape	0	0	1 123	2 758	4 233	20 001
Distance à la grande agglomération (m)	commune	0	4 138	17 210	26 270	35 841	40 001
	shape	0	714	3 122	3 998	5 818	23 863
Proportion de surface bâtie (communale)	commune	0	0,1	0,3	2,4	0,8	61,7
	shape	0.009	0,14	0,37	1,84	1,16	61,7
Nombre moyen d'habitants par ménage	commune	3,3	4,1	4,1	4,1	4,5	4,5
	shape	3,3	4,1	4,1	4,1	4,5	4,5
Densité de bâti	shape	0	2	3	2,9	4	10
Surface (m^2)	shape	0	200	400	1675	1100	2 683 164
Surface Google (m^2)	shape	0	0	190	698	400	1 483 017

Sources : Décomptes de population du RGPH3 (INSTAT) et variables issues du programme TeleCense

Dans un souci d'exploration plus approfondie des relations entre les variables susceptibles d'influencer la modélisation de la population, nous avons choisi d'appliquer une Analyse en Composantes Principales (ACP). Cette approche permet de résumer l'information tout en identifiant les liens entre les variables, et d'identifier ainsi celles qui seront intégrées à nos modèles. À cet égard, nous avons également procédé à un ajustement des variables de distance et de l'intensité lumineuse nocturne par la transformation logarithmique, en raison de la nature fortement asymétrique de leurs distributions (tableau 12).

La figure 28, qui présente le cercle de corrélations de l'ACP, met en évidence que le premier plan factoriel explique 71 % de l'inertie du jeu de données, avec une répartition de 50 % pour le premier axe et 21 % pour le second. Le fait que seulement deux axes capturent l'essentiel de la variance du jeu de données confirme que l'ACP est efficace ici comme méthode de réduction dimensionnelle.

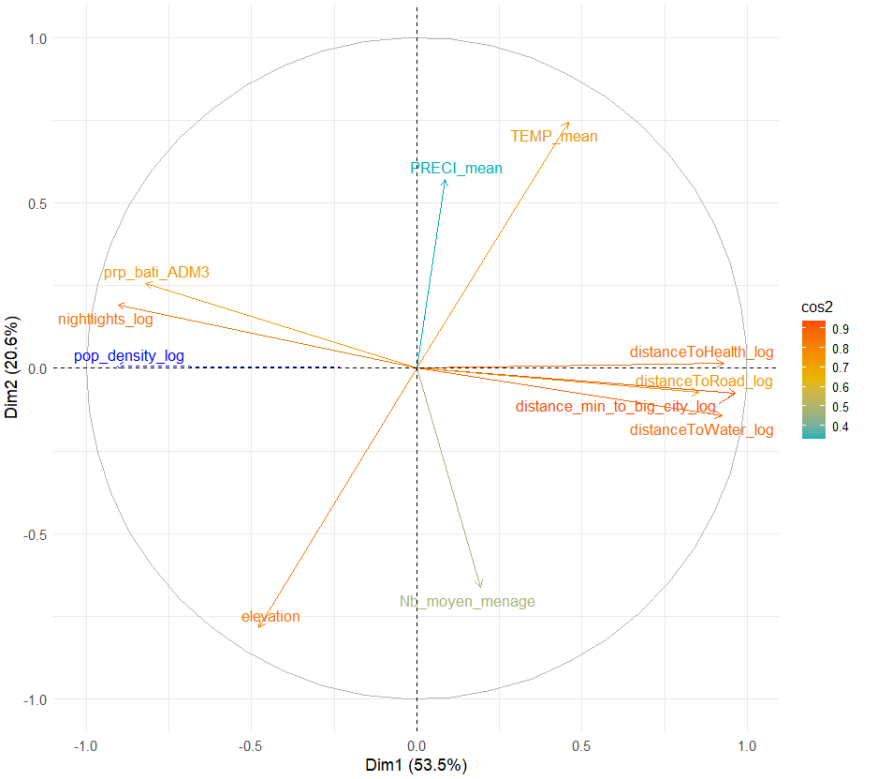
La variable de densité de population, ajustée par le logarithme (notée "pop_density_log"), a été incluse comme variable supplémentaire dans cette analyse afin d'examiner ses relations avec les autres variables. L'ACP révèle que la densité de population est fortement corrélée avec le premier axe, indiquant que les zones à forte densité de population se distinguent clairement des zones moins peuplées le long de cet axe. La représentation graphique suggère que les zones densément peuplées sont celles caractérisées par une forte proportion de bâti et une intensité lumineuse nocturne élevée. En outre, ces zones semblent être associées à de faibles valeurs de distance (les variables de distance ont une direction inverse à la variable de densité de population sur la figure 28), ce qui est cohérent avec l'idée que les communes dotées d'infrastructures telles que des routes, des établissements de santé et des grandes villes attirent une population plus importante, renforçant ainsi les facteurs d'urbanisation.

Une observation notable est la forte proximité entre les variables de distance, qui semblent partager une interprétation similaire. Bien que cela semble logique, cela soulève la question de la redondance de ces variables, et il sera important d'évaluer si certaines doivent être exclues sans perdre d'information essentielle dans le cadre de la modélisation.

Quant à la proportion de bâti et à l'intensité lumineuse nocturne, leur forte association est un aspect positif, car elle conforte l'idée que ces deux variables sont étroitement liées et que peut être que la variable d'intensité lumineuse nocturne suffira pour la modélisation. En effet, la proportion de bâti ne pourra pas être utilisée dans la modélisation, car elle est agrégée au niveau des communes et ne se transpose pas aux unités d'analyse des shapes, lesquelles auraient alors des valeurs identiques, rendant cette variable inutile, voire perturbant la modélisation.

Avec ce graphique, il semble difficile d'analyser les impacts des variables climatiques et environnementales sur la densité de population, car la densité de population est très fortement liée à l'axe 1, alors que les autres variables sont surtout liées à l'axe 2. Cependant, on observe que la variable de température est complètement opposée à celle d'altitude, ce qui est logique car à mesure que l'altitude augmente, les températures ont tendance à diminuer. On les conserve pour la suite de l'analyse et identifierons leur impact plus précisément lors de la mise en œuvre des modèles.

Figure 28 : Analyse en Composante Principales des variables des communes de Madagascar



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et des variables issues du programme TeleCense

Au niveau des shapes, il est intéressant de noter que la majorité de notre jeu de données est constituée de shapes relativement petites, avec 75 % d'entre elles ayant une superficie inférieure à 1100 m². Cependant, aux côtés de ces petites shapes, on trouve également des shapes bien plus grandes, comme l'illustrent les exemples des figures précédentes (figure 25, également figure 14 – chapitre 2) d'agglomérations urbaines où une grande densité de bâti est concentrée au sein d'une même shape.

Également, lorsque l'on passe de la surface identifiée par TeleCense à la surface pondérée par les identifications de Google Open Buildings, on remarque que l'on réduit considérablement la taille des surfaces des shapes détectées. Notamment,

de nombreuses shapes ont une surface pondérée estimée comme nulle, ce qui pourrait indiquer des faux positifs du côté de la détection par TeleCense. Si l'on décide de désagréger en utilisant cette surface pondérée, les shapes avec une surface pondérée nulle ne sont pas prises en compte dans la répartition de la population lors de la désagrégation.

4. Calcul des pondérations Ws par modélisation linéaire et forêt aléatoire

Pour rappel, quelque soit la méthode utilisée, nous commençons par diviser aléatoirement les données en deux ensembles distincts : un ensemble d'entraînement composé de 70 % des communes, et un ensemble de test constitué des 30 % de communes restantes.

4.1 Modélisation linéaire des communes de Madagascar

Pour construire le modèle linéaire à partir de l'échantillon d'entraînement, nous avons adopté une approche progressive de sélection des variables. Nous avons d'abord testé individuellement chaque variable, en nous basant sur les corrélations préalablement observées lors de l'ACP. L'objectif était d'identifier celles qui sont le plus liées à la densité de population et d'évaluer leur pertinence dans le modèle. Ensuite, nous avons intégré progressivement les variables les plus significatives, en analysant leur contribution et en vérifiant l'amélioration du modèle à chaque étape.

Finalement, la variable de distance au point d'eau n'a pas été retenue, car elle s'est avérée non significative. L'altitude, quant à elle, a été exclue en raison de sa multicolinéarité avec les variables climatiques, notamment les précipitations. En effet, dans ce contexte, les précipitations suffisent généralement à capter l'effet de l'altitude sur la densité de population.

Enfin, bien que la proportion de surface bâtie de la commune présente une relation forte avec la densité de population, elle n'a pas été intégrée au modèle. En effet, cette variable, ainsi que le nombre moyen d'habitants par ménage, prennent des valeurs identiques pour toutes les shapes d'une même commune, ce qui limite leur pertinence pour affiner les estimations à une échelle plus fine lors de l'exercice de désagrégation. L'un des meilleurs modèles obtenus prend en compte six variables :

$$Y_i = \beta_0 + \beta_1 \log(distanceToCity) + \beta_2 \log(distanceToRoad) + \beta_3 \log(distanceToHealth) + \beta_4 \log(nightlights) + \beta_5 precipitations + \beta_6 temperature + \varepsilon_i$$

Avec Y_i la densité de population de la commune i et ε_i le terme résiduel aléatoire censé suivre une distribution aléatoire indépendante qui nous permet de faire des inférences sur les paramètres du modèle. Les estimations des coefficients β associés à chaque variable ont une p-valeur inférieure à 0.05 (tableau 13), garantissant que tous les facteurs pris en compte dans ce modèle jouent un rôle significatif dans la modélisation de la densité de population des communes de Madagascar.

Un signe positif du coefficient indique que plus la variable a une valeur élevée, plus la densité de population sera élevée elle aussi, à niveau fixé des covariables. À l'inverse, un signe négatif indique que plus la variable prend une valeur élevée, moins la commune sera densément peuplée. Pour comprendre l'intensité de l'effet d'une variable sur la population, on examine le coefficient standardisé, qui, étant normalisé et sans unité, permet de comparer directement l'effet de chaque variable, à niveau fixé des covariables.

Les variables de distance à la route, de distance à la grande agglomération, et d'intensité lumineuse nocturne sont les plus importantes. Les signes négatifs de ces coefficients suggèrent que la proximité des routes est associée à une densité de population plus élevée pour la et que plus une commune est éloignée des grandes villes, plus sa densité de population est faible. Cela reflète l'attrait des grandes agglomérations en termes d'opportunités et d'accès à certains services, comme l'accès aux centres de santé, par exemple qui est une autre des variables du modèle et qui, avec son coefficient négatif, traduit la même dynamique. Enfin, le coefficient relatif à la densité lumineuse est positif indiquant que plus une commune a une intensité lumineuse forte, plus elle aura une densité de population élevée.

En ce qui concerne les variables climatiques, le coefficient négatif de la température suggère que des températures plus élevées sont associées à une densité de population plus faible. Cela pourrait s'expliquer par le fait que des températures excessives ne sont pas favorables à l'habitation humaine. En revanche, le coefficient positif des précipitations indique que des précipitations plus élevées sont associées à une densité de population légèrement plus élevée. Cela pourrait être lié à l'agriculture, car les régions avec des précipitations adéquates sont souvent plus propices à l'agriculture, attirant ainsi une population plus dense, notamment dans les hautes terres (Raison, 1974).

Il est important de noter que ces résultats pourraient être influencés par le choix restreint des régions étudiées. Bien que cela ne soit pas nécessairement un problème si l'objectif est de désagréger au mieux les données, cela pourrait poser problème si l'objectif final est de mettre en place un modèle ascendant, qui, rappelons-le, vise à s'entraîner sur les résultats de la désagrégation afin de pouvoir estimer la population d'autres zones géographiques en dehors de l'entraînement. Dans ce cas, on risquerait une sur-adaptation et sans doute un surapprentissage du modèle qui pourrait rendre difficile son exportation. Il faut garder cela à l'esprit et si nous retenons cette méthode comme méthode finale pour désagréger la population, il serait judicieux de centrer et de réduire ces variables.

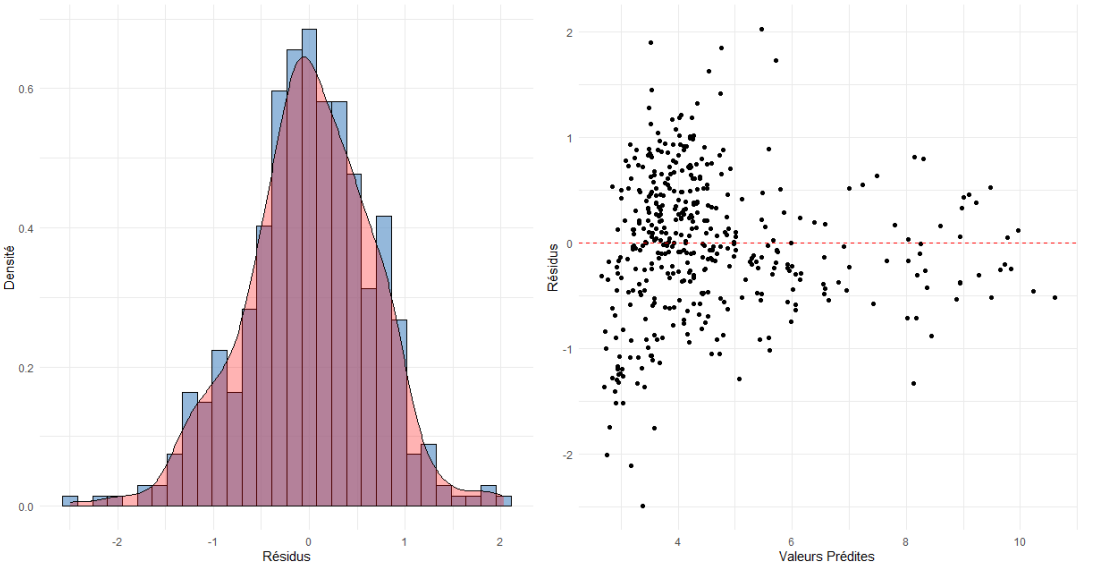
Pour finir, avec un R^2 ajusté de 0,83, le modèle est capable d'expliquer 83 % de la variance de la densité de population dans l'échantillon des communes de Madagascar à l'aide des variables sélectionnées. De plus, l'analyse des résidus présentée en figure 29 valide les hypothèses fondamentales de la régression linéaire. D'une part, la distribution des résidus suit une forme approximativement normale, sans asymétrie marquée ni valeurs extrêmes, ce qui suggère que l'hypothèse de normalité des erreurs est respectée. D'autre part, l'analyse des résidus en fonction des valeurs prédites ne révèle aucun motif particulier si ce n'est une hétérogénéité plus marquée des résidus lorsque les valeurs prédites sont faibles, mais il n'y a pas de tendance marquée, indiquant une variance constante des résidus quelle que soit la densité de population prédite. Ces éléments permettent ainsi de valider a priori l'hypothèse d'homoscédasticité.

Tableau 13 : Résultats et coefficients de la modélisation de la densité de population des communes de Madagascar

Variable	Estimation du coefficient	Coefficient standardisé
Constante = Densité de population (au log)	8.049***	
Moyenne de la distance à la grande ville la plus proche (en mètres et au log)	-0.13***	-0.24
Moyenne de la distance à la route la plus proche (en mètres et au log)	-0.20***	-0.33
Moyenne de la distance au point de santé le plus proche (en mètres et au log)	-0.079***	-0.09
Nightlights (au log)	0.49***	0.24
Moyenne des précipitations	0.00026***	0.09
Moyenne de la température	-0.11***	-0.19
R^2 sur l'ensemble de test	0.834	

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et des variables issues du programme TeleCense

Figure 29 : Évaluation de la normalité et de l'homogénéité des résidus de la modélisation linéaire



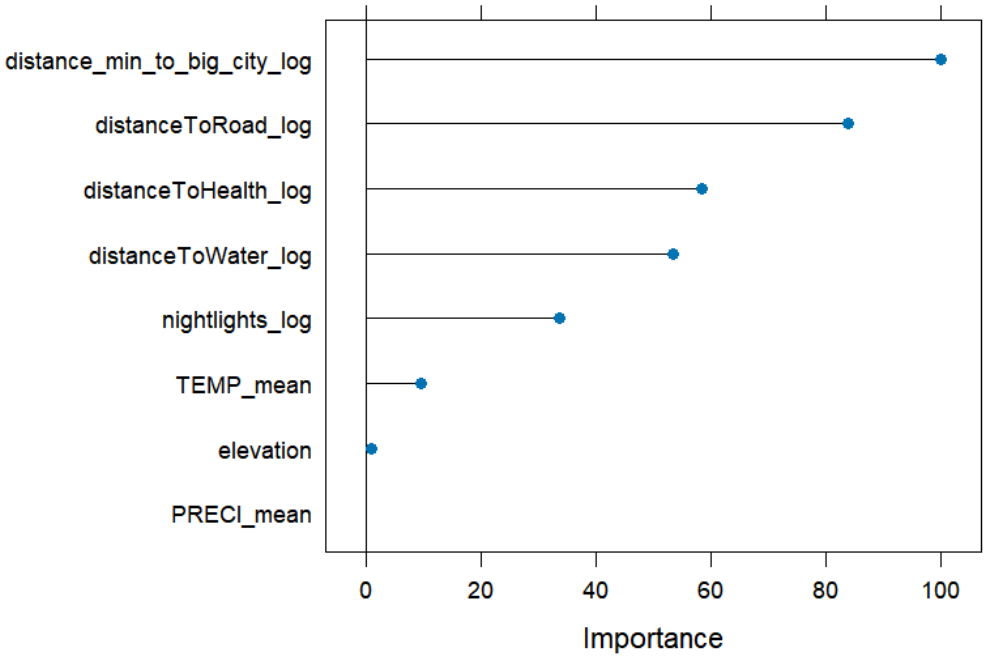
Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et des variables issues du programme TeleCense

4.2 Forêts aléatoires pour la modélisation des communes de Madagascar

En ce qui concerne l'algorithme de forêt aléatoire, le modèle a été ajusté sur l'échantillon d'entraînement en utilisant la validation croisée avec cinq plis, comme détaillé dans la section « 3.1 Approche descendante – La désagrégation » du chapitre 2. L'un des avantages de l'algorithme de forêt aléatoire est sa faible sensibilité à la multicolinéarité, ce qui nous a conduit à inclure toutes les variables dans le modèle pour observer les résultats (notamment l'altitude et la distance au point d'eau).

La figure 30 illustre l'importance des différentes variables dans le modèle, c'est-à-dire leur contribution à la prédiction du modèle. La variable de distance à la grande agglomération s'avère être la plus influente, suivie par les variables de distance à la route, au point de santé, et au point d'eau. L'intensité de la lumière nocturne a une importance moindre, tout comme la température. Enfin, les variables de précipitations et d'altitude ont une importance nulle ou quasi nulle dans le modèle de forêt aléatoire, contrairement à ce que nous avions observé pour les précipitations dans la régression linéaire.

Figure 30 : Importance des variables de l'algorithme de forêt aléatoire



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et des variables issues du programme TeleCense

4.3 Comparaison des performances des deux modèles :

Dans la suite de notre analyse, nous utilisons l'ensemble de test (30 % des communes) pour prédire la densité de population logarithmique et comparer les performances de nos deux modèles. Comme le montre le tableau 14 ci-dessous, l'algorithme de forêt aléatoire présente un RMSE plus faible et un R^2 plus élevé que la régression linéaire, indiquant une meilleure capacité à estimer la densité de population logarithmique.

Tableau 14 : Comparaison des performances des modèles de régression linéaire et de forêt aléatoire pour la prédiction de la densité de population

Modèle	Root Mean Square Error (RMSE)	R^2
Régression linéaire	0.69	0.829
Forêt aléatoire	0.51	0.91

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et des variables issues du programme TeleCense

Bien que ces deux modèles soient très performants pour estimer le logarithme de la densité de population des communes de Madagascar, il faudra évaluer si les relations trouvées au niveau des communes peuvent se transférer au niveau des shapes et si cela est adéquat pour la répartition de la population.

5. Mise en place des différentes méthodes de désagrégation

À la suite du calcul des deux pondérations précédentes, nous avons en tout quatre méthodes de désagrégation à tester et à évaluer. La première repose sur l'interpolation surfacique, qui consiste à répartir la population des communes entre les shapes qui leur sont rattachées, proportionnellement à leur surface. La seconde méthode est l'allocation proportionnelle à la densité de bâti, où la population est redistribuée en fonction de la surface et de la variable de densité de bâti des shapes. Les deux dernières approches reposent sur une modélisation de la densité de population à partir de plusieurs variables explicatives, dont l'estimation est ensuite utilisée pour pondérer la répartition de la population au niveau des shapes. Cette modélisation a été réalisée selon une modélisation linéaire et un algorithme de forêt aléatoire.

Nous appliquons ces méthodes afin de répartir dans les shapes les décomptes de population des 430 communes de Madagascar. Chaque méthode produit une distribution différente de la population. Cependant, il est difficile de déterminer, uniquement à partir de ces distributions, la validité de chaque méthode et de les comparer. Nous décidons donc d'évaluer les méthodes en suivant la procédure expliquée dans le chapitre 2.

5.1 Evaluation des méthodes de désagrégation

Pour évaluer la précision de la désagrégation, nous procédons à une désagrégation à partir d'un niveau administratif plus large : celui des districts, qui est le deuxième niveau administratif à Madagascar, avec un total de 32 districts répartis sur six régions. Une fois la répartition de la population dans les shapes effectuées, nous additionnons ces prédictions à l'échelle des communes afin de les comparer aux décomptes officiels du RGPH3. La précision de chaque méthode sera mesurée par le pourcentage d'erreur relative pour chacune des communes, selon la formule suivante :

$$Pct.Erreur_{relative} = \frac{Decomptes_{RGPH} - Predictions}{Decomptes_{RGPH}} * 100$$

Le tableau 15 présente la distribution de la population dans les shapes selon chaque méthode de pondération, toujours à partir des districts et non des communes et en utilisant la surface ajustée par la détection Google (le choix de cette surface est justifié dans la partie qui suit). Indépendamment de la méthode choisie, le premier quart des surfaces bâties détectées affiche une population estimée nulle, en raison des

shapes ayant une surface nulle. Cela s'explique par le fait que, dans notre approche actuelle utilisant la surface ajustée par la détection de Google, certaines shapes ont une surface nulle, probablement en raison de faux positifs issus de notre détection initiale.

Tableau 15 : Distribution de la population selon les différentes méthodes de pondération et à partir de la désagrégation de la population des districts

Méthode de pondération	1Q	Médiane	3Q	90 %	95 %	99 %	99.9 %	Maximum
1. Surface	0	3.60	10.64	28	60	252	1 495	30 904
2. Densité de bâti	0	2.98	9.92	26.5	56	259	1 685	32 340
3. Modélisation linéaire	0	3.37	9.57	26.3	57	260	1 672	38 895
4. Forêt aléatoire	0	3.33	9.47	26.2	57	259	1 666	37 691

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection Google Open Buildings et des variables issues du programme TeleCense

Dans l'ensemble, les estimations demeurent similaires pour les quartiles supérieurs, avec des valeurs particulièrement élevées dans le dernier percentile, en raison de la présence de quelques shapes de grande taille dans la base de données. Une différence notable se situe dans les valeurs maximales qui semblent bien plus élevées pour les méthodes modélisation linéaire et forêt aléatoire.

Il est néanmoins difficile de déterminer quelle méthode de désagrégation fournit la répartition la plus précise des populations uniquement à partir de ces distributions. Afin de déterminer quelle méthode est la plus efficace pour répartir la population dans les shapes nous nous appuyons sur les résultats des tableaux 16 et 17 qui présentent la distribution des erreurs selon les quatre méthodes de pondération. L'objectif était aussi d'évaluer si l'erreur moyenne diffère significativement entre la surface détectée par TeleCense et celle ajustée à partir de la détection Google. Pour rappel, cette dernière est calculée à partir des données du programme Google Open Buildings, qui exploite des images dont la résolution est dix à vingt fois plus fine que celle des satellites Sentinel. Cette approche vise principalement à affiner l'estimation de la surface des grandes shapes, en corrigeant d'éventuelles zones non bâties qui auraient pu être intégrées lors de leur détection initiale.

Tableau 16 : Distribution de l'erreur relative selon les différentes méthodes de pondération avec la surface TeleCense

Méthode de pondération	Erreur moyenne (%)	25 %	50 %	75 %	90 %
Surface	30,9	10,8	24,9	40,8	61,3
Densité de bâti	26,3	9,7	20,1	36,8	54,3
Modélisation linéaire	33,9	14,3	27,9	46,9	68,1
Forêt aléatoire	32,2	12,7	27,4	45,0	63,9

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et des variables issues du programme TeleCense

Tableau 17 : Distribution de l'erreur relative selon les différentes méthodes de pondération avec la surface TeleCense ajustée à partir de la détection Google

Méthode de pondération	Erreur moyenne (%)	25 %	50 %	75 %	90 %
Surface	23,2	7,2	19,4	32,8	51,9
Densité de bâti	22,8	9,4	17,1	32,3	46,8
Modélisation linéaire	26,5	9,9	22,8	39,1	53
Forêt aléatoire	26,0	9,3	20,9	38,2	56,7

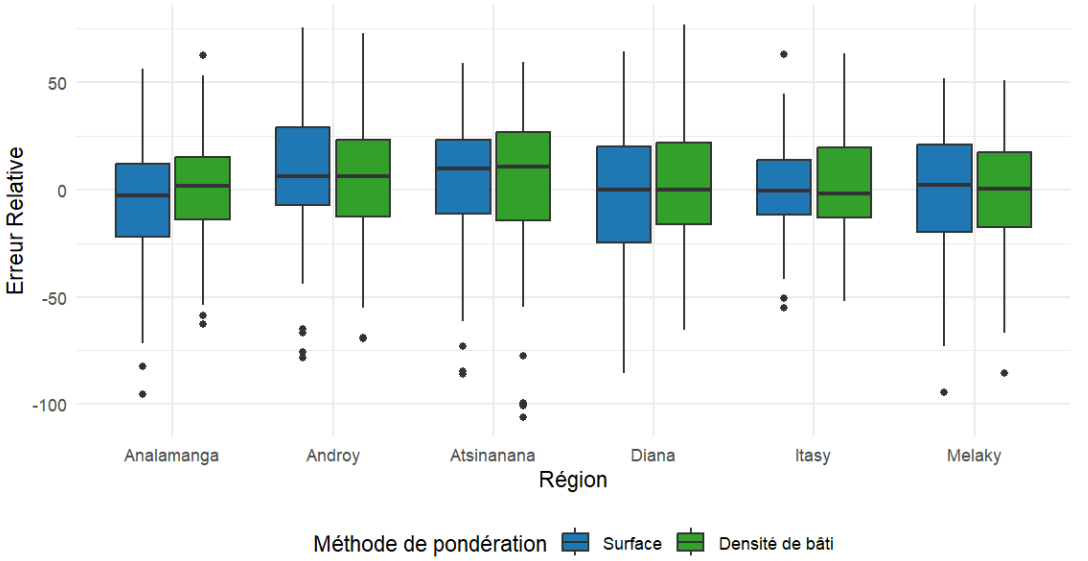
Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection Google Open Buildings et des variables issues du programme TeleCense

Les résultats des tableaux 16 et 17 montrent une différence marquée dans la distribution des erreurs en fonction des types de détections utilisés. Lorsqu'on utilise la surface ajustée par Google, l'erreur moyenne est systématiquement plus faible, avec une réduction de 3,5 à 7,7 % par rapport à la surface TeleCense. Cette tendance est observée pour toutes les méthodes de pondération, ce qui suggère que la surface modifiée par Google améliore la précision de la désagrégation de la population. En conséquence, nous choisissons d'utiliser exclusivement la surface modifiée par Google pour la suite de l'analyse, étant donné l'erreur significativement plus faible qu'elle génère.

À première vue, et en tenant compte de la performance prédictive évaluée au niveau des communes, il peut sembler surprenant que les méthodes basées sur des couches de pondération obtenues par modélisation linéaire et forêt aléatoire soient, en moyenne, moins précises que celles reposant uniquement sur la surface ou sur la combinaison de la surface et de la densité de bâti. Cette différence de performance pourrait s'expliquer par le fait que nous avons supposé que la relation identifiée au niveau communal se maintiendrait au niveau des shapes, ce qui n'est pas nécessairement le cas. Une autre explication réside dans le nombre relativement faible de communes utilisées pour créer les modèles (environ 300 communes sur 430, soit 70 %).

Il est donc difficile de départager les deux premières approches de désagrégation, car, comme l'indique le tableau 17 (colonne 4), plus de 50 % des communes présentent une erreur inférieure à 20 %. Plus précisément, comme le montre la figure 31, les erreurs associées aux deux méthodes semblent globalement similaires, indépendamment de la région.

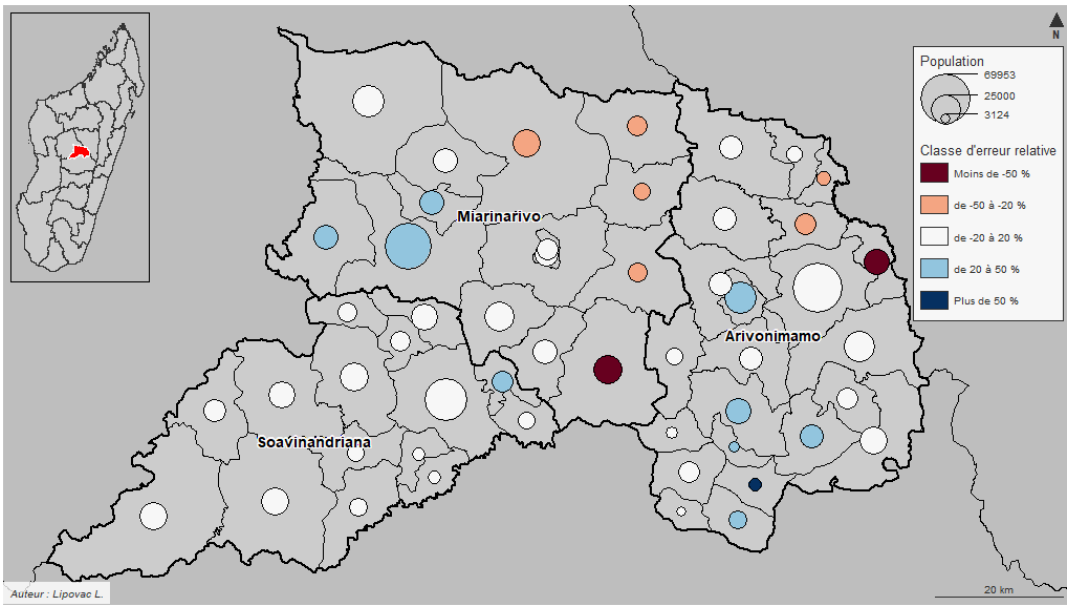
Figure 31 : Comparaison des erreurs au niveau régional en fonction des deux premières méthodes de désagrégation



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection Google Open Buildings et des variables issues du programme TeleCense

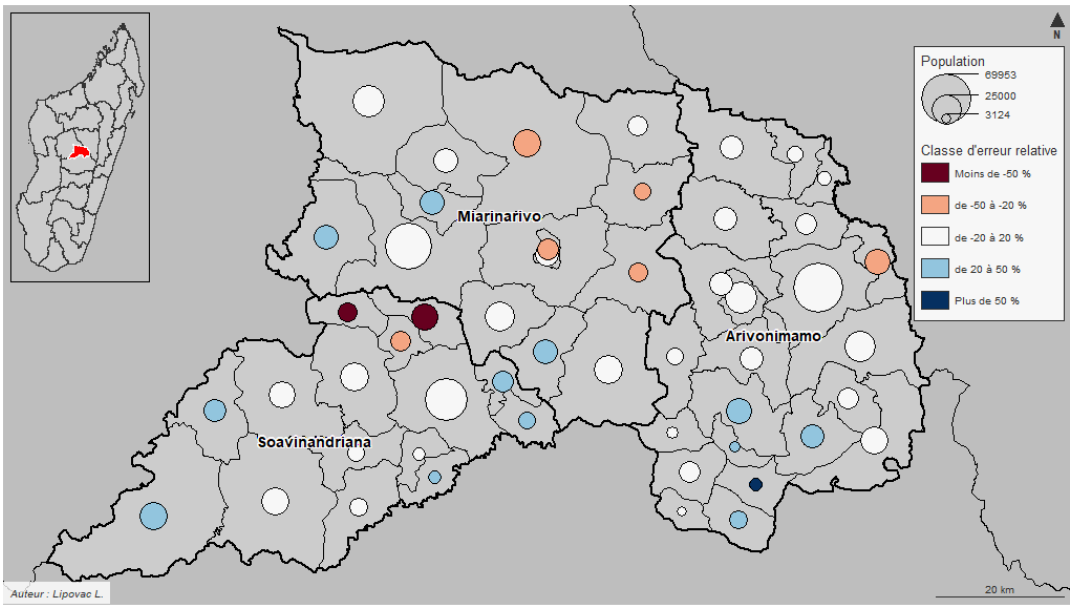
Nous choisissons finalement d'analyser les résultats sous forme de cartes en classifiant les erreurs selon cinq catégories : grande surestimation (erreurs inférieures à -50 %), surestimation (-50 % à -20 %), estimation correcte (-20 % à +20 %), sous-estimation (+20 % à +50 %) et forte sous-estimation (erreurs supérieures à +50 %). La palette de couleurs, allant du rouge au bleu, sera utilisée pour symboliser cette catégorisation dans les deux cartes suivantes (figures 32 et 33).

Figure 32 : Carte des erreurs relative d'Itasy avec la méthode d'interpolation surfacique



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et de la détection de bâti TeleCense dont la surface est ajustée par la détection Google

Figure 33 : Carte des erreurs relative d'Itasy avec la méthode utilisant la densité de bâti



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et de la détection de bâti TeleCense dont la surface est ajustée par la détection Google

Bien que nous ne nous concentrions ici que sur la région d'Itasy, les cartes ne permettent pas non plus de faire un choix clair entre les deux méthodes. Certaines spécificités par district se dégagent, comme dans le cas du district de Soavindraiana, où les résultats sont presque tous correctement estimés par l'interpolation surfacique, contrairement à la méthode basée sur la densité de bâti. En revanche, dans les autres districts, notamment la commune de Mandiavato, située au sud du district de Miarinaviro, les erreurs sont plus marquées, avec une surestimation importante de la première méthode par rapport à la méthode fondée sur la densité de bâti.

Le fait que la méthode d'interpolation surfacique soit aussi performante peut s'expliquer par le type d'habitations, qui est à 83 % composé de maisons individuelles, suivi par 5 % de concessions et 12 % d'appartements ou maison collectives (INSTAT, 2021f). Bien que les maisons individuelles et concessions puissent avoir un étage, comme c'est parfois le cas dans les Hautes Terres centrales, la plupart n'en ont pas. Les résultats de la désagrégation par interpolation surfacique montrent ainsi qu'à partir du moment où les bâtiments sont correctement identifiés, avec une surface précise, il n'est pas nécessaire d'avoir de nombreuses variables ou des modèles plus complexes. Cela fonctionnerait donc particulièrement bien lorsque les constructions sont peu élevées.

A l'inverse, dans la région d'Analamanga, plus urbaine, la désagrégation basée sur la densité de bâti s'avère légèrement plus performante (figure 31), ce qui souligne l'importance de la prise en compte des structures verticales à travers la variable de densité de bâti pour capter la population urbaine dense. Il est donc possible que cette méthode soit plus adaptée à d'autres régions ou pays présentant une configuration similaire.

En raison de la similarité des résultats obtenus avec les deux méthodes, nous avons décidé de retenir les deux et de les appliquer en fonction des spécificités des régions sur lesquelles porte le projet TeleCense. Cependant, pour la suite de cette étude, j'ai choisi de privilégier la méthode d'interpolation surfacique, principalement pour sa simplicité d'implémentation et pour éviter de dépendre de la variable de densité de bâti. Il semble plus prudent de retenir la méthode d'interpolation surfacique, car elle repose sur des données plus simples et directes, sans dépendre d'une variable dont la validité scientifique n'a pas encore été confirmée.

5.2 Comparaison des résultats de désagrégation avec les estimations de WorldPop

Maintenant que nous avons sélectionné la méthode de désagrégation, nous pouvons comparer les résultats avec ceux fournis par le programme WorldPop, qui constitue la référence actuelle pour les estimations de population. Pour cette comparaison, nous utilisons les estimations « Unconstrained » du programme WorldPop pour Madagascar pour l'année 2018[50]. Ces estimations fournissent des décomptes de population par carreau de 100x100 mètres.

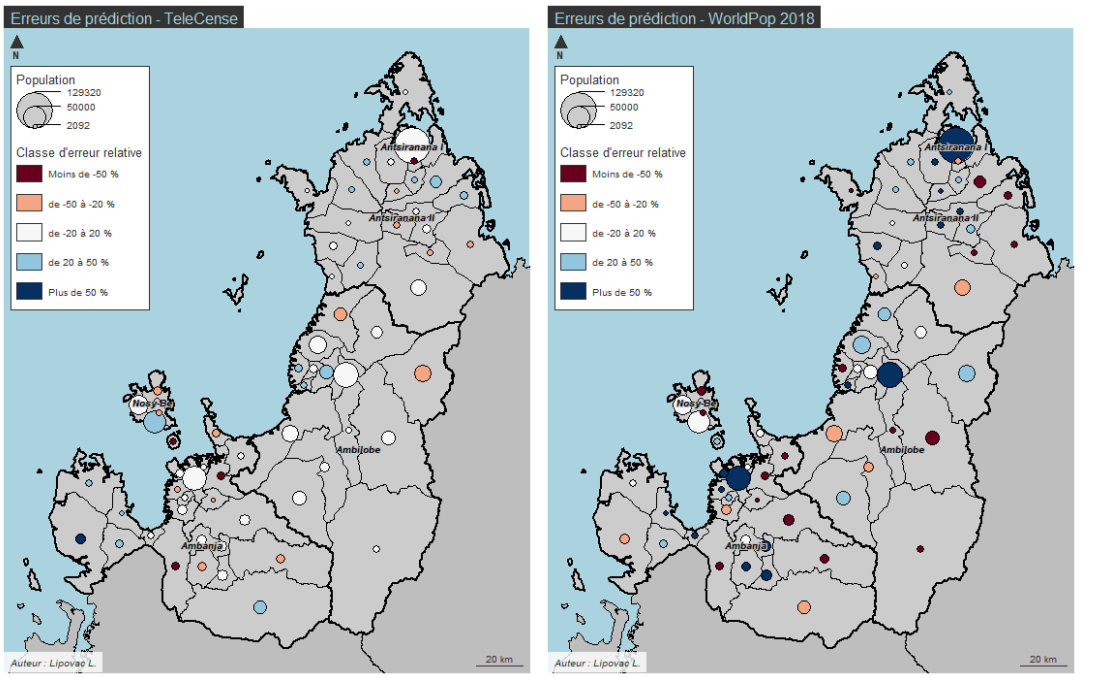
Idéalement, il aurait été préférable de comparer les résultats avec les estimations contraintes de WorldPop, qui intègrent des informations détaillées sur la cartographie des bâtiments, et donc plus proches des estimations de population générées par le programme TeleCense. Les estimations non contraintes ont en effet une erreur moyenne bien plus élevée que les estimations contraintes avec l'empreinte du bâti prise en compte (Darin et al., 2022). Toutefois, les estimations contraintes ne sont disponibles que pour l'année 2020, ce qui rend impossible leur utilisation pour cette analyse en 2018.

Afin de réaliser cette comparaison de manière cohérente, il est nécessaire d'agréger les données de WorldPop au niveau communal pour les rendre compatibles avec les résultats obtenus à partir de la désagrégation des shapes de TeleCense par

[50] https://hub.worldpop.org/geodata/summary?id=5898

interpolation surfacique. Une fois cette agrégation effectuée, une catégorisation similaire des erreurs est appliquée au niveau des communes, dans le but d'évaluer la correspondance des erreurs entre les deux ensembles de données. Cette analyse est illustrée par un exemple détaillé dans la région de Diana, qui met en évidence que les résultats issus de notre méthode de désagrégation surpassent globalement ceux de WorldPop pour la même année (Figure 34).

Figure 34 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – exemple dans la région de Diana - 2018

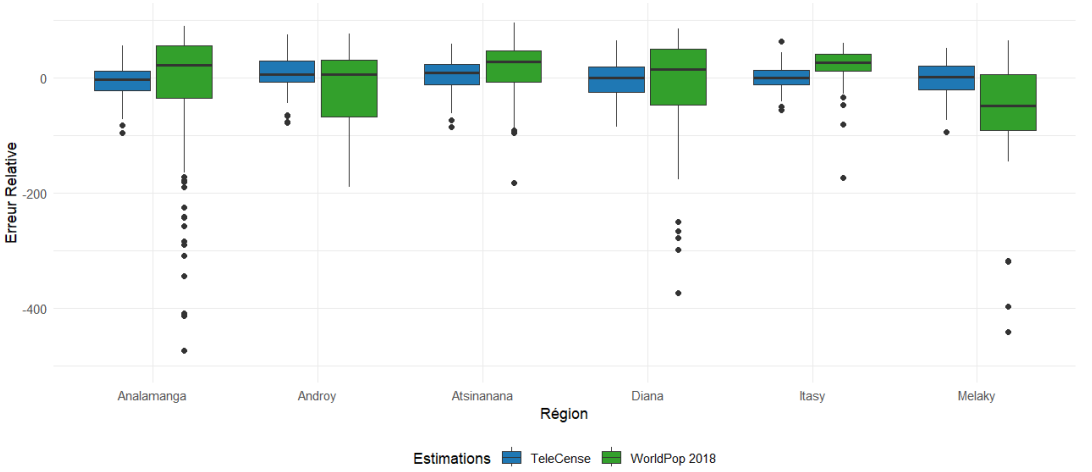


Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection de bâti TeleCense dont la surface est ajustée par la détection Google et des estimations non contraintes de WorldPop en 2018

Ce constat est confirmé à l'échelle des autres régions de Madagascar, comme en témoigne la figure 35 ainsi que les tableaux 18 et 19, qui présentent une comparaison détaillée de la distribution des erreurs des communes pour chaque région (les comparaisons en cartes avec les cinq autres régions sont disponibles en annexe 3). Il ressort de cette comparaison que, dans toutes les régions étudiées, les estimations de TeleCense semblent en moyenne plus proche de zéro que ne le sont celles via WorldPop. De plus, à l'exception de la région d'Itasy, les estimations issues de WorldPop présentent des amplitudes d'erreur nettement plus élevées par rapport à celles issues de la méthode TeleCense. Ce constat est également corroboré par les résultats numériques présentés dans les tableaux, où l'on observe que 7,7 % des communes sont fortement surévaluées avec la méthode TeleCense, un chiffre déjà important, tandis que pour WorldPop, ce pourcentage atteint 22,8 %. Plus

significativement, ces tableaux révèlent que TeleCense présente une proportion beaucoup plus élevée de communes pour lesquelles l'erreur se situe dans une plage relativement réduite (-20 % à 20 %) : 51,2 % des communes sont classées dans cette catégorie avec TeleCense, contre seulement 19,3 % avec WorldPop.

Figure 35 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop pour toutes les régions – 2018



Sources : TeleCense & WorldPop – Madagascar population 2018

Tableau 18 : Distribution de l'erreur au niveau des communes pour la désagrégation TeleCense

Catégorie d'erreur	Nombre de communes (n)	Pourcentage (%)	Pourcentage cumulé (%)
Moins de -50 %	33	7.7	7,7
de -50 à -20 %	72	16.7	16,7
de -20 à 20 %	220	51.2	51,2
de 20 à 50 %	90	20.9	20,9
Plus de 50 %	15	3.5	3,5

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et de la désagrégation TeleCense

Tableau 19 : Distribution de l'erreur au niveau des communes à partir des estimations non contraintes de WorldPop en 2018

Catégorie d'erreur	Nombre de communes (n)	Pourcentage (%)	Pourcentage cumulé (%)
Moins de -50 %	98	22.8	22,8
de -50 à -20 %	42	9.8	9,8
de -20 à 20 %	83	19.3	19,3
de 20 à 50 %	118	27.4	27,4
Plus de 50 %	89	20.7	20,7

Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT) et des estimations non contraintes de WorldPop en 2018

Ainsi, les résultats indiquent que la couche de population produite à partir des données de bâti TeleCense est de meilleure qualité que celle produite par WorldPop. Ce constat soutient l'hypothèse selon laquelle l'intégration des caractéristiques bâties dans le cadre de la désagrégation (Darin et al., 2022; Thomson, Leasure, Bird, Tzavidis, Tatem, 2022) fournit une approximation plus fiable des populations comparée aux méthodes de modélisation globales, telles que celles utilisées par les estimations non contraintes de WorldPop. Nous ne pouvons néanmoins pas conclure que l'approche adoptée soit meilleure que celle adoptée par WorldPop. En effet, bien que d'un point de vue spatial les deux couches utilisent le niveau des districts en entrée pour la désagrégation, les données en entrée de population ne sont pas identiques et nous utilisons dans nos estimations des données de population bien plus récentes.

En effet, comme précisé dans le chapitre 1, afin d'estimer la population des pays africains entre 2000 et 2020, le programme WorldPop utilise les résultats du recensement le plus récent comme base. Cependant, pour Madagascar, les résultats du recensement général de la population et de l'habitat (RGPH-3) de 2018 n'étaient pas encore disponibles au moment de l'estimation pour cette année-là. En conséquence, WorldPop a dû se baser sur le recensement le plus récent, à savoir le RGPH-2 de 1993[51], afin d'estimer la population des zones administratives et procéder à la désagrégation pour 2018. Pour estimer la population pour des années postérieures au recensement, ils projettent la population dans le temps en appliquant des méthodes de projection démographique. Ces méthodes utilisent principalement les taux de croissance intercensitaire, avec l'intention de correspondre aux projections annuelles fournies par la division de la population des Nations Unies. Les modèles de régression sont appliqués de manière itérative pour chaque année, en ajustant les prédictions en fonction des taux de croissance projetés.

Dans ce contexte, il est logique que nos méthodes de désagrégation, qui s'appuient sur des données plus récentes et intègrent la cartographie des bâtiments, présentent moins d'erreurs, et des erreurs de plus petite ampleur, par rapport aux estimations fournies par WorldPop. Il aurait d'ailleurs même été surprenant de ne pas obtenir de meilleurs résultats.

Comme l'illustrent les défis rencontrés lors de la préparation des zones sources de Madagascar, le travail de modélisation descendante nécessite une grande rigueur et un temps considérable pour collecter et croiser les informations. Notre travail n'est donc pas être réalisé dans les mêmes conditions pour le programme WorldPop, qui a

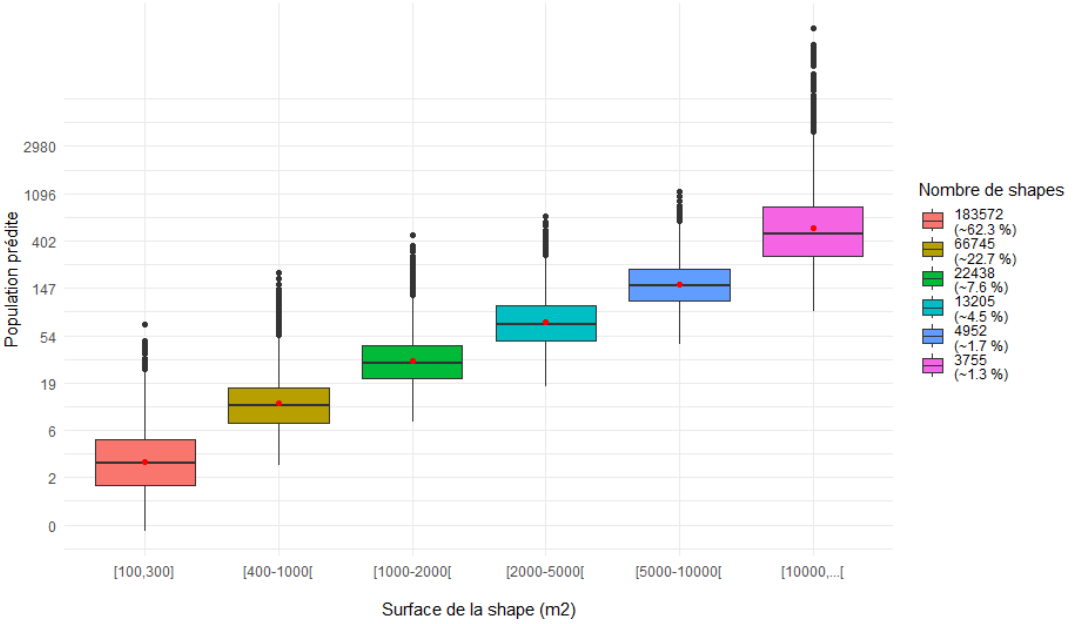
[51] https://hub.worldpop.org/resources/docs/national_boundaries/global-input-population-data-summary.xlsx

pour vocation d'être semi-automatique. Il est aussi important de noter que l'ambition de WorldPop diffère : ils ont produit des estimations pour presque tous les pays d'Afrique sur une période couvrant deux décennies (2000-2020). En revanche, notre approche, centrée sur un seul pays, nous a permis de consacrer plus de temps à la récolte des données et ainsi à l'amélioration de la précision des estimations. Sans pouvoir conclure définitivement de l'apport d'une estimation par shapes, cette application montre que nos estimations réalisées en 2018 sont plus performantes que celles disponibles en accès libre à cette même date, renforçant l'idée qu'avoir un projet de cartographie permettant de suivre quasi-en temps réel l'évolution des zones bâties et l'estimation de la population est primordial.

5.3 Distribution de la population au niveau des shapes

Concernant l'estimation de la population au niveau de la shape – sans pouvoir la valider en l'absence de données précises à ce niveau – il est rassurant de constater que pour la catégorie de shapes la plus représentée, c'est-à-dire les shapes de très petite taille (62 % du jeu de données !) l'estimation pour 50 % d'entre elles se situe entre deux et cinq habitants (1ere boite à moustaches figure 36). Sachant que ces habitations en milieu rural ont de très grandes chance d'être des maisons individuelles, comme souligné par le rapport sur l'habitation et le cadre de vie de la population (INSTAT, 2021c), il est pertinent que le modèle de désagrégation donne une estimation proche du nombre moyen d'habitants par ménages, soit 4,1 habitants en moyenne.

Figure 36 :
Distribution de l'estimation de la population suivant les shapes, selon les catégories de surface



Sources : Calculs de l'auteur à partir de la détection de bâti TeleCense dont la surface est ajustée par la détection Google

5.4 Vérifier les résultats au niveau des shapes : un défi inatteignable à relever

L'utilisation de la méthode de désagrégation par interpolation surfacique, ainsi que celle basée sur la pondération par densité, soulève plusieurs défis en matière de vérification et de validation des résultats. En effet, lors de la désagrégation, on passe d'une échelle à une autre (de celle des communes à celle des shapes), mais on ne dispose pas d'observations au niveau des shapes pour valider cette répartition.

Une approche potentiellement pertinente pour valider la méthode de désagrégation par densité de bâti pourrait être de s'inspirer des travaux qui calculent des estimations de population en fonction de catégories de couverture terrestre ou d'occupation du sol (Hallot et al., 2019; Linard et al., 2012). En suivant cette approche, il serait possible de relier la variable de densité de bâti à une densité de population réelle pour chaque valeur de la variable de densité. Toutefois, cette méthode nécessite des données réelles de population, voire des données issues de micro-recensements, qui sont souvent très difficiles à obtenir.

L'accès à des données de validation précises reste une contrainte majeure, partagée par de nombreux projets de cartographie de population, dont WorldPop. Pour évaluer précisément la fiabilité des estimations au niveau des shapes, il faudrait disposer de données de terrain, telles que celles provenant d'enquêtes ou de recensements locaux, recensant spécifiquement la population dans les zones bâties identifiées par TeleCense. En ce qui concerne les estimations de WorldPop, il faudrait obtenir des données de population au niveau des carreaux de 100x100 mètres. Ces données n'existent pas et nécessiteraient une collecte ou une création spécifique.

Cependant, la question se pose quant à la pertinence d'une telle collecte. D'une part, ce travail serait colossal et soulèverait des interrogations sur la représentativité des données obtenues, notamment de savoir à partir de combien de régions et combien de zones bâties recensées, l'échantillon est censé être représentatif et suffisant pour valider nos estimations de manière fiable ? D'autre part, si ce type de collecte et données existait, la pertinence de nos études perdrait de son sens : à quoi bon créer des modèles pour estimer des informations déjà existantes ? L'objectif de ces approches est d'offrir des estimations qui permettent de mieux comprendre les dynamiques démographiques à différentes échelles spatiales et temporelles, qu'il s'agisse d'une initiative locale et ponctuelle ou de la préparation d'un recensement à grande échelle. Dans cette perspective, il semble peu justifié de consacrer des efforts considérables à la validation locale des résultats, surtout si cette validation ne permet

pas d'améliorer significativement la fiabilité des estimations à une échelle plus globale.

L'absence de données de recensement précises empêche une évaluation rigoureuse des estimations de population au sein des unités géographiques analysées, qu'il s'agisse des carreaux de 100x100 mètres dans le cas de WorldPop (Thomson et al., 2022), ou des shapes pour notre étude. En conséquence, la vérification exacte des résultats d'estimation de population, comme ceux finaux observés dans la figure 36 précédente, n'est pas possible ici et n'est généralement pas menée à bien dans les autres travaux actuels mobilisant une approche descendante.

6. Estimer la population après 2020 : quelles solutions pour Madagascar et au-delà ?

Au fur et à mesure de nos analyses, plusieurs points émergent. Nos résultats montrent clairement que disposer d'une cartographie du bâti améliore significativement la qualité et la précision des estimations de population par désagrégation. Cette amélioration est mise en évidence par notre méthode de validation, qui, bien que perfectible et reposant sur une comparaison à un niveau agrégé, fournit néanmoins des indications pertinentes pour évaluer les performances des différentes méthodes.

Grâce à cela, nous avons tout d'abord observé que plus la cartographie du bâti est précise, meilleures sont les estimations, comme en témoigne l'utilisation des surfaces issues de Google pour ajuster celles de TeleCense. Par ailleurs, nos résultats surpassent ceux de WorldPop dans le contexte spécifique étudié. Si cet écart était prévisible au regard des différences entre les deux approches, il mérite néanmoins d'être souligné. Enfin, la littérature montre que WorldPop obtient de meilleures performances lorsque ses estimations sont contraintes, notamment par des données de cartographie du bâti (Darin et al., 2022), bien que ces estimations ne soient disponibles que pour 2020.

Ainsi, nos analyses confirment que la disponibilité d'une cartographie du bâti, la plus précise possible et si possiblement correctement datée, constitue un levier essentiel pour améliorer la qualité des estimations démographiques par désagrégation.

Au-delà de chercher à déterminer qui, de TeleCense ou d'autres organismes de cartographie de la population, fournit les meilleures estimations, on peut s'interroger sur la capacité de notre champ de recherche à produire des estimations sur le long terme. Cela est essentiel, que ce soit dans une optique de suivi de la population après un recensement ou pour pallier l'absence d'un recensement à venir.

Dès lors, quelles sont alors les données actuellement disponibles ? À la date de finalisation de ce chapitre (1er mars 2025), on peut notamment consulter la plateforme *Humanitarian Data Exchange* (HDX), qui, bien que non exhaustive, est reconnue pour centraliser de nombreuses données sur la population en compilant divers projets externes. Cela permet d'obtenir un premier état des lieux : à Madagascar, quatre sources d'estimation de population sont accessibles, celles de Kontur[52], de Meta[53], de WorldPop et de GHSL[54] (dont l'actualisation est sortie en janvier 2025).

Nous n'avons pas abordé Kontur dans le chapitre 1, mais il s'agit d'une entreprise privée qui agrège plusieurs bases de données existantes (GHSL, Facebook, Microsoft Buildings, Copernicus Global Land Service, Land Information New Zealand et OpenStreetMap). Leurs données, uniquement disponibles pour 2022, offrent des estimations à une échelle de 400 m, une résolution bien inférieure à celle de WorldPop ou de TeleCense. Quant aux données de Meta, elles datent de 2020 pour Madagascar, mais elles s'avèrent inutilisables en raison d'une détection incomplète du bâti, comme illustré dans l'annexe 4. Les estimations de WorldPop sont disponibles sans contrainte jusqu'en 2020, tandis que les estimations contraintes ne couvrent que l'année 2020. Toutefois, GHSL a sorti le 1er janvier 2025 sa mise à jour d'estimation de population, et il est désormais possible de télécharger des données d'estimation de population à l'échelle du carreau de 100x100 mètres pour Madagascar.

Bien sûr, il existe les données de l'INSTAT, qui sont précieuses. Cependant, elles n'existent que pour 2018 et, surtout, ce sont des données de recensement, lesquelles, comme déjà mentionné précédemment, ne seront probablement pas mises à jour avant au moins dix ans. Puis, comme c'est souvent le cas avec les recensements, elles incluent également des projections démographiques. Par exemple, l'INSTAT fournit des projections jusqu'en 2023 par région et des estimations nationales jusqu'en 2050 (INSTAT, 2021h). Cependant, ces chiffres calculés pour le premier niveau administratif, ne permettent pas de répondre à l'objectif de suivi de la croissance et de la répartition de la population. Surtout, la question ici porte sur l'estimation de la population à distance, à partir d'images satellites et de télédétection

Si l'on récapitule, il existe deux ensembles de données d'estimation de population post-2020 : celles de Kontur, qui ont une résolution assez faible et ne sont disponibles que pour l'année 2022, et celles de GHSL, dont la mise à jour est publiée

[52] https://data.humdata.org/dataset/kontur-population-madagascar
[53] https://data.humdata.org/dataset/highresolutionpopulationdensitymaps-mdg
[54] https://data.humdata.org/dataset/mdg-ghsl

tous les cinq ans et qui viennent de sortir, avec une résolution satisfaisante (100 m, la même que celle des estimations de WorldPop), mais qui ne couvrent également que l'année 2025. Le constat est clair : il n'existe pas de données d'estimation de population permettant un suivi à jour après 2020 pour Madagascar.

Dès lors, comment assurer le suivi de l'évolution de la population après le recensement de 2018 ? Ce questionnement est important pour le gouvernement de Madagascar et l'anticipation des politiques publiques ou pour d'autres acteurs locaux et internationaux engagés dans divers enjeux de développement. C'est dans cette optique que des programmes d'estimation de population couplés à des initiatives de cartographie du bâti ont été développés. Or, jusqu'en septembre 2024, la situation post-2020 n'était guère plus favorable en matière de données sur le bâti. On l'avait mentionné dans le chapitre 1 ; on trouve les détections de GHSL puis celles de Google Open Buildings pour l'année 2022. Ces dernières présentent cependant des limites importantes : elles concernent une seule année et, surtout, ne sont pas véritablement datées. Il s'agit d'une mosaïque d'informations issues des images satellitaires disponibles, sans indication précise sur la temporalité. Cette limite a été partiellement levée en septembre 2024, lorsque Google Research a proposé une extension de sa base de données sous la forme d'un nouveau projet : *Open Buildings 2.5D – Temporal Dataset*. En combinant des images satellitaires à basse et haute résolution, ce projet a permis de créer une base de données temporelle couvrant la période de 2016 à 2023. Il s'agit d'une avancée majeure, car elle permet de suivre l'évolution du bâti dans le temps tout en assurant une cohérence spatiale et temporelle, répondant ainsi à la principale limite de la mosaïque d'images.

Néanmoins, ce jeu de données, publié tardivement dans le cadre de ce travail de thèse, est présenté ici à titre informatif. Il fait l'objet d'un chapitre dédié en fin de thèse, dans lequel il est mobilisé pour enrichir nos approches d'estimation de la population.

Pour l'heure et pour les deux prochains chapitres, nous continuons de nous concentrer sur les identifications des zones bâties de TeleCense. Le programme TeleCense, associé aux méthodes développées dans cette thèse, vise à produire des estimations de population à partir d'images satellites avec un pas de 6 mois. De nouveau, hormis des contraintes de temps et de stockage, il n'y a aucune difficulté à actualiser la détection du bâti de Madagascar année après année. Une fois les zones bâties identifiées sous la forme de shapes, nous pouvons adopter une approche similaire à WorldPop : appliquer des taux de croissance démographique aux niveaux administratifs les plus fins pour projeter la population année après année (d'abord en

2019, puis en 2020, et ainsi de suite), puis désagréger ces projections au sein des shapes de l'année concernée.

Cependant, comment nos estimations évolueront-elles au fil du temps, à mesure que nous nous éloignons de l'année de référence du recensement ? Il est raisonnable de penser que, dans les premières années, les projections resteront relativement proches de la réalité. Mais au-delà, dans des pays comme Madagascar ou plus largement en Afrique, où la croissance démographique est relativement rapide et la mobilité spatiale importante la stabilité des projections devient plus incertaine. Des modèles plus précis seraient nécessaires, intégrant des facteurs clés tels que la natalité, la mortalité et les migrations. Cependant, la collecte et l'actualisation de ces données restent complexes et ne correspondent pas aux objectifs des projets comme TeleCense ou WorldPop, qui visent avant tout des approches reproductibles et semi-automatisées.

En résumé, jusqu'à la sortie de Google 2.5D il manquait définitivement une cartographie fiable et actualisée des zones bâties, et c'est précisément ce que le programme TeleCense cherche à combler. Pour estimer la population à un niveau local, deux éléments essentiels sont nécessaires : d'une part, les décomptes de population, qui peuvent être désagrégés, et d'autre part, des données de bâti. Nous avons vu que la précision des estimations est nettement améliorée lorsqu'on dispose de l'identification des zones bâties.

Si ces deux informations sont disponibles, l'enjeu principal réside alors dans la durée pendant laquelle ces estimations peuvent rester fiables et précises. Dans le prochain chapitre, nous proposerons une étude de cette question en nous appuyant sur le cas du Bénin. En utilisant le recensement de 2013, nous chercherons à estimer la population de 2017 à 2023, en analysant l'évolution des estimations de population et du bâti. Nous tenterons ainsi de déterminer combien d'années après un recensement il est raisonnable de continuer à utiliser ses projections pour estimer la population.

Cette analyse soulignera probablement la nécessité de méthodes alternatives. En effet, que faire lorsque l'on dispose de données sur le bâti, mais pas de décomptes de population ? Par exemple, estimer la population de Madagascar en 2026, huit ans après le dernier recensement, semble peu réaliste en s'appuyant uniquement sur des projections. De même, pour des pays comme la République Démocratique du Congo, non recensée depuis 1984, il serait souhaitable de développer une modélisation qui n'ait pas besoin de données de population comme entrée. C'est ce que nous proposons d'étudier dans la suite de la thèse.

7. Conclusion

Ce chapitre a présenté l'adaptation des approches descendantes de désagrégation, généralement appliquées pour une visualisation finale carroyée, à une visualisation directement dans les zones bâties identifiées à partir d'images satellites. Pour répartir la population à l'intérieur des shapes identifiés à Madagascar, nous avons d'abord récupéré les zones administratives les plus précises disponibles dans le recensement, à savoir les communes. Associer les décomptes de population aux communes malgaches a constitué un premier défi majeur, en raison de l'absence de shapefiles actualisés correspondant au découpage utilisé lors du RGPH-3. Malgré ces difficultés, la reproductibilité de notre démarche et la transparence des sources utilisées montrent que cette approche peut être précieuse pour répondre à des besoins spécifiques de cartographie. Cependant, elle demeure fastidieuse et souffre encore d'un manque de validation complète. Ce premier travail souligne ainsi la nécessité d'une mise à jour régulière des données géographiques pour garantir la pertinence des études spatiales contemporaines.

Une fois les zones sources (les communes) et les zones cibles (shapes) définies, nous avons pu procéder à la désagrégation de la population. Quatre méthodes ont été testées, et ce sont finalement les deux les plus simples – l'interpolation surfacique et la pondération par densité de bâti – qui ont obtenu les erreurs moyennes les plus faibles. À l'inverse, les deux approches basées sur la modélisation de la densité de population des communes ont systématiquement donné des résultats avec des erreurs moyennes plus élevées. Bien que les caractéristiques retenues soient fortement corrélées à l'échelle des communes, elles peinent à se répercuter avec la même pertinence au niveau des shapes lors de l'exercice de désagrégation. Ces résultats mettent en avant l'importance d'une cartographie détaillée des zones bâties pour améliorer la précision de la désagrégation.

Nous avons ensuite comparé nos résultats à ceux de WorldPop et démontré que notre désagrégation surpassait largement leurs estimations disponibles pour 2018, mettant en évidence l'importance d'utiliser les décomptes de recensement les plus récents. De plus, de la même manière que WorldPop améliore ses modèles en passant d'une approche sans contrainte à une modélisation intégrant des contraintes spatiales liées aux zones bâties, notre méthode tire parti d'une meilleure délimitation des espaces construits pour l'amélioration des estimations démographiques.

Enfin, ce chapitre a mis en évidence l'enjeu essentiel du suivi de la population malgache après le recensement de 2018. Plus largement, après 2020 l'absence de

données temporelles sur l'évolution du bâti et de la population constitue un enjeu majeur en Afrique et dans notre champ de recherche. Les chapitres suivants exploreront cette problématique et proposeront des solutions, d'abord en identifiant le bâti sur plusieurs années au Bénin (chapitre 4), puis en créant un modèle dit ascendant, permettant d'essayer d'estimer la population sans données de recensement en entrée, cela testé et validé dans le département de Mayotte puis dans la sous-préfecture d'Abidjan (Partie 3).

Chapitre 4 – Cas d'application sur plusieurs années de la désagrégation de la population du Bénin

Ce chapitre trouve son origine dans un projet que nous avons mené au Bénin avec l'entreprise Diginove. Réalisé en partenariat avec l'Agence Nationale d'Aménagement du Territoire (ANAT), ce projet mobilise la solution TeleCense afin de répondre aux enjeux croissants de l'urbanisation et de l'organisation des services publics. Bien que cette étude n'ait pas été initialement pensée pour s'intégrer à la thèse, l'approfondissement du travail et l'élargissement de la zone d'analyse ont révélé un intérêt scientifique certain, justifiant son inclusion. Son intégration dans la thèse permet d'explorer de nouvelles perspectives, notamment en évaluant la précision de l'identification du bâti par TeleCense sur plusieurs années, ainsi que la pertinence des approches de désagrégation démographique dans ce contexte temporel étendu.

L'ANAT, acteur clé dans ce contexte, est chargée d'élaborer une vision stratégique pour le développement territorial du Bénin. Elle assure la coordination et l'harmonisation des politiques d'aménagement du territoire, en mobilisant des compétences en planification urbaine, en gestion de projets et en urbanisme. Cette collaboration entre une entreprise privée (Diginove) et une institution publique illustre concrètement l'application de notre recherche et son potentiel impact sur les politiques d'aménagement et le développement urbain au Bénin.

Plus précisément, ce projet est issu d'une initiative proactive de Diginove, qui a identifié au Bénin un besoin en matière d'aménagement du territoire et de gestion urbaine. Grâce au FASEP[55], fonds d'études et dispositif de soutien à l'internationalisation des entreprises françaises, Diginove a proposé, via la solution TeleCense, une réponse innovante aux enjeux liés à l'urbanisation croissante et à l'organisation des services publics. Cette collaboration avec l'ANAT s'inscrit pleinement dans les ambitions du Programme d'Actions du Gouvernement (PAG 2).

En combinant plusieurs domaines d'expertise, la solution TeleCense cherche à contribuer à améliorer la gestion de l'urbanisation, de l'étalement urbain, de la pression foncière, ainsi que des problématiques environnementales et des risques d'inondation. Dans le cadre de cette thèse, nous nous concentrons uniquement sur l'aspect estimation de la population.

[55] https://www.tresor.economie.gouv.fr/services-aux-entreprises/le-fasep

Le Bénin, pays d'Afrique subsaharienne, partage ses frontières avec le Togo, le Burkina Faso, le Niger et le Nigéria. Sa population, recensée à 10 008 749 habitants lors du RGPH-4 en 2013 (INSAE, 2015a), est estimée à environ 14,5 millions d'habitants en 2024 (World Population Prospect, 2024). Le dernier recensement officiel, RGPH-5, est actuellement en cours et les résultats devraient être disponibles en 2025 ou 2026. Administrativement, le pays est structuré en trois niveaux, comprenant 12 départements (dont la répartition est illustrée à la figure 37), 77 communes et 545 arrondissements. Les résultats des recensements sont disponibles à ces trois niveaux administratifs, sous format agrégé.

Figure 37 : Carte de situation du Bénin et localisation de ses départements

Sources : Carte réalisée à partir des tracés des limites administratives fournis par GADM

Le projet porte sur la région du Grand Nokoué dans le sud du pays, qui s'étend principalement dans les départements du Littoral, de l'Atlantique et de l'Ouémé. Cette région, stratégique en raison de sa densité de population et de son potentiel

économique, connaît une urbanisation rapide, notamment dans ses cinq principales communes : Abomey-Calavi, Porto-Novo, Ouidah, Cotonou et Sèmè-Kpodji. La figure 38 illustre l'état de l'urbanisation en présentant une image aérienne Sentinel-2 de 2023, mettant en évidence la densité et la continuité du bâti dans les communes de Cotonou, Porto-Novo et Abomey-Calavi**.** Cependant, cette croissance urbaine accélérée s'accompagne de défis majeurs pour la gestion urbaine et le développement durable, avec des préoccupations environnementales telles que les inondations, la dégradation de l'environnement et la gestion des déchets (Maximenne, Djego, Lougbegnon, Sinsin, 2017).

Figure 38 : Zone du Grand Nokoué – image sentinel-S2 optique 2023, résolution 10m

Sources : TeleCense, image issue du logiciel QGIS

Ce projet, qui comprend plusieurs volets de travail (risques d'inondation, suivi des espaces naturels), nécessite l'estimation de la population dans les zones administratives, puis dans les zones bâties pour les années 2017 à 2023. C'est sur ce dernier point que ce chapitre se focalise.

Afin de mettre en œuvre ce projet, les premières étapes sont similaires à celles effectuées à Madagascar : il s'agit de récupérer les données du dernier recensement et de les associer aux zones administratives composant la région du Grand Nokoué. Ensuite, la région est cartographiée pour identifier et caractériser les zones bâties, puis cette opération est répétée pour chaque année de 2017 à 2023.

La principale différence réside dans les données de population, car nous avons besoin d'estimations pour plusieurs années. Or, au Bénin, le dernier recensement

accessible est le RGPH4 de 2013 et le RGPH5 étant en cours de réalisation, ses résultats ne sont pas encore disponibles. Cela signifie que même pour obtenir une estimation de la population en 2017 pour chacune des zones administratives, une projection à partir des données officielles de 2013 est nécessaire.

Nous devons donc nous appuyer conjointement sur les recensements de 2002 et 2013 pour proposer une projection de la meilleure qualité possible. L'objectif est de rester dans une approche pragmatique, sans mobiliser de données détaillées sur la natalité, la mortalité ou la migration, qui ne sont généralement pas accessibles à un niveau aussi fin que celui des arrondissements (niveau administratif 3 équivalent aux communes de Madagascar).

Cette approche repose donc sur l'utilisation des taux de croissance démographique. Or, en raison de la nature même d'une projection, on peut s'attendre à ce que la désagrégation de la population pour les années postérieures à 2013 devienne progressivement moins précise. Dès lors, ce chapitre en forme de cas d'application est particulièrement utile, car il peut permettre d'évaluer la pertinence de l'estimation et de l'identification du bâti dans le temps. Ce point constitue un enjeu fondamental non seulement pour le programme TeleCense et pour notre approche et notre champ de recherche de manière plus générale

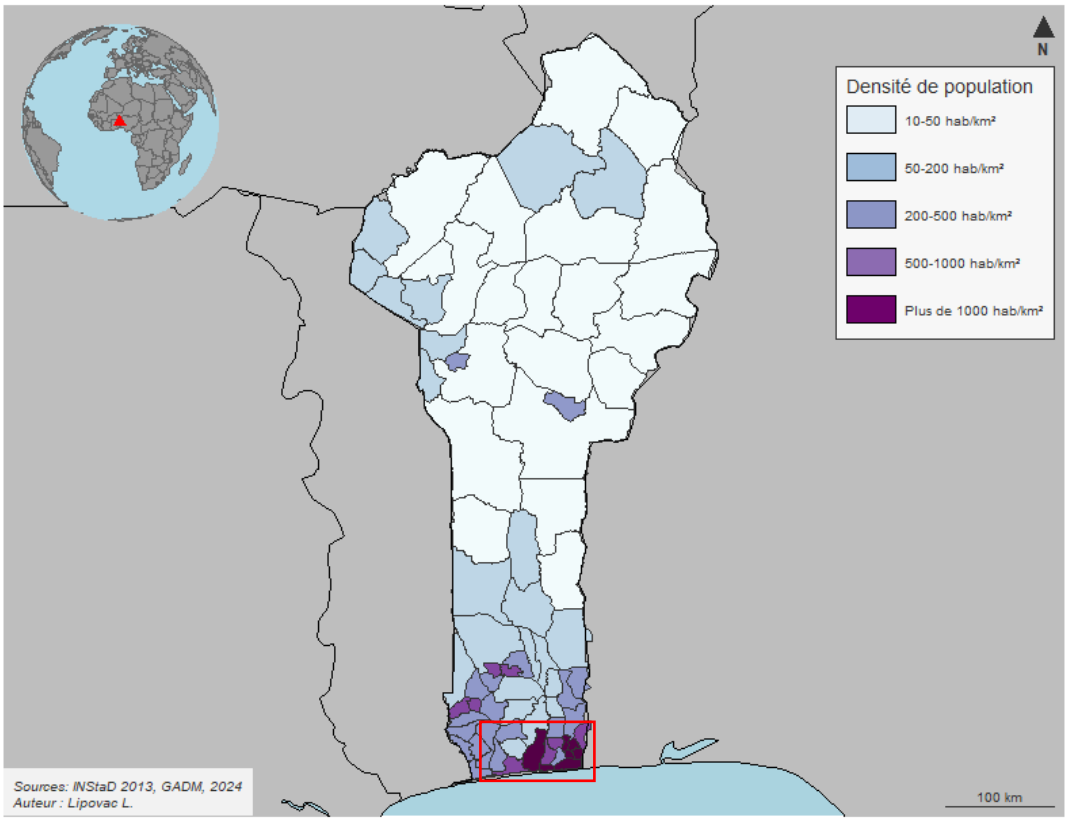
1. Préparation des zones sources : les arrondissements du Bénin

Dans ce chapitre nous nous concentrons sur une région spécifique du Bénin : le Grand Nokoué, illustrée dans la figure 39 ci-dessus. Nous avons choisi une zone légèrement élargie, incluant 13 communes et 92 arrondissements. Ce choix est dicté principalement par les objectifs du projet, mais il s'avère également pertinent pour notre recherche, car il nous permet de travailler sur la partie sud du pays, qui est plus densément peuplée et urbanisée, ce qui constitue une différence notable par rapport à Madagascar. Cette différence de densité entre le nord et le sud du pays est visible dans la figure 39 où les communes deviennent de plus en plus petites en termes de surfaces, mais de plus en plus densément peuplées. L'encadré en rouge met en évidence la zone sur laquelle nous nous concentrons dans ce chapitre dont le focus sur la population en 2013 est détaillé dans la figure 40.

À noter que le découpage des différentes zones administratives, qu'elles soient récupérées via GADM ou geoBoundaries, ne se superpose pas parfaitement. C'est pourquoi, dans la figure 39, nous observons une légère redondance sur les frontières du pays. De plus, dans les figures 39 et 40, nous avons choisi de ne pas afficher simultanément les départements et les communes, ni les communes et départements

par-dessus les arrondissements, car comme ces couches ne se superposent pas correctement, cela crée de la confusion et masque l'information principale.

Figure 39 : Densité de population des communes (niveau 2) du Bénin – décompte de population du RGPH4 - 2013



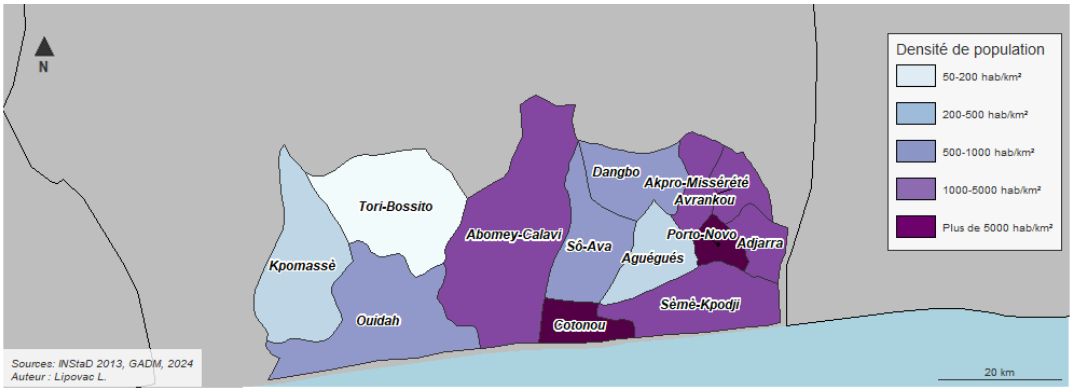
Sources : Calculs de l'auteur à partir des décomptes de population issus du RGPH4 (INSTAD) et tracés des limites administratives fournis par GADM

La densité de population des 13 communes est présentée dans la figure 40, où nous avons légèrement modifié les catégories de densité pour mettre en évidence que les communes de Cotonou et Porto-Novo, respectivement capitale économique et politique, et capitale officielle, sont les plus densément peuplées, avec environ 9 000 et 5 300 habitants/km².

Bien que le département du Littoral, comprenant Cotonou et ses 13 arrondissements, soit très densément peuplé, sa croissance démographique est désormais très faible en raison de son caractère de centre économique. De fait, sa part dans la population totale a diminué entre 2002 et 2013, passant de 9,8 % à 6,8 %, au profit notamment des communes d'Abomey-Calavi et de Sèmè-Kpodji. Ces dernières ont connu un fort développement urbain et sont devenues des cités-dortoirs (INSAE, 2015a). Cette évolution est déjà visible à travers les densités de population de 2013 qui y dépassent plus de 1000 habitants/km².

Le phénomène observé à Cotonou, où la croissance de la population dans les périphéries est très marquée, est fréquent dans de nombreuses villes africaines. Cette expansion se fait souvent de manière spontanée, sans un accompagnement suffisant de politiques d'aménagement urbain, pouvant rendre ainsi la gestion de ces zones périurbaines complexe (*IFRI*, 2024).

Figure 40 : Densité de population des communes de la zone élargie du Grand Nokoué – décompte de population du RGPH4 - 2013



Sources : Calculs de l'auteur à partir des décomptes de population issus du RGPH4 (INSTAD) et tracés des limites administratives fournis par GADM

Au total, nous travaillons sur 13 communes, qui sont ensuite subdivisées en 92 arrondissements, le niveau administratif le plus fin au Bénin. Ce niveau correspond en quelque sorte aux communes de Madagascar, également classées comme niveau administratif 3. La principale différence réside dans la taille des zones : la superficie moyenne des 92 arrondissements sélectionnés est d'environ 27 km², contre 292 km² en moyenne pour les communes sélectionnées à Madagascar, une différence notamment due à la localisation dans le sud du Bénin, où les arrondissements sont plus petits et plus densément peuplés, contrairement au nord du pays où les arrondissements sont plus grands.

Dans la suite, nous récupérons les données nécessaires à la mise en œuvre de la désagrégation : les zones sources, les zones cibles et nous procédons à la projection de la population des arrondissements de 2017 à 2023

1.1 Le recensement de la population du Bénin

Le cinquième recensement de la population du Bénin est actuellement en cours de réalisation. Par conséquent, pour obtenir des informations détaillées sur la population, nous devons nous appuyer sur les données du RGPH4, réalisé en 2013. Pour récupérer les décomptes de population par niveau administratif, il convient de se

rendre sur le site officiel de l'Institut National de la Statistique et de la Démographie (INStaD) du Bénin, où sont disponibles les résultats du RGPH-4 (2013) ainsi que ceux du RGPH-3 (2002). Le tableau 2, intitulé « Effectifs de population par département, commune et arrondissement au RGPH-4 » dans la plaquette du RGPH-4 (INSAE, 2013), présente les décomptes de population organisés par département, commune et arrondissement. À partir de ces données, il s'agit ensuite de se restreindre aux 13 communes retenues et présentées dans les figures précédentes, puis d'exporter les informations dans un format compatible afin de constituer la base de données de la population pour les 92 arrondissements de ces communes en 2013.

Comme nous allons projeter les décomptes de population des arrondissements, nous récupérons également les décomptes de population de 2002 du RGPH-3 à partir des cahiers des villages-2002 des départements de l'Atlantique, du Littoral et de l'Ouémé (INSAE, 2004a, 2004b, 2004c). Ces décomptes permettent de calculer les taux intercensitaires pour les arrondissements.

1.2 Le tracé des limites administratives du Bénin

L'association des décomptes de population aux zones administratives ne requiert pas un travail aussi complexe que pour Madagascar. Au Bénin, malgré quelques problèmes de superposition entre les couches géographiques, les limites sont à jour et facilement accessibles via des bases de données telles que GADM ou geoBoundaries. Pour les arrondissements, nous retenons le fichier issu de geoBoundaries (Runfola et al., 2020), tout en apportant des ajustements nécessaires à certains noms pour garantir une correspondance précise avec les données de population. Voici les corrections effectuées :

Département de l'Atlantique :

- Commune d'Abomey-Calavi
    - 1er Arrondissement devient Ouidah i
    - 2e Arrondissement devient Ouidah Ii
    - 3e Arrondissement devient Ouidah Iii
    - 4e Arrondissement devient Ouidah Iv
- Commune de Kpomasse
    - Tokpa Dome devient Tokpa-Dome
- Commune de Ouakpe-Daho
    - Ouakpe Daho devient Ouakpe-Daho
- Commune de So-Ava
    - Ahome-Lokpo devient Ahomey-Lokpo
    - Dekanme devient Dekanmey

    - Ganvie I devient Ganvie i
    - Ganvie II devient Ganvie Ii
- Commune de Tori-Bossito
    - Azohoue-Kada est devenu Azohoue-Cada
    - Bossito est devenu Tori-Bossito
    - Tori-Kada est devenu Tori-Cada

Département de l'Ouémé :

- Commune d'Adjarra
    - Adjarra I devient Adjarra i
    - Adjarra II devient Adjarra Ii
- La commune Akpo-Misserete devient Akpro-Misserete
- Commune de Seme-Kpodji
    - Aglangandan est devenu Agblangandan

Après ces corrections, nous pouvons associer les décomptes de population aux entités administratives afin de constituer une base de données géoréférencée des populations recensées en 2013 au niveau des arrondissements. Afin d'avoir les zones sources prêtes à être désagrégées pour les années 2017 à 2023 il faudra donc projeter ces décomptes de population pour les années concernées.

Contrairement aux analyses menées à Madagascar, où nous avions testé des modèles linéaires et de forêts aléatoires à l'échelle communale, nous ne prévoyons pas ici de recréer des modèles spécifiques pour les arrondissements. Aucune variable supplémentaire ne nécessite donc d'être recalculée à ces niveaux administratifs. Nous appliquerons directement la méthode d'interpolation surfacique, retenue comme approche de désagrégation dans le chapitre 2. Pour cela nous devons dans un premier temps cartographier le bâti pour les années 2017 à 2023.

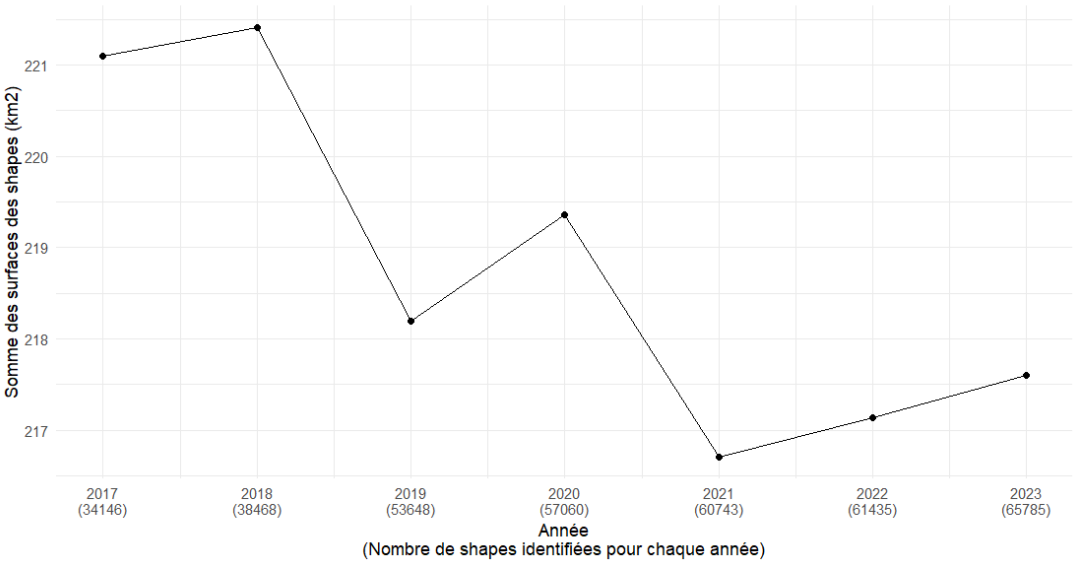
2. Identification des zones bâties des années 2017 à 2023

En 2017, environ 34 000 shapes sont identifiées dans la zone élargie du Grand Nokoué, avec une forte présence dans les communes de Cotonou et Porto-Novo, suivies par Abomey-Calavi et l'ouest de Ouidah. Nous poursuivons ensuite la cartographie du bâti pour les années suivantes, de 2018 à 2023.

L'évolution du bâti identifié par TeleCense au Bénin révèle cependant un résultat inattendu. Bien que la tendance générale attendue soit respectée en termes du nombre de shapes, avec une augmentation nette passant de 34 146 zones bâties

détectées en 2017 à 65 758 en 2023, la surface totale bâtie estimée en 2017 est bien plus importante qu'en 2023. De plus, cette surface fluctue selon les années et ne suit pas une évolution linéaire, à l'exception des trois dernières années (figure 41). Notons que nous utilisons toujours la surface ajustée par la détection de Google Open Buildings afin d'affiner l'estimation de l surface des shapes que l'on détecte (l'annexe 5 présente par ailleurs la même figure que la 41 avec la surface TeleCense, où l'on voit une très grande différence de surface entre les deux détections).

Figure 41 : Évolution du nombre de shapes et de leur surface totale au fil des années au Bénin



Sources : Calculs de l'auteur à partir de la détection de bâti TeleCense dont la surface est ajustée par la détection Google

2.1 Un problème d'identification des shapes au fil des années ?

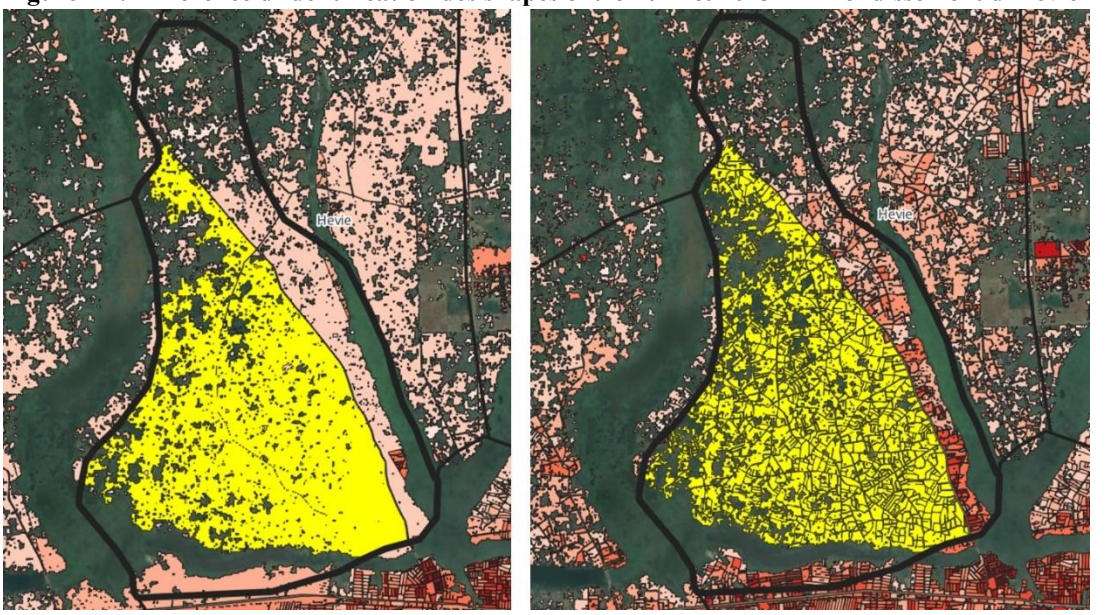
Dans le chapitre 2, concernant la présentation des variables du programme TeleCense, nous avions déjà évoqué un premier problème lié à l'homogénéité de l'identification des shapes. En effet, parmi les milliers de shapes identifiées par TeleCense, la grande majorité sont de petite ou moyenne taille, tandis que certaines sont grandes, voire très grandes (exemple de la figure 14 du chapitre 2 section 2.1). Ce phénomène est souvent observé dans des zones à forte densité de bâti, où les routes ne sont pas bien recensées par Open Street Map (OSM).

La détection des shapes par TeleCense est en effet très sensible à l'état du réseau routier fourni par OSM. Si ce dernier n'est pas à jour, il devient difficile de découper et de départager correctement les grandes structures bâties. Ce point peut légitimement poser problème, car OSM repose sur un système dynamique mis à jour par une communauté de contributeurs qui peut varier d'une année à l'autre et d'un pays à l'autre.

L'arrondissement d'Hêvié dans la commune d'Abomey-Calavi – 67 218 habitants en 2013 selon les décomptes officiels du RGPH4 – illustre parfaitement ce phénomène (figure 42). TeleCense identifie 327 shapes en 2017, puis 2 004 shapes en 2023, et ce alors que la surface bâtie détectée de la commune est sensiblement la même aux deux dates. La figure 42 l'illustre : sur la figure de gauche en 2017, tout l'élément surligné en jaune est une seule shape comprenant un très grand nombre de zones bâties. Bien qu'elles fassent figure d'exception, ces très grandes shapes demeurent néanmoins un problème, car elles sont beaucoup plus grandes que les autres et ne permettent pas de caractériser précisément les zones bâties qu'elles contiennent.

En 2023, sur la figure de droite (figure 42), au lieu d'avoir une seule grande shape, celle-ci est correctement découpée en centaines de shapes, permettant une bien meilleure identification. Dès 2019, le réseau routier accessible via OSM est bien plus complet, ce qui assure une meilleure détection. Cela est par ailleurs confirmé par la figure 41, où le passage de 2018 à 2019 montre une légère chute de la surface totale bâtie (de 222 km² à 218,5 km²) couplée à une forte augmentation du nombre de shapes (de 38 500 à 54 600 shapes). Un phénomène similaire est observé de 2020 à 2021 et ensuite il semble avoir une meilleure cohérence, avec une augmentation linéaire pour les dernières années observées.

Figure 42 : Différence d'identification des shapes entre 2017 et 2023 – Arrondissement d'Hêvié



Sources : Tracés des limites administratives fournis par geoBoundaries et détection du bâti par TeleCense, images issues du logiciel QGIS

L'arrondissement d'Hêvié met en évidence des défauts d'identification qui nécessitent une prise en compte dans l'analyse, difficiles à corriger dans l'immédiat et

auxquels il faudra s'accommoder dans le cadre de notre projet. Cependant, ce type d'erreurs reste extrêmement rare, et dès 2017, la majorité de la zone du Grand Nokoué, y compris le sud-ouest d'Hêvié (montré par la figure 43), avec les communes d'Abomey-Calavi et Cotonou, ainsi que la commune de Porto-Novo (figure 44), ont un bâti identifié et découpé de manière homogène, répondant ainsi aux exigences nécessaires pour produire des modèles fiables et homogènes, ainsi qu'une estimation précise de la population.

Figure 43 : Identification satisfaisante des shapes dans la région sud du Grand Nokoué – 2017

Sources : Tracés des limites administratives fournis par geoBoundaries et détection du bâti par TeleCense, image issue du logiciel QGIS

Figure 44 : Identification satisfaisante des shapes dans la commune de Porto-Novo – 2017

Sources : Tracés des limites administratives fournis par geoBoundaries et détection du bâti par TeleCense, images issues du logiciel QGIS

En somme, dès que le réseau routier récupéré via OSM est de meilleure qualité, la détection des zones bâties devient plus efficace, et les shapes sont mieux segmentées et identifiées. Cela est primordial, principalement dans notre étude, car nous effectuons une désagrégation basée sur l'interpolation surfacique. Ainsi, la superficie des shapes joue un rôle clé dans l'estimation finale et la répartition de la population à partir des arrondissements.

Cependant, le problème de la non-augmentation linéaire des shapes demeure. Dans quelle mesure cela peut-il impacter l'estimation de la population ? Concernant l'approche descendante et la désagrégation, l'effet pourrait être limité : même en l'absence d'identification de nouvelles constructions, puisque nous redistribuons une population fixe chaque année pour chacun des arrondissements, cela se traduit simplement par une population plus élevée au sein des shapes existantes.

Le risque principal réside plutôt dans une possible incohérence. Il est compréhensible que, dans une commune comme Cotonou, où « un début probable de saturation de la ville, en termes de densité » a été documentée (INSAE, 2015b), l'augmentation des surfaces bâties soit limitée. En revanche, pour une commune comme Abomey-Calavi et ses environs, considérée comme une cité-dortoir en pleine expansion, il est essentiel de capturer correctement la dynamique de l'extension du bâti au fil des années.

Cette question est encore plus importante pour nos approches ascendantes et soulève des interrogations pour l'avenir. Étant donné que notre méthode repose fortement sur la cartographie du bâti, il est essentiel d'assurer une détection et une identification précise des shapes, une homogénéité dans leur surface et une cohérence temporelle. Sans cela, on risque d'entraîner un modèle sur une année où le bâti a été sous- ou surestimé, ce qui biaiserait l'ensemble du processus et compromettrait la fiabilité des estimations produites par l'approche ascendante. C'est donc un point à garder en tête pour la suite.

Ainsi, nous conservons pour l'instant ces identifications du bâti et passons à la projection de la population des arrondissements, étape essentielle avant la mise en œuvre de la désagrégation.

3. Projection des décomptes de population des arrondissements du Bénin

Les méthodes de projection démographique se divisent généralement en deux grandes catégories : les projections simples des effectifs à partir de taux de croissance ou les projections par sexe et âge, par la méthode des composantes. Dans les deux cas, les projections reposent sur une modélisation des tendances sous-jacentes (Vallin, Méslé, Toulemon, Véron, 2011; Wattelar, Caselli, Vallin, Wunsch, 2004). Pour mettre en œuvre la méthode des composantes, il faut disposer de statistiques précises sur les naissances, décès et flux migratoires chaque année (Pelletier, Spoorenberg, 2016). De

plus, cette méthode est généralement appliquée au niveau national ou régional où ces données sont généralement plus accessibles.

Au Bénin, nous disposons des recensements de 2002 et 2013, mais l'absence de données annuelles sur les dynamiques démographiques empêche l'utilisation de la méthode des composantes. Plus précisément, c'est finalement fréquent de ne pas avoir accès à ce type de données, mais généralement la méthode est accommodée aux indicateurs qui existent. Comme nous travaillons sur la zone du Grand Nokoué, bande côtière dont les dynamiques liées à la migration et à l'urbanisation sont plus importantes que celles liées à l'accroissement naturel, la méthode fait moins de sens du fait de manque de données précises sur la migration à l'échelle des arrondissements du Bénin.

En revanche, les deux points d'observation entre 2002 et 2013 permettent de calculer un taux moyen de croissance annuelle sur cette période. Nous décidons de faire l'hypothèse forte que ce taux se maintient dans le temps, quelque soit sa valeur initiale. Cela nous permet de l'appliquer aux années suivantes (de 2014 à 2023) afin d'obtenir une estimation de la population pour cette période.

Parmi les modèles de croissance possibles — linéaire, géométrique et exponentielle — la croissance géométrique est la plus appropriée. En effet, contrairement à la croissance linéaire, qui suppose une augmentation constante en valeur absolue chaque année, la croissance géométrique prend en compte un taux de variation relatif, ce qui est plus réaliste pour des phénomènes démographiques où la population évolue en proportion de sa taille. En retenant la croissance géométrique on assume donc le fait que l'augmentation annuelle est proportionnelle à la taille de la population existante, cela veut dire qu'une population plus grande génère une augmentation absolue plus importante chaque année.

Puis, pour différencier les croissances exponentielle et géométrique — qui sont généralement assez proches en termes de résultats finaux — nous avons calculé les projections selon les deux méthodes et comparé le taux d'erreur moyen final pour nos 92 arrondissements lors des projections entre 2002 et 2013. Comme nous disposons des valeurs réelles, il est possible de comparer les résultats, ce qui n'est plus possible à partir de 2014. Étant donné que la méthode géométrique présente un taux d'erreur légèrement plus faible nous avons retenu cette méthode pour la suite de notre analyse. Voici comment elle se met en œuvre :

Nous devons d'abord calculer le taux géométrique de variation annuelle moyenne r comme suivant :

$$r = \left[\left(\frac{P_0}{P_b}\right)^{\frac{1}{y}}\right] - 1$$

Où P_0 est la population de départ (en 2002),
P_b est la population initiale (en 2013) et,
y est le nombre d'année entre 2002 et 2013, soit 11,25 années (le RGPH3 a eu lieu en février 2002 et le RGPH4 en mai 2013).

Une fois le taux de variation r calculé, la population projetée P_t pour chaque année t est donnée par :

$$P_t = P_0 * (1 + r)^t$$

Le tableau 20 présente la projection de la population des huit arrondissements de la commune d'Abomey-Calavi (le reste est disponible en annexe 6). Prenons l'exemple de de la projection pour l'arrondissement d'Abomey-Calavi. Avec une population de départ de 61 450 en 2002 et de 117 824 en 2013, le taux de croissance annuel est estimé comme suit :

$$r = \left[\left(\frac{117\,824}{61\,450}\right)^{\frac{1}{11,25}}\right] - 1 = 0{,}0596$$

Cela correspond à un taux de croissance de 5,96 % par an. En appliquant la formule de projection pour 2014, nous obtenons :

$$P_{2014} = 117\,824 * (1 + 0{,}0596)^{2014-2013} = 124\,843$$

Enfin, cette méthode est répétée pour les années suivantes, jusqu'en 2023, et appliquée à l'ensemble des arrondissements. Cette démarche semi-automatique permet d'obtenir les bases de données géoréférencées prêtes à être désagrégées dans le but répartir la population de chaque arrondissement entre 2017 et 2023 au sein des shapes identifiées.

Tableau 20 : Projections des décomptes de population des arrondissements de la commune d'Abomey-Calavi au Bénin pour les années 2014 à 2023

Pays	Département	Commune	Arrondissement	Effectif total RGPH4 (2013)	Effectif total RGPH3 (2002)	r	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Benin	Atlantique	Abomey-Calavi	Abomey-Calavi	117824	61450	0,060	124843	132280	140160	148509	157356	166730	176662	187186	198337	210152
Benin	Atlantique	Abomey-Calavi	Akassato	61262	17197	0,120	68586	76785	85965	96242	107748	120629	135050	151195	169271	189507
Benin	Atlantique	Abomey-Calavi	Glo-Djigbe	28103	12827	0,072	30132	32308	34641	37142	39824	42699	45782	49088	52633	56433
Benin	Atlantique	Abomey-Calavi	Godomey	253262	153447	0,046	264797	276858	289467	302652	316436	330849	345918	361673	378146	395369
Benin	Atlantique	Abomey-Calavi	Hevie	67218	13450	0,154	77553	89477	103234	119106	137419	158547	182924	211049	243498	280937
Benin	Atlantique	Abomey-Calavi	Ouedo	27522	10067	0,094	30096	32910	35988	39353	43033	47058	51458	56270	61532	67287
Benin	Atlantique	Abomey-Calavi	Togba	73331	18674	0,129	82812	93518	105609	119263	134682	152095	171759	193965	219042	247362
Benin	Atlantique	Abomey-Calavi	Zinvie	18157	13212	0,029	18677	19213	19764	20330	20913	21512	22129	22763	23416	24087

Sources : Calculs de l'auteur à partir des décomptes de population issus du RGPH3 et RGPH4 (INSTAD)

4. Répartition de la population à partir de la méthode d'interpolation surfacique entre 2017 et 2023 dans les shapes

Afin de répartir la population dans les shapes, nous appliquons la méthode d'interpolation surfacique, en partant des 92 arrondissements répartis sur les 13 communes de la zone du Grand Nokoué. Pour valider les résultats au niveau des arrondissements, nous effectuons une nouvelle désagrégation à partir du niveau administratif plus large des communes (c'est le niveau administratif 2 au Bénin). Cette étape permet d'agréger les estimations de population des shapes au niveau des arrondissements, puis de calculer le pourcentage d'erreur relative pour chaque arrondissement. Ce processus est répété pour les années 2017 à 2023.

Afin d'analyser la distribution des erreurs relatives dans les arrondissements, nous décidons de classer de nouveau ces erreurs en cinq catégories, comme défini précédemment dans le chapitre 3, sur Madagascar :

- Grande surestimation : erreurs inférieures à -50 %
- Surestimation : erreurs entre -50 % et -20 %
- Estimation correcte : erreurs entre -20 % et +20 %
- Sous-estimation : erreurs entre +20 % et +50 %
- Forte sous-estimation : erreurs supérieures à +50 %

En 2017, année la plus proche du recensement de 2013 et nécessitant déjà quatre années de projections démographiques, seulement 43 % des arrondissements présentent des erreurs inférieures à 20 % en valeur absolue et sont donc considérés comme correctement estimés. Par ailleurs, 12 % des arrondissements sont fortement sous-estimés et 10 % fortement surestimés. Le tableau 21 compare la distribution des erreurs pour l'année 2017 au Bénin avec les résultats obtenus à Madagascar en 2018, année du recensement. Une différence notable apparaît dans ces trois catégories, Madagascar comptant une proportion plus élevée de communes correctement estimées (51 %) et moins d'erreurs importantes.

Tableau 21 : Comparaison de la distribution de l'erreur entre le Bénin et Madagascar

	Bénin (2017)		Madagascar (2018)	
Catégorie d'erreur	Nombre d'arrondissements (n)	Pourcentage (%)	Nombre de communes (n)	Pourcentage (%)
Moins de -50 %	11	12,0	33	7.7
de -50 à -20 %	13	14,1	72	16.7
de -20 à 20 %	40	43,5	220	51.2
de 20 à 50 %	19	20,7	90	20.9
Plus de 50 %	9	9,8	15	3.5

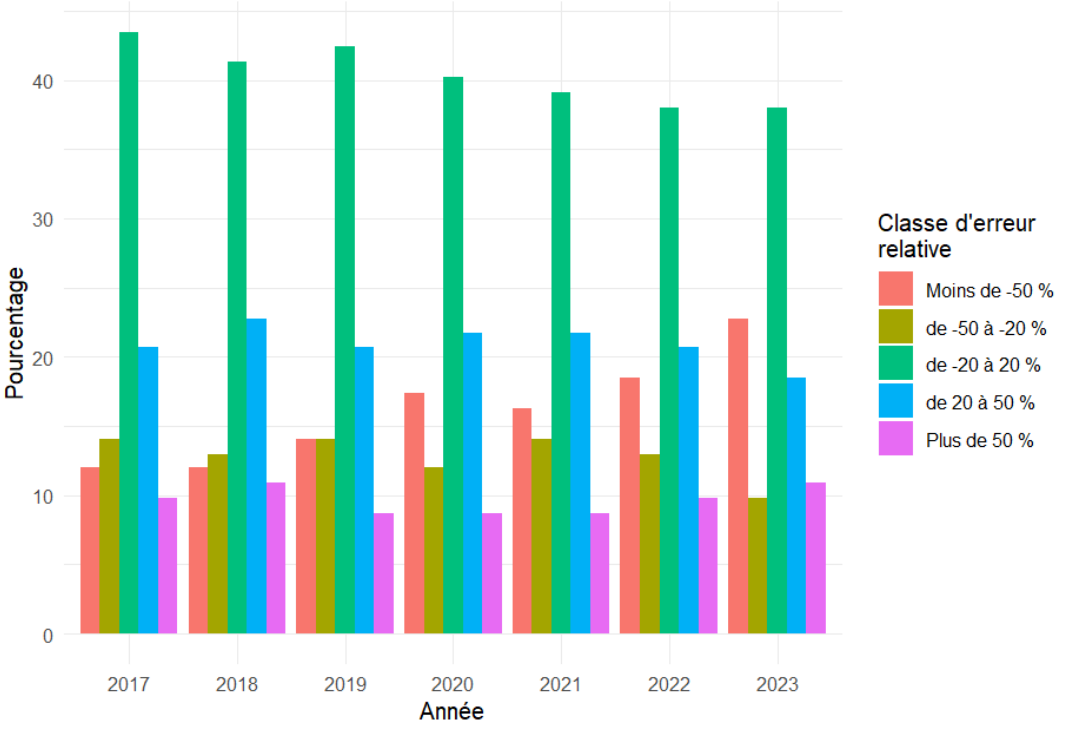
Sources : Calculs de l'auteur à partir des données de population de l'INSTAD et de l'INSTAT ainsi que des détections de bâti de TeleCense.

Un premier constat est que, dès la première année d'estimation au Bénin, les résultats sont moins précis qu'à Madagascar. Il aurait été plus pertinent de comparer pour l'année des recensements au Bénin et à Madagascar (c'est-à-dire 2013 pour le Bénin et 2018 pour Madagascar) mais comme les images Sentinel ne datent que de 2016, il n'est pas possible de remonter jusqu'en 2013 pour évaluer les différences dans la désagrégation. Ainsi, cette différence s'explique en partie par la nécessité de projeter la population sur quatre ans mais il est également possible que l'interpolation surfacique soit moins adaptée à la zone du Grand Nokoué, influençant ainsi la qualité des estimations.

L'évolution des erreurs dans le temps (figure 45) montre que les catégories d'estimations incorrectes ne progressent pas de manière spectaculaire en effectif, mais une tendance claire se dégage : le nombre d'arrondissements correctement estimés diminue progressivement (hormis pour 2019), tandis que les erreurs extrêmes deviennent plus fréquentes. Entre 2017 et 2023, la proportion d'arrondissements correctement estimés passe de 43,5 % à 38 %, tandis qu'en 2023, 11 % et 23 % des arrondissements sont fortement sous- ou surestimés, soit 32 des 92 arrondissements, marquant une augmentation significative depuis 2017.

Si l'évolution des catégories reste modérée, l'augmentation de l'erreur moyenne est en revanche bien plus marquée. En 2017, l'erreur relative moyenne en valeur absolue était de 36,2 %, atteignant 48,1 % en 2023, avec une hausse quasi linéaire de 2 à 3 points par an. Cette augmentation était attendue, notamment en raison du modèle de projection géométrique utilisé, qui tend à accumuler les erreurs d'une année sur l'autre. Par ailleurs, l'évolution du bâti, qui, comme nous l'avons vu dans la section 2 de ce chapitre, ne suit pas une progression strictement linéaire, constitue un facteur supplémentaire d'incertitude dans la désagrégation spatiale.

Figure 45 : Pourcentage d'arrondissements appartenant à chaque catégorie d'estimation selon les années



Sources : Calculs de l'auteur à partir des données de population de l'INSTAD et de l'INSTAT ainsi que des détections de bâti de TeleCense.

Il apparaît que l'utilisation de la désagrégation sur le long terme – en devant projeter la population sur plusieurs années – présente des limites importantes. Bien que cette méthode puisse être pertinente pour l'année du recensement, son efficacité décroît rapidement lorsqu'il s'agit de projections à plus long terme, comme le montre l'erreur relative moyenne de 36 % observée dans le cas du Bénin en 2017. Cette observation souligne clairement les limites de la désagrégation seule pour suivre l'évolution démographique dans le temps. Ainsi, il semble nécessaire d'explorer d'autres approches pour estimer la population dans des zones géographiques sans données de recensement récentes ou fiables.

5. Proposition de projections démographiques en prenant en compte l'évolution de la surface bâtie

Dans cette dernière partie, nous proposons une autre approche de projection de population basée sur l'évolution de la surface bâtie. Au vu de l'évolution incohérente de la surface bâtie au fils des années vue en section 2 et 2.1 et notamment

par la figure 41, nous n'avons pas pu mettre en place cette démarche. Elle reste néanmoins intéressante afin de répondre à certaines limites de notre démarche initiale et elle pourra dans le futur être mobilisée avec des programmes tels que Google 2.5D qui permettent une détection des zones bâties de manière continue de 2016 à 2023.

En effet, la démarche de projection de population retenue repose sur l'hypothèse que le taux d'accroissement démographique reste constant sur la période projetée, ce qui constitue une hypothèse forte et sans doute difficile à maintenir partout. Prenons l'exemple de l'arrondissement d'Abomey-Calavi, situé à proximité de Cotonou et déjà considéré en 2013 comme une cité-dortoir. Est-il raisonnable de penser que le taux de variation de la population de 6 % par an restera constant pendant les dix prochaines années ? Ce taux va sans doute évoluer et même plutôt décroitre en raison de la saturation progressive des bâtis de la zone, un phénomène similaire à celui observé à Cotonou, sans pour autant atteindre un point de saturation totale. En revanche, certains autres arrondissements, comme ceux des communes de Tori-Bossito ou de Ouidah, devraient voir leur taux de croissance augmenter avec les années. Ces zones, dont les taux de croissance connus sont relativement faibles, pourraient connaître une dynamique démographique plus marquée à mesure que le développement urbain et les infrastructures s'étendent.

Sans avoir les facteurs démographiques de natalité, mortalité et migration, nous souhaitons tout de même pouvoir faire évoluer le taux de croissance démographique. L'idée pour cela est d'y intégrer l'évolution de la surface bâtie dans l'ajustement du taux de croissance démographique. Pour ce faire, nous devons d'abord calculer un taux de croissance de la surface bâtie pour chaque année, en utilisant les données disponibles entre 2017 et 2023. Cela nous permettra de connaître l'évolution de la surface construite au fil du temps.

Le taux de croissance annuel de la surface bâtie est calculé pour chaque année comme suit :

$$r_{bâti} = (\frac{S_{t+1} - S_t}{S_t})$$

où :

- S_t est la surface bâtie à l'année t,
- S_{t+1} est la surface bâtie à l'année t+1.

Cette formule permet de calculer le taux de variation annuel de la surface bâtie entre deux années consécutives. Le taux de croissance du bâti $r_{bâti}$ peut être calculé pour chaque année de la période entre 2014 et 2023, en utilisant les valeurs de surface bâtie disponibles pour chaque arrondissement.

Une fois le taux de croissance du bâti calculé pour chaque année, ce taux peut être utilisé pour ajuster le taux de croissance démographique. L'idée est que les zones avec une forte croissance de la surface bâtie peuvent connaître une croissance démographique plus rapide.

Dès lors, le taux de croissance ajusté de la population $r_{ajusté}$ pour chaque année est calculé comme suit :

$$r_{ajusté} = r + \propto * r_{bâti}$$

Où :

- $r_{ajusté}$ serait le taux de croissance ajusté de la population,
- r est le taux de croissance démographique calculé sur la période 2002-2013,
- $r_{bâti}$ est le taux de croissance annuel de la surface bâtie calculé précédemment,
- $\propto$ est un paramètre de pondération qui ajuste l'importance du taux de croissance du bâti dans l'ajustement du taux démographique.

Pour finir, le taux de croissance ajusté $r_{ajusté}$ est utilisé pour projeter la population dans les années suivantes. La formule pour calculer la population projetée P_t est la suivante :

$$P_t = P_0 * (1 + r_{ajusté})^t$$

Cette méthode permettrait ainsi d'ajuster les projections démographiques en tenant compte de l'évolution réelle du bâti, offrant ainsi une estimation plus réaliste de la croissance de la population.

6. Conclusion

Dans ce chapitre, nous avons exploré l'application de la désagrégation à la population du Bénin, en nous concentrant particulièrement sur le sud du pays, la région du Grand Nokoué, qui est fortement urbanisée et densément peuplée. Cette zone, caractérisée par une urbanisation rapide et une saturation des espaces bâtis à certains endroits, comme à Cotonou, offre un terrain d'étude intéressant pour observer l'évolution de la population et des zones bâties au fil des années.

Afin de mettre en œuvre la désagrégation, nous avons dû préparer les zones sources et cibles du Bénin. Pour les zones sources, c'est-à-dire les arrondissements du Bénin, la tâche s'est avérée beaucoup moins complexe qu'à Madagascar, et il a seulement ici fallu résoudre des problèmes de mise à jour des noms des zones administratives. Toutefois, ces travaux ont de nouveau mis en lumière les difficultés rencontrées lors de la récupération des zones administratives en accès libre. Dans le cas du Bénin, les zones administratives à différents niveaux souffrent d'un manque de clarté dans leur délimitation. En effet, les frontières administratives du pays, des départements, des communes ou des arrondissements ne se superposent pas toujours parfaitement, ce qui peut poser des problèmes, non seulement du point de vue de la visualisation, mais également, et surtout, dans le suivi temporel des données spatiales. Ce chapitre souligne donc une nouvelle fois l'importance de disposer de données spatiales homogénéisées et en accès libre, afin de permettre des analyses spatiales cohérentes et fiables.

En ce qui concerne les zones cibles, c’est-à-dire les shapes identifiées par le programme TeleCense, la situation s’est avérée plus complexe en raison de certaines approximations et incohérences découvertes au fil de ce travail. Au Bénin, bien que le nombre de zones bâties ait montré une augmentation linéaire, l’augmentation de la surface bâtie n’a pas suivi cette tendance. Ce phénomène s’explique principalement par des périodes où la base de données OSM était incomplète ou de qualité insuffisante, en particulier en ce qui concerne le réseau routier, ce qui a conduit à des estimations erronées, incluant des zones non bâties à l’intérieur de certaines grandes shapes. Bien que ce problème n’ait pas pu être résolu immédiatement, il était essentiel de le mettre en évidence, non seulement pour la suite des travaux de thèse, mais également pour le programme TeleCense.

Ensuite, il a été nécessaire de projeter la population dans le temps afin de pouvoir procéder à la désagrégation de la population sur la période de 2017 à 2023. Faute de données démographiques classiques (natalité, mortalité, migration)

permettant des projections par composantes, nous avons opté pour la méthode de la croissance géométrique à taux fixe, en utilisant les taux de croissance de la population observés entre les recensements de 2002 et 2013 pour les arrondissements. Cette approche comporte certains défis, notamment l'accumulation progressive des erreurs au fil du temps, ce qui entraîne une diminution de la précision des projections à mesure que l'horizon temporel s'étend. Toutefois, en l'absence de données plus détaillées ou de solutions plus adaptées, elle reste une méthode efficace et pragmatique pour les premières années. Néanmoins, ces projections pourraient être améliorées, par exemple, en prenant en compte l'évolution des surfaces bâties au fil des années. L'émergence d'images satellitaires de plus en plus précises pourrait constituer un atout majeur pour affiner ces projections, en permettant de suivre l'évolution exacte de l'urbanisation et du bâti, année après année.

Enfin, il semble difficile de retenir la méthode de désagrégation comme méthode principale d'estimation de la population dès lors que l'on s'éloigne des données de recensement. Comme nous l'avons observé avec le cas du Bénin, après seulement quatre années de projection, l'erreur relative moyenne atteint déjà 36 %, et, avec le temps, cette erreur ne peut et ne fait que croître. Ces résultats soulignent la nécessité d'explorer d'autres approches pour estimer la population dans les zones géographiques où les données de recensement sont absentes ou trop anciennes. C'est dans cette perspective que nous développerons, dans les chapitres suivants, un modèle ascendant, visant à estimer la population dans ces zones, et que nous testerons notamment à Mayotte et dans la sous-préfecture d'Abidjan.

PARTIE 3 : Estimation de population par méthode ascendante

Chapitre 5 – Développement du modèle ascendant à partir des données de bâti TeleCense et évaluation à Mayotte puis à Abidjan

La troisième et dernière partie de cette thèse explore la possibilité de développer un modèle capable d'estimer la population sans s'appuyer sur des données démographiques préexistantes. Nous passons ainsi d'un modèle descendant, qui repose sur la répartition dans les shapes de la population des plus petites zones administratives disponibles, à un modèle directement fondé sur les caractéristiques des shapes elles-mêmes. Ce modèle, qualifié d'ascendant, s'appuie sur l'identification des zones bâties pour estimer les effectifs au niveau des shapes et remonte ensuite vers les zones administratives, qui servent de référence pour valider les résultats en l'absence de données démographiques précises à l'échelle des shapes.

Quel que soit l'approche – descendante ou ascendante –, l'objectif demeure le même : estimer la population des zones bâties déterminées à partir d'images satellites. Dans ce chapitre, les zones bâties utilisées sont les shapes identifiées par TeleCense. Toutefois, lors de la mise en place du modèle de désagrégation, nous disposions des données de population issues du dernier recensement national de Madagascar. Or, aucun jeu de données officiel ne fournit la population directement au niveau des zones bâties ou des shapes, contrairement aux données généralement disponibles au niveau des zones de dénombrement ou des unités administratives lors des recensements de population. Ainsi, en l'absence de données de référence à l'échelle des shapes, il n'est pas possible d'entraîner un modèle prédictif de la population des shapes, comme nous l'avions fait dans le chapitre 3 pour la modélisation de la densité de population des communes de Madagascar.

Dès lors, nous proposons d'utiliser les résultats obtenus avec la méthode descendante afin de développer la méthode ascendante. Concrètement, la population estimée dans les shapes à partir de la désagrégation des données du recensement de 2018 sert de point de référence. Cela nous permet d'entraîner un modèle qui pourra estimer la population directement à partir des caractéristiques des shapes, sans dépendre directement des données administratives existantes.

Ce chapitre se décline en trois sections principales. La première concerne la création du modèle ascendant, en nous appuyant sur les résultats de la désagrégation à Madagascar. Ensuite, nous testons ce modèle dans deux contextes distincts : tout d'abord, le département de Mayotte, géographiquement proche de Madagascar et qui

présente des caractéristiques similaires en termes de diversité de bâti et d'environnement. Le second, bien plus éloigné en distance et en type de bâti, est la sous-préfecture d'Abidjan, l'une des régions les plus densément peuplées d'Afrique (ONU-Habitat, 2023). Abidjan, qui se distingue par sa forte urbanisation et ses bâtiments à plusieurs étages, soulève la question des limites d'un modèle ascendant créé dans un contexte différent.

1. Création du modèle ascendant à Madagascar

Pour développer un modèle capable d'estimer la population des shapes, nous utilisons comme référence la population désagrégée dans les six régions de Madagascar. Bien que nous n'ayons pas pu évaluer directement la précision de la répartition de la population à l'échelle des shapes, les résultats obtenus dans le chapitre 3 apportent des garanties. En effet, lorsque nous avons évalué l'erreur relative des communes à partir de la désagrégation par interpolation surfacique réalisée au niveau des districts, les prédictions se sont révélées satisfaisantes avec plus de 50 % des communes qui présentent une erreur inférieure à 20 %. Ces éléments nous permettent de considérer que ces données désagrégées peuvent être utilisées comme population de référence pour la construction du modèle ascendant.

L'intérêt principal de cette approche est qu'elle constitue un moyen unique de disposer d'un vaste ensemble de données, soit près de 300 000 individus statistiques (les shapes des six régions de Madagascar), sur lequel entraîner le modèle. Ces régions choisies pour leurs caractéristiques environnementales et de bâti différentes, suggèrent qu'en utilisant leurs shapes nous pourrions avoir un modèle généralisable que nous pourrons exporter pour estimer la population d'autres contextes géographiques.

Afin de mettre en œuvre l'approche ascendante, nous appliquons les mêmes méthodes que celles présentées dans le chapitre 2, à savoir la modélisation linéaire et les forêts aléatoires avec validation croisée. Pour cela, nous constituons deux échantillons aléatoires à partir des 430 communes, en réservant les shapes de 70 % des communes pour l'entraînement du modèle et celles des 30 % restantes pour la validation. Après la création de ces deux échantillons, nous obtenons pour l'entrainement 206 480 shapes réparties dans 298 communes et 89 886 shapes réparties dans 132 communes pour la validation.

La variable à estimer est donc la population désagrégée à partir des communes de Madagascar dans les shapes, dont la distribution a été brièvement présentée dans

la section 5.3 du chapitre 3. Nous y avions observé que la majorité des shapes sont de petite taille — 85 % d'entre elles couvrent une surface inférieure à 1 000 m². Cependant, la présence de shapes de très grande superficie, associées à des populations élevées, impose un ajustement de la variable à travers une transformation logarithmique.

Les variables testées dans les modélisations sont les variables présentées dans le tableau 12 du chapitre 3, à savoir les variables physiques propre à la shape (la surface, la densité de bâti, la proportion de bâti communale) les variables climatiques et environnementales (précipitation, température, altitude) et les variables liées à une distance à une entité proche de la zone bâtie (route, santé, grande ville) et la variable d'intensité de la lumière la nuit.

1.1 Modèle ascendant par modélisation linéaire

Les variables explicatives ont été incorporées une à une afin de tester leur effet sur la prédiction de la population logarithmique des shapes. Hormis la variable de distance au point de santé le plus proche, toutes se sont révélées significatives et ont un impact sur la variable dépendante. Le modèle retenu est le suivant :

$$\begin{aligned} Y_i = \ & \beta_0 + \beta_1 \log(surface) + \beta_2 altitude + \beta_3 proportion_{bâti} \\ & + \ \beta_4 \log(distanceToCity) + \beta_5 \log(distanceToRoad) \\ & + \beta_6 precipitations \ + \beta_7 temperature + \ \beta_8 \text{densité}_{\text{bâti}} \\ & + \beta_9 \log(nightlights) + \varepsilon_i \end{aligned}$$

Avec Y_i le logarithme de la population de la shape i et ε_i un terme résiduel aléatoire supposé indépendant et distribué selon une distribution $N(0, \sigma^2)$. Le tableau 22 présente les résultats de la modélisation linéaire, notamment les coefficients des paramètres estimés, leur significativité statistique ainsi que le coefficient de détermination (R^2) obtenu sur l'ensemble de test. Toutes les variables sont significatives, ce qui confirme leur pertinence dans l'explication de la population désagrégée. En ce qui concerne l'interprétation des coefficients des variables, comme prévu, dans un modèle fondé sur une désagrégation par interpolation surfacique, la surface joue un rôle crucial et est la variable la plus influente, démontrée par le coefficient standardisé le plus élevé (0,91). Cela veut dire que plus la surface d'une shape est grande, plus sa population estimée est élevée.

Le coefficient positif associé à la distance à la grande ville indique qu'une shape située à proximité d'une grande ville ou intégrée à son agglomération a, en moyenne, une population plus faible que les shapes plus éloignées. À l'inverse, les shapes situées à plus grande distance tendent à avoir une population plus élevée.

Pour rappel, la définition des grandes villes s'inspire directement du concept d'agglomération utilisé par Africapolis, où une agglomération est formée par des shapes situées à moins de 200 mètres les unes des autres. En ajoutant la contrainte selon laquelle une grande ville doit regrouper des shapes couvrant plus de 1 km², on s'assure qu'il s'agit bien d'un ensemble structuré reflétant un phénomène d'urbanisation.

Ainsi, l'effet observé sur cette variable est cohérent avec les données issues des recensements, qui montrent que la taille moyenne des ménages est généralement plus faible en milieu urbain qu'en milieu rural. Cette différence est en effet bien documentée en Afrique, notamment à Madagascar, où en moyenne, un ménage compte 3,9 habitants en milieu urbain contre 4,3 en milieu rural (INSTAT, 2021g). Cette tendance se retrouve dans les régions étudiées : 3,9 et 4,1 à Analamanga, 4,0 et 4,5 à Itasy, 3,6 et 4,1 à Atsinanana, 4,0 et 4,5 à Melaky, et 3,3 et 3,6 à Diana. Seule la région d'Androy présente une exception, avec une taille moyenne de 4,5 habitants par ménage, quelle que soit la nature du milieu. Bien qu'on ne puisse pas dire qu'une shape est équivalente à un ménage, l'effet global de cette variable semble logique au vu des différences entre milieu rural et urbain.

Si l'on suit l'effet de la variable précédente, une shape située très loin d'une grande ville aurait théoriquement une population de plus en plus grande, ce qui pourrait sembler contre-intuitif. Cependant, cet effet est en grande partie contrebalancé par l'influence d'autres variables moins importantes dans le modèle mais qui vont dans la même direction.

L'une de ces variables est la distance à la route la plus proche, qui a un coefficient négatif. Cela signifie qu'une shape éloignée d'une route a une population estimée légèrement inférieure à celle d'une shape proche d'une route. Par construction, cette variable reflète l'idée que les shapes non isolées, situées à proximité des routes ou des pistes, ont tendance à avoir une population plus élevée en raison d'un probable meilleur accès à des infrastructures.

De manière similaire, l'intensité de la lumière nocturne est un autre facteur positif dans le modèle. Ce coefficient positif indique qu'une shape ayant une intensité lumineuse élevée aura une population estimée légèrement plus importante. Enfin, la proportion de bâti communal a également un coefficient positif. Cela signifie que plus la surface bâtie dans une commune est importante, plus une shape située dans cette commune aura une population estimée élevée.

Les variables climatiques étudiées ont toutes un impact similaire sur la population estimée des shapes. En particulier, les deux variables les plus importantes sont l'altitude et la température, qui sont également fortement corrélées entre elles. Les coefficients négatifs associés à l'altitude et à la température indiquent que, plus

une shape est située à une altitude élevée ou dans une zone à température élevée, moins sa population estimée est importante. La variable de précipitations a également un coefficient négatif mais son effet est plus faible en comparaison avec l'altitude et la température.

Enfin, la variable de densité de bâti, avec un coefficient standardisé de 0,0071, a un impact plus que négligeable, notamment en raison de la méthode de construction de la population désagrégée dans le modèle descendant utilisé, uniquement fondé sur la surface et non la densité de bâti.

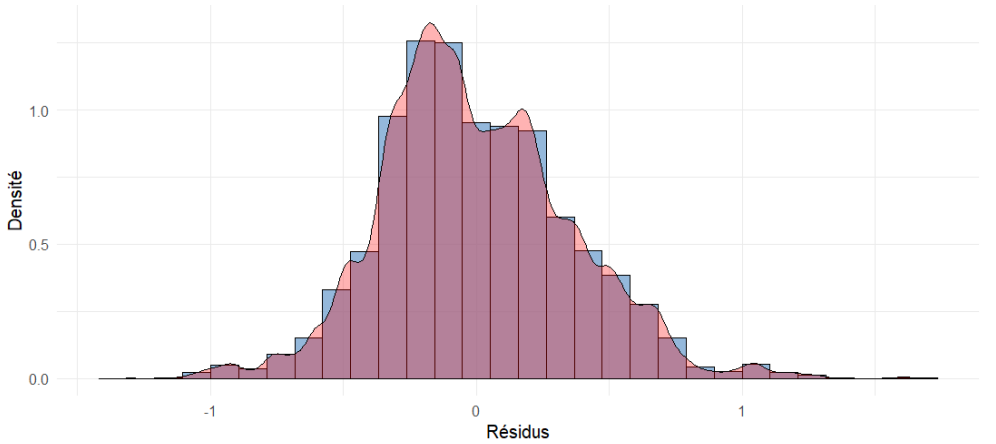
Tableau 22 : Coefficients associés à la modélisation linéaire de la population des shapes de Madagascar

Variable	Estimation du coefficient	Coefficient standardisé
Constante = Population désagrégée (au log)	- 0,059***	
Surface (en m² et au log)	0,87***	0,91
Altitude (en mètres)	-0,0008***	-0,43
Proportion de surface bâtie dans la commune (les shapes d'une même commune ont la même valeur)	0,024***	0,083
Distance à la grande ville la plus proche (en mètres et au log)	0,048***	0,18
Distance à la route la plus proche (en mètres et au log)	-0,0098***	-0,035
Précipitations (mm)	-0,00015***	-0,08
Température (°C)	-0,11***	-0,32
Densité de bâti	0,0048***	0,0071
Intensité de la lumière la nuit (au log)	0,032***	0,025
R^2 sur l'ensemble de test	0,896	

Sources : Calculs réalisés à partir des variables issues du programme TeleCense

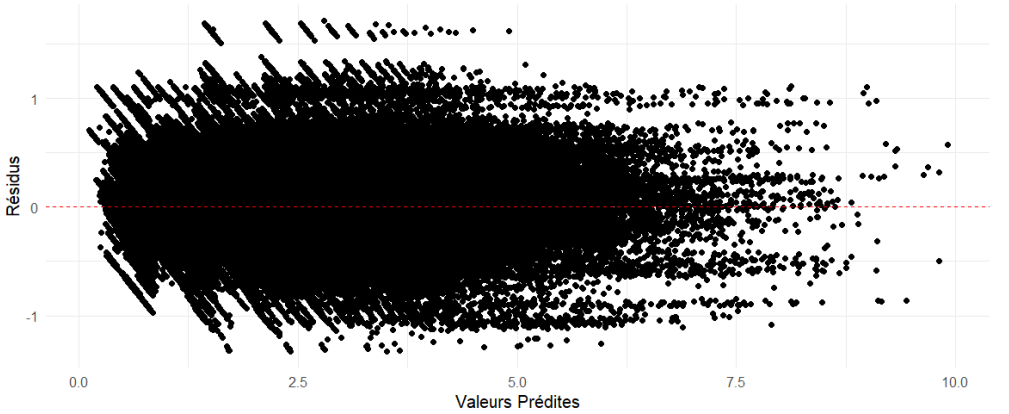
Le coefficient de détermination R^2 égal à 0,896 calculé sur l'ensemble de test indique que 89,6 % de la variance observée dans la population estimée peut être expliquée par le modèle, suggérant que le modèle fournit une estimation très précise des valeurs réelles. Le modèle semble donc performant pour estimer des populations proches de celles obtenues par désagrégation. Cette performance est confirmée par les deux figures 46 et 47 suivantes, qui montrent une distribution des résidus centrée autour de zéro, avec des écarts qui restent relativement faibles et réguliers à travers les différentes valeurs prédites.

Figure 46 : Histogramme des résidus du modèle ascendant



Sources : Calculs réalisés à partir des variables issues du programme TeleCense

Figure 47 : distribution des Résidus en fonction des valeurs prédites



Sources : Calculs réalisés à partir des variables issues du programme TeleCense

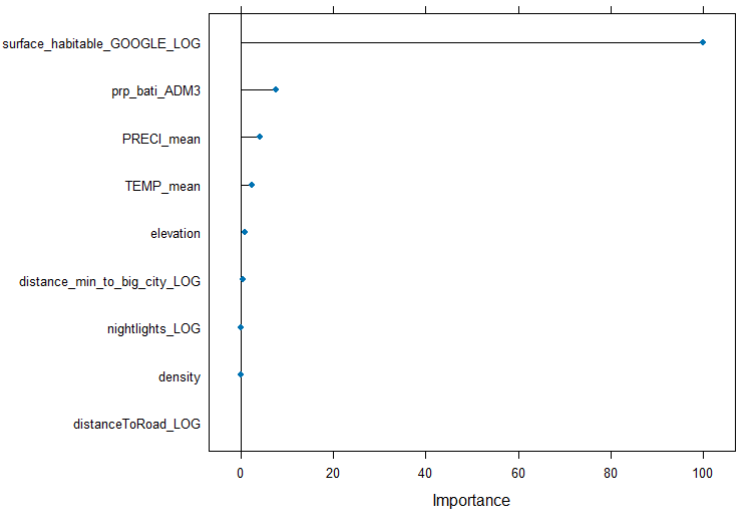
1.2 Modèle ascendant par forêt aléatoire

Nous développons ensuite le modèle ascendant en utilisant l'algorithme de forêt aléatoire. L'échantillon d'entraînement reste identique à celui utilisé pour la modélisation linéaire, soit 206 480 shapes réparties dans 298 communes. Comme dans le chapitre 3, l'ajustement du modèle a été réalisé à l'aide d'une validation croisée à cinq plis.

Les résultats d'importances des variables, illustrés par la figure 48, confirment que la variable de surface joue un rôle prépondérant dans l'estimation de la population. Elle apparaît comme la principale variable explicative, surpassant largement les autres facteurs du modèle. Contrairement à la régression linéaire, où

certaines variables complémentaires conservaient une influence notable, la forêt aléatoire accentue davantage la domination de la variable de surface, réduisant considérablement l'apport explicatif des autres. Seules les variables de proportion de bâti et les variables climatiques conservent une contribution modérée à l'estimation, les autres présentent une influence négligeable.

Figure 48 : Importance des variables de l'algorithme de forêt aléatoire pour le modèle ascendant



Sources : Calculs réalisés à partir des variables issues du programme TeleCense

1.3 Comparaison et discussion des deux modèles ascendants

Comme dans le chapitre 3, afin de comparer les performances du modèle, nous utilisons l'ensemble de test (89 886 shapes réparties dans 132 communes) pour prédire la population logarithmique. Les résultats présentés dans le tableau 23 montrent que le modèle de forêt aléatoire affiche un RMSE plus faible et un R^2 plus élevé que la régression linéaire, indiquant une meilleure capacité à estimer la population logarithmique.

Tableau 23 : Comparaison des performances des modèles ascendants pour la prédiction de la population

Modèle	Root Mean Square Error (RMSE)	R^2
Régression linéaire	0.38	0.90
Forêt aléatoire	0.28	0.94

Sources : Calculs réalisés à partir des variables issues du programme TeleCense

Les deux modèles s'avèrent performants pour estimer la population désagrégée dans les shapes de Madagascar en 2018. Toutefois, il est important de rappeler une

limite méthodologique intrinsèque à notre méthode globale : les valeurs modélisées ne correspondent effectivement pas à des chiffres officiels de population, mais aux estimations obtenues par interpolation surfacique dans l'approche descendante du chapitre 3. Le modèle, bien qu'efficace sur les données de test de Madagascar, repose sur des hypothèses concernant la relation entre la surface bâtie et la population, ce qui peut ne pas être entièrement transposable à d'autres zones aux contextes géographiques ou socio-économiques différents.

Notamment, bien que le modèle de forêt aléatoire présente de meilleures performances (RMSE plus faible et R^2 plus élevé), il accorde une importance encore plus marquée à la variable de surface bâtie que la régression linéaire et pourrait amener à une moins bonne estimation de la population dans d'autres zones géographiques. C'est pourquoi une validation sur d'autres régions, avec des contextes démographiques et géographiques variés, est nécessaire pour confirmer la robustesse et la fiabilité de ces modèles dans des environnements différents.

Il n'est pas nécessaire de valider davantage les modèles sur Madagascar car ce territoire constitue notre terrain d'expertise initial et ce n'est pas du tout notre objectif de savoir ou de démontrer qu'il est possible d'estimer correctement la population à Madagascar à partir de ce modèle ascendant, ce serait redondant. Notre objectif est bien d'évaluer si un modèle développé à partir des données de Madagascar peut être appliqué à d'autres pays ou régions, aux contextes environnementaux parfois très différents, et surtout dans des zones où la diversité du bâti peut varier de manière significative. C'est ce que nous mettons en œuvre dans les deux sections suivantes du chapitre 5, d'abord à Mayotte, puis en Côte d'Ivoire.

2. Application du modèle ascendant dans le département de Mayotte

Dans le cadre de cette thèse, la meilleure stratégie de validation aurait été d'élaborer et de réaliser un micro-recensement de nombreuses zones bâties afin de tester et valider directement le modèle créé. Cependant, cette approche a rapidement été abandonnée en raison de la complexité logistique et du coût élevé qu'elle impliquerait. Une alternative consistait à sélectionner plusieurs régions d'un même pays où un recensement récent aurait été réalisé, permettant ainsi d'évaluer la performance des modèles développés. L'idée principale était d'exploiter les données des zones de dénombrement, plus petite unité opérationnelle du recensement, dont l'effectif décompté dépasse rarement 1000 habitants. Toutefois, l'accès à ces données s'est révélé complexe en raison de leur caractère sensible, ces informations étant généralement protégées par les gouvernements nationaux.

Puis, durant l'année 2023-2024, la préfecture de Mayotte avec l'Insee à Mayotte, confrontée à des besoins croissants en matière d'identification rapide du bâti, notamment au vu de la forte évolution du bâti précaire (Insee, 2023) a sollicité l'entreprise Diginove. Depuis, le service TeleCense a été mis en place et les shapes du département de Mayotte ont été cartographiées sur plusieurs années. Dans le cadre de cette collaboration, l'Insee a fourni les décomptes de population issus du recensement de 2017 au niveau des îlots. Ces derniers correspondent à un niveau administratif fin (niveau 5 ou 6) et constituent, de fait, des unités comparables aux zones de dénombrement évoquées juste précédemment. Ces données au format unique permettent de mettre en œuvre l'évaluation du modèle ascendant à partir de données de population très précises.

Mayotte, localisée dans l'océan Indien et faisant parte de l'archipel des Comores (Figure 49), est administrativement située en France, elle-même divisée en 18 régions et 101 départements. Mayotte est justement devenue ce 101e département français en 2011, et est subdivisée en 17 communes, elles-mêmes composées de 72 villages et 811 îlots. Ces derniers peuvent être considérés comme le niveau administratif 5.

Figure 49 : Carte de situation du département de Mayotte

Crédits : Encyclopædia Universalis France

En 2017, d'après le dernier recensement réalisé à Mayotte, la population était de 256 518 habitants. Il s'agissait du dernier recensement exhaustif. Depuis 2021, Mayotte a adopté le même mode de recensement que le reste du territoire français, avec une enquête annuelle couvrant chaque année une portion différente de son territoire (Insee, 2023). D'ailleurs, à partir de fin 2025 ou début 2026, sa population ne sera plus comptabilisée séparément mais intégrée aux statistiques nationales (Bardy, 2023). Comme en métropole, l'enquête annuelle est organisée et contrôlée par l'Insee et contient deux phases légèrement différentes, qui permettent une collecte successive de toutes les communes sur une période de 5 ans. D'une part, on recense sur cinq ans de manière exhaustive toutes les communes de moins de 10 000 habitants. Pour cela, l'Insee recense 20 % de ces communes tous les ans et au bout de cinq ans, toutes seront recensées (INSEE, 2022). D'autre part, les communes de plus de 10 000 habitants voient chaque année 20 % des logements recensés. De plus, au vu de la diversité des types d'habitats à Mayotte, notamment ceux précaires, 20 % des logements en tôle qui sont recensés chaque année, soit la totalité au bout de 5 ans (INSEE, 2022).

Pour autant, à Mayotte – qui connait des tensions politiques importantes et une croissance démographique rapide, de l'ordre de 4 % (à titre de comparaison, Madagascar est autour de 3 %) – le recensement de la population fait l'objet de critiques. Ces interrogations ne sont pas récentes et avaient déjà été évoquées par des membres de l'Insee, dans un article de blog qui résumait le ressenti de certains élus et d'une partie de la population (Seguin, Granjon, Thibault, 2023). L'un des points soulevés concerne la cartographie censitaire, avec le message de certains élus ou de la population locale que : « il serait illusoire de croire que tous les logements puissent être repérés et donc enquêtés. Non seulement le rythme de construction est très élevé, mais les constructions nouvelles peuvent être localisées dans des espaces jusqu'ici inhabités et peu accessibles ». En plus de cela, depuis le passage du cyclone Chido fin décembre 2024, de nouvelles critiques ont émergé, notamment de la part de députés et de membres du gouvernement Bayrou instauré depuis le 13 décembre 2024. Celles-ci portent en particulier sur une possible sous-estimation de l'immigration clandestine[56] dans les chiffres du recensement.

Concernant cette dernière critique, l'Insee, par la voix de son directeur ou de la cheffe du département de la démographie, répond à travers différents articles de presse que de nombreux contrôles sont effectués et qu'il n'existe pas de raison scientifique de penser que la population serait fortement sous-évaluée. Ces précisions visent notamment à répondre aux inquiétudes d'une partie de la population en

[56] https://www.francetvinfo.fr/france/mayotte/mayotte-les-vrais-chiffres-du-recensement_7027652.html

affirmant que « les personnes en situation illégale sont bien prises en compte dans le recensement », puisque, en France, « le recensement ne s'intéresse pas à la situation, régulière ou non, des personnes ». Ainsi, chaque individu présent sur le territoire, quel que soit son statut ou son âge, est recensé (Alexandre, 2025; Bardy, 2025; Deseille, 2025).

Bien entendu, des incertitudes existent, et des marges d'erreur sont inhérentes à toute estimation démographique. Toutefois, ces marges ne sont pas aussi importantes que certaines affirmations suggérant que Mayotte compterait autour de 500 000 habitants. Plusieurs contrôles externes, notamment l'analyse de la consommation de riz, d'huile, d'électricité ou encore le nombre de cartes SIM en circulation, fournissent des indices solides qui, associés à la rigueur méthodologique de l'Insee, rendent peu probable une sous-estimation aussi massive de la population.

Puis, concernant les incertitudes liées au recensement des habitats, l'Insee revendique au contraire, ne pas avoir omis de logements au cours de la collecte, notamment grâce à la méthode mise en œuvre, qui a vu une quinzaine d'agents recenseurs locaux sillonner le département à la recherche de tout nouveau logement, de tout type de bâti (Insee, 2023).

Dans les faits, la question est effectivement légitime. Il est d'ailleurs probable, au vu de la rapidité de construction des logements, couplée à leur nombre important - quatre logements sur dix sont en tôle en 2017 (Thibault, 2019) – que l'Insee n'ai pas décompté tous les logements. Cependant, en tenant compte du nombre important d'agents recenseurs mobilisés et des multiples contrôles réalisés tout au long du processus, en plus des différents croisements effectués, il est aussi probable que les omissions soient limitées voire marginales.

Quelle est donc l'utilité d'un outil tel que TeleCense, ou plus généralement des recherches liées à l'imagerie satellitaire dans le cadre du recensement à Mayotte ? TeleCense, dans ce contexte, n'a pas pour vocation de mesurer les flux ou d'estimer directement la population. Son rôle est plutôt informatif : il permet de signaler presque en temps réel l'apparition de nouveaux bâtiments. Grâce aux images fournies par Sentinel-2, avec une fréquence moyenne de 5 à 6 images par mois, et en l'absence de couverture nuageuse excessive, il est possible de repérer ces nouvelles constructions. TeleCense pourrait ainsi contribuer à la préparation du travail de l'Insee et de ses agents recenseurs, sans pour autant se substituer à l'indispensable travail de terrain. Il permettrait plutôt de l'alléger et de le rendre plus efficace, notamment en aidant à ne pas oublier certains bâtiments. En conclusion, TeleCense ne vise pas à remplacer les agents sur le terrain, mais peut :

1. Faciliter le travail, en fournissant des cartes régulièrement mises à jour, indiquant les zones où de nouveaux bâtiments ont été détectés.
2. Offrir une validation supplémentaire, en comparant les informations recueillies par les agents recenseurs avec celles issues des algorithmes de détection de TeleCense.
3. Réduire le risque d'omission, en générant une carte finale avant la soumission du recensement, permettant ainsi de vérifier si de nouveaux bâtiments ont été construits après le passage des agents.

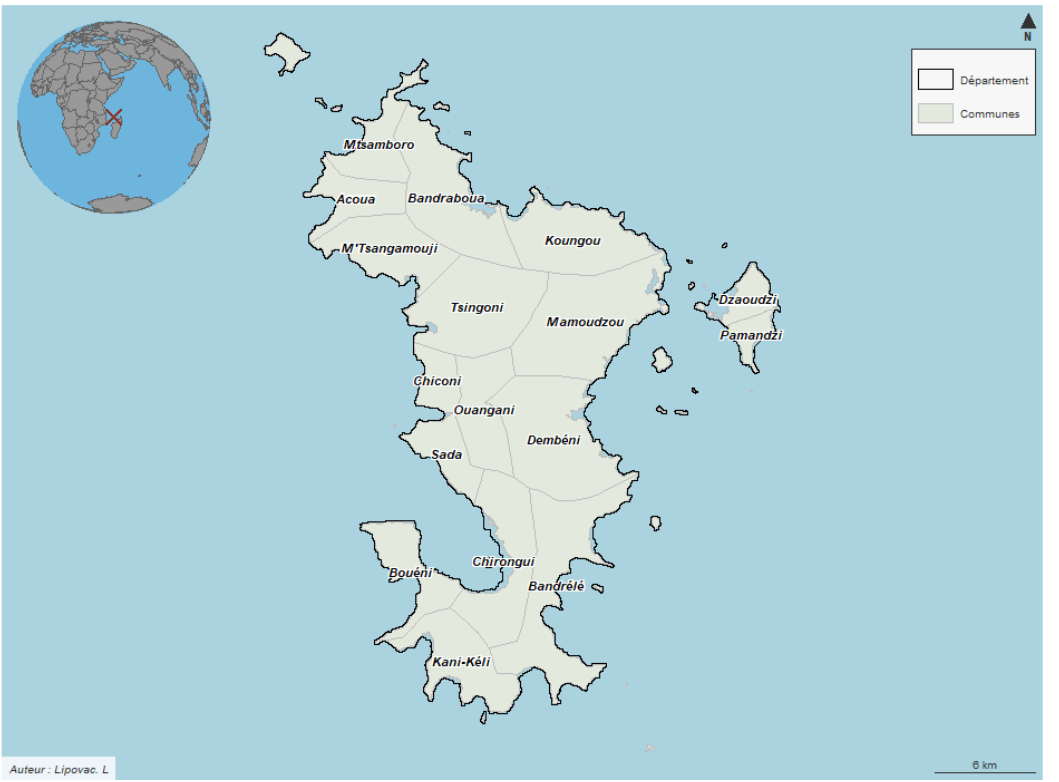
Quoi qu'il en soit, ce partenariat avec l'Insee nous permet d'accéder aux décomptes de population du recensement de 2017 ainsi qu'aux tracés des limites administratives correspondantes, les îlots. Ces données sont utilisées pour évaluer la performance du modèle ascendant à Mayotte en 2017.

2.1 La population et les zones administratives de Mayotte

Les recensements de population pour les années 2012[57] et 2017[58] sont disponibles à l'échelle des niveaux administratifs 3 et 4, correspondant aux 17 communes et 72 villages (figures 50 et 51) que compte Mayotte. Dans le cadre de ce projet, l'Insee à mis à notre disposition la population recensée en 2017, ainsi que le nombre de logements, dans les 811 ilots. Ces ilots, dont la répartition est visible figure 51, ne sont pas des niveaux administratifs officiels, mais plutôt des zones de dénombrement, définies pour les besoins du recensement. Leur petite taille facilite une collecte de données plus précise. Bien qu'en 2017, on décompte 256 518 habitants, en prenant les fichiers partagés et en sommant la population totale des 811 ilots, nous trouvons 254 926 habitants, ce qui sera donc notre marque de référence et de vérification pour la suite. Contrairement à Madagascar ou au Bénin, aucun travail de jointure n'est ici nécessaire dans le sens où les 811 ilots sont déjà liés à leur population.

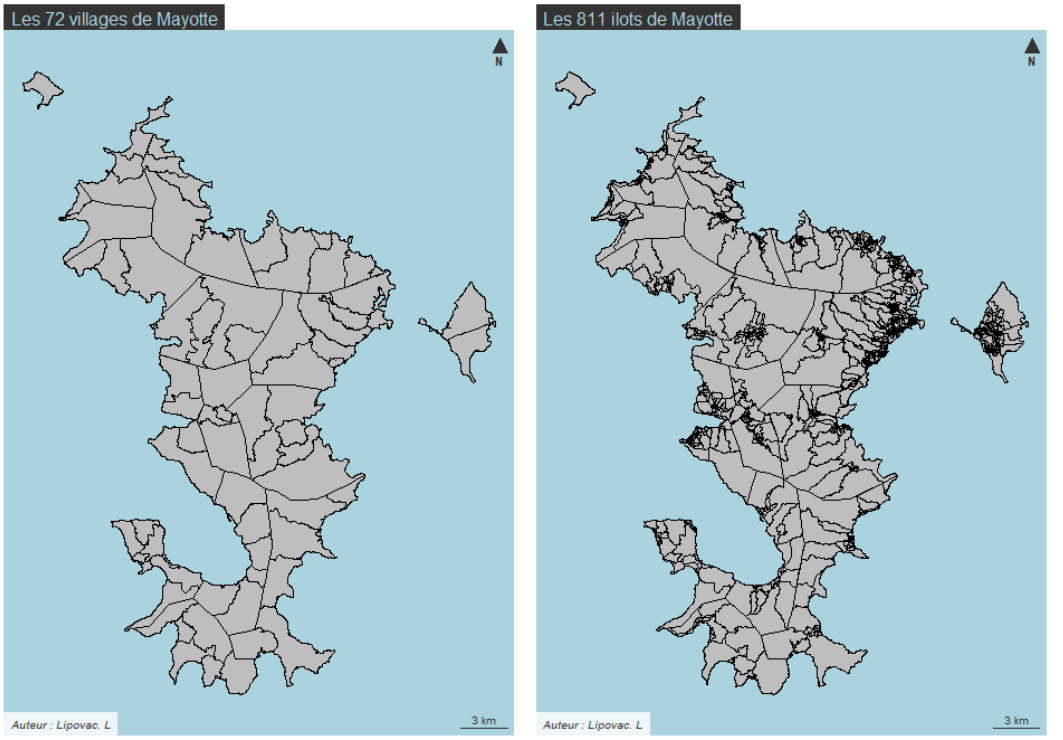
[57] https://www.insee.fr/fr/statistiques/2409395?sommaire=2409812
[58] https://www.insee.fr/fr/statistiques/4223807#documentation

Figure 50 : Les 17 communes de Mayotte



Sources : Carte réalisée par l'auteur à partir de la couche des communes fournie par l'Insee

Figure 51 : Répartition des villages et ilots de Mayotte



Sources : Carte réalisée par l'auteur à partir de la couche des villages et des ilots fournis par l'Insee

Comme le montre la figure 51 (image de droite), un îlot peut être très petit en termes de superficie, avec une superficie moyenne de 0,45 km², tandis que la médiane est encore plus réduite, à seulement 0,04 km². En comparaison, à Madagascar et au Bénin, les superficies moyennes d'une commune et d'un arrondissement (les plus petites unités administratives disponibles) sont respectivement de 292 km² et 27 km², avec des médianes de 151 km² et 21 km². Les unités administratives disponibles à Mayotte sont donc nettement plus petites que celles utilisées lors de la désagrégation à Madagascar ou au Bénin et offrent une perspective intéressante dans l'optique d'une vérification et validation du modèle ascendant.

La figure 52 illustre par ailleurs ce propos. On peut y voir la distribution de la population dans une partie de la commune de Mamoudzou : pour de nombreux ilots, le décompte de population ne dépasse pas 300 habitants. Cette ressource partagée par l'Insee est donc précieuse, dans le sens où c'est la première fois que nous avons à disposition des données aussi précises.

Figure 52 : Vue satellitaire de la répartition de la population dans les ilots – Commune de Mamoudzou – 2017

Sources : Image issue du logiciel QGIS à partir de la couche des ilots fournie par l'Insee

Comme ces décomptes de population datent de 2017, il est nécessaire d'identifier les shapes correspondant à cette même année afin d'estimer la population dans les zones bâties.

2.2 La cartographie des zones bâties en shapes

En 2017, TeleCense a cartographié 7 175 zones bâties à Mayotte. Comme les ilots sont des surfaces particulièrement petites, il y a de nombreuses shapes qui intersectent deux ou plusieurs ilots. Pour ces shapes, nous les divisons afin qu'elles n'appartiennent plus qu'a à un seul et même ilot.

Ensuite, nous calculons la surface modifiée à partir de l'identification des zones bâties de Google Open Buildings, ce qui produit un total de 8 385 Shapes réparties dans les 811 ilots du département de Mayotte.

2.3 Méthodes de validation employées

Les zones de dénombrement permettent d'évaluer le modèle ascendant selon deux approches distinctes. Dans les deux cas, nous commençons par estimer la population des shapes identifiées en 2017 en appliquant un des deux modèles ascendants. Pour ce faire, nous réutilisons les coefficients obtenus par la modélisation linéaire ou par la forêt aléatoire sur Madagascar et les appliquons aux shapes de Mayotte en tenant compte des valeurs spécifiques de leurs variables.

Ensuite, 1) pour la première méthode de vérification, nous agrégeons ces résultats au niveau des îlots, des villages et des communes, puis nous les comparons aux décomptes de population officiels. Cela permet d'évaluer les écarts avec les chiffres officiels pour chaque niveau administratif. Il s'agit du même type de vérification que celui réalisé dans les chapitres précédents.

2) Concernant la deuxième méthode, nous formulons l'hypothèse que, compte tenu de la faible superficie des îlots, une désagrégation réalisée à ce niveau devrait conduire à une répartition plus fidèle de la population. En effet, attribuer une population de 300 habitants à une zone restreinte avec peu de bâtiments est fondamentalement différent de la répartition de 10 000 à 50 000 habitants sur un grand nombre de shapes. Dans ce contexte, la probabilité d'une estimation proche de la réalité est plus élevée. Cette approche constitue donc notre deuxième méthode de validation, permettant d'évaluer l'écart entre les estimations issues du modèle et les données de référence obtenues par désagrégation par interpolation surfacique. De plus, compte tenu du type de bâti à Mayotte, qui est majoritairement horizontal et avec peu de bâtiments de forte hauteur de faible hauteur (Insee, 2023), l'interpolation surfacique devrait offrir des résultats particulièrement pertinents.

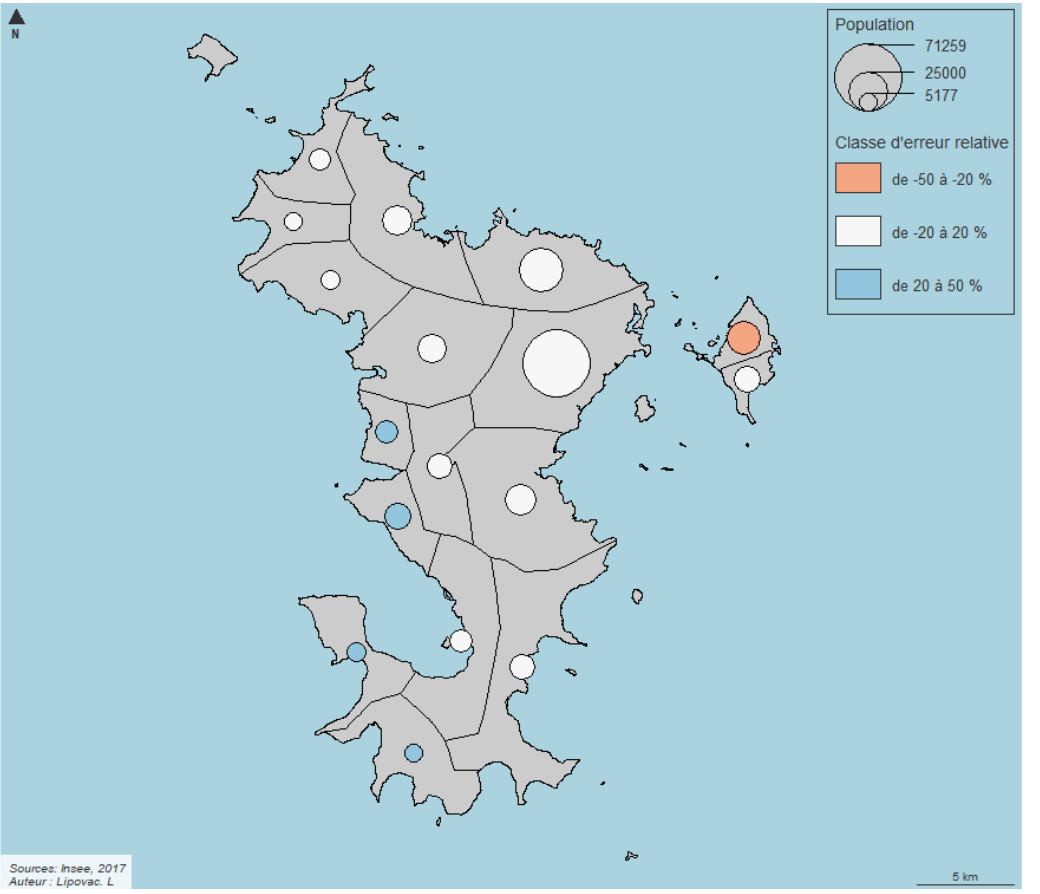
2.4 Résultats de l'approche ascendante à Mayotte

Dans cette partie, encore une fois c'est le modèle linéaire qui a donné les meilleurs résultats ; c'est donc celui-ci qui est retenu pour la suite des analyses. Il a permis d'estimer la population des 8 385 shapes identifiées à Mayotte en 2017.

2.4.1 Évaluation du modèle en agrégeant les estimations de population au niveau des zones administratives

Lorsque l'on agrège les estimations au niveau du département de Mayotte, on obtient un total de 246 470 habitants, soit une erreur d'estimation d'environ 3 %, ce qui constitue un résultat très satisfaisant. Ensuite, en agrégeant les estimations de population des shapes au niveau des 17 communes, les résultats restent relativement précis, dans la mesure où seules cinq d'entre elles présentent une légère sous- ou surestimation (plus de 20 % d'erreur relative en valeur absolue), comme illustré en figure 53.

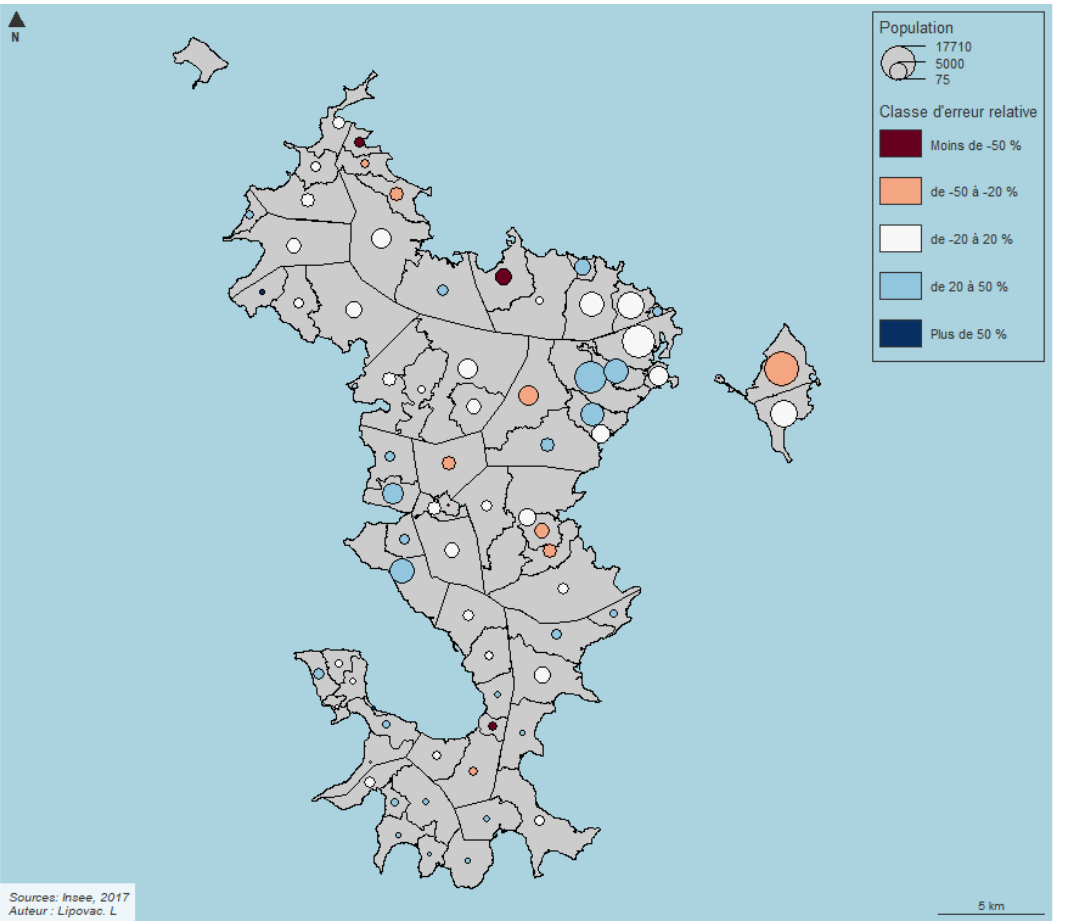
Figure 53 : Erreur relative lors de l'agrégation des estimations pour les 17 communes de Mayotte



Sources : Carte réalisée par l'auteur à partir de la couche des communes et des données de recensement de population fournis par l'Insee (2017)

Dès lors que l'on agrège les estimations au niveau des villages, la précision diminue légèrement. En effet, 43 % des villages sont considérés comme correctement estimés, tandis que 46 % présentent une légère sous- ou surestimation. Enfin, huit des 72 villages affichent une erreur plus marquée (villages en rouge et bleu foncé, figure 54).

Figure 54 : Erreur relative lors de l'agrégation des estimations pour les 72 villages de Mayotte



Sources : Carte réalisée par l'auteur à partir de la couche des villages et des données de recensement de population fournis par l'Insee (2017)

Enfin, l'évaluation des estimations au niveau des îlots révèle une précision encore plus limitée. En effet, seuls 243 des 811 îlots sont correctement estimés (environ 30 %). De plus, 28 % des îlots présentent une forte surestimation ou sous-estimation de leur population agrégée (plus de 50 % d'erreur). Il convient de souligner qu'à une échelle aussi fine, une erreur relative importante peut rapidement survenir. Par exemple, pour un îlot abritant 200 habitants, une estimation de 300 personnes représente déjà une erreur de 50 %.

Le tableau 24 récapitule ces résultats selon le niveau administratif d'agrégation. Comme entrevu avec la présentation des cartes, on comprend qu'aux niveaux plus grossiers correspondant aux communes et aux villages, la majorité des zones administratives semblent correctement estimées ; en témoigne la corrélation avec la population très forte, mais surtout les médianes relativement faibles (respectivement, la moitié des communes et la moitié des villages ont une erreur de moins de 17 et 22 %) .

Tableau 24 : Erreurs d'estimation selon le niveau administratif d'agrégation des résultats

Zone administrative	Corrélation avec décompte de population	Erreur moyenne	25 %	50 %	75 %	90 %
Communes	0,98	15,9	9,9	16,6	21,6	28
Villages	0,957	32,2	13,4	22,2	36,1	50,5
Ilots	0,493	40,3	16,9	33,3	52,4	74,3

Sources : Calculs de l'auteur à partir des données de recensement de population de l'Insee (2017) et des données issues du programme TeleCense

Concernant les ilots, 50 % d'entre eux sont estimés avec une erreur de moins de 33,3 %, ce qui correspond à une sous-estimation ou surestimation d'en moyenne 52 habitants (ou de médiane de 48 habitants). Cependant, pour la deuxième partie beaucoup moins bien estimée, nous avons une sous-estimation ou surestimation d'en moyenne 201 habitants (ou de médiane de 150 habitants).

Les différents niveaux d'agrégation et de vérification montrent que, à l'échelle nationale et communale, les erreurs semblent se compenser. Cependant, au niveau des ilots, il apparaît qu'un grand nombre de shapes pourraient être mal estimées. Étant donné le faible nombre de shapes par ilot (en médiane 7 shapes par ilot), ces erreurs peuvent rapidement entraîner une estimation incorrecte de la population de l'îlot. Il est donc essentiel d'examiner ces résultats de manière plus approfondie pour identifier les sources d'erreur et comprendre les raisons de ces écarts. C'est ce que nous faisons dans la partie suivante avec la deuxième méthode de validation qui va directement comparer les résultats de la désagrégation à partir des ilots avec ceux du modèle ascendant.

2.4.2 Évaluation du modèle en comparant les résultats avec ceux de la désagrégation

Désormais, nous comparons la population prédite par le modèle ascendant dans les shapes avec celle prédite par désagrégation par interpolation surfacique dans les shapes. Le tableau 25 montre que les estimations sont globalement assez proches. Cette observation est confirmée par la figure 55, qui illustre la relation entre ces deux ensembles de données avec une corrélation de 0,83. En général, le modèle ascendant

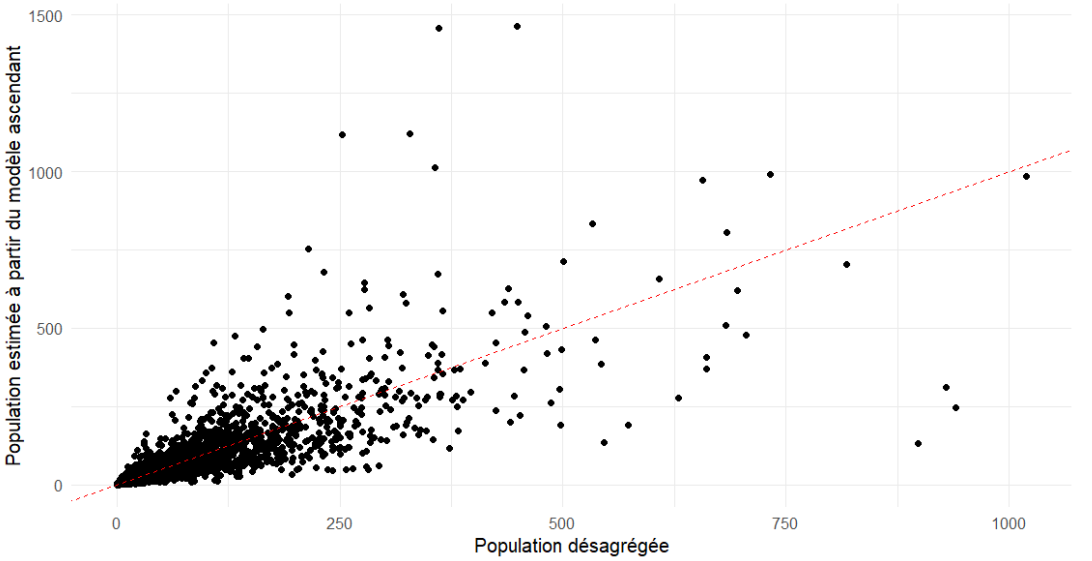
sous-estime légèrement les valeurs pour la première moitié des shapes (les plus petites en surface) et continue à fournir des estimations légèrement inférieures jusqu'à environ 95 % de l'échantillon. Cependant, pour les 2 à 3 % restants, correspondant aux shapes les plus grandes, le modèle ascendant tend à surévaluer la population, parfois de manière significative.

Tableau 25 : Comparaison des distributions de population entre l'estimation par désagrégation et par modèle ascendant

Méthode d'estimation	25 %	50 %	75 %	80 %	85 %	90 %	95 %	98 %	99 %	99,5 %	Max
Désagrégation	2	5	27	39	59	88	148	244	320	389	1020
Modèle ascendant	1.6	4.6	22.6	34	51	78	145	257	347	451	1463

Sources : Calculs de l'auteur à partir des données de recensement de population de l'Insee (2017) et des données issues du programme TeleCense

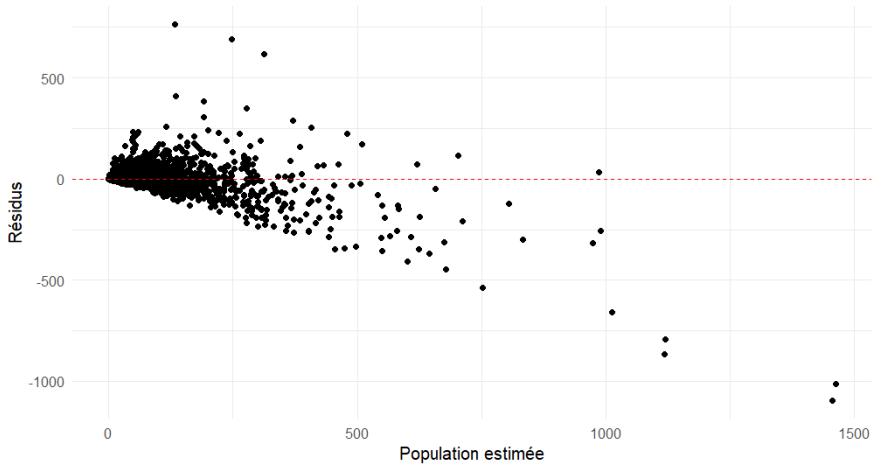
Figure 55 : Relation entre la population désagrégée et celle estimée par le modèle ascendant



Sources : Calculs de l'auteur à partir des données de recensement de population de l'Insee (2017) et des données issues du programme TeleCense

Ensuite, le graphique des résidus présenté en figure 56 vient appuyer ce constat : les shapes les plus peuplées ont davantage tendance à être surestimées que sous-estimées.

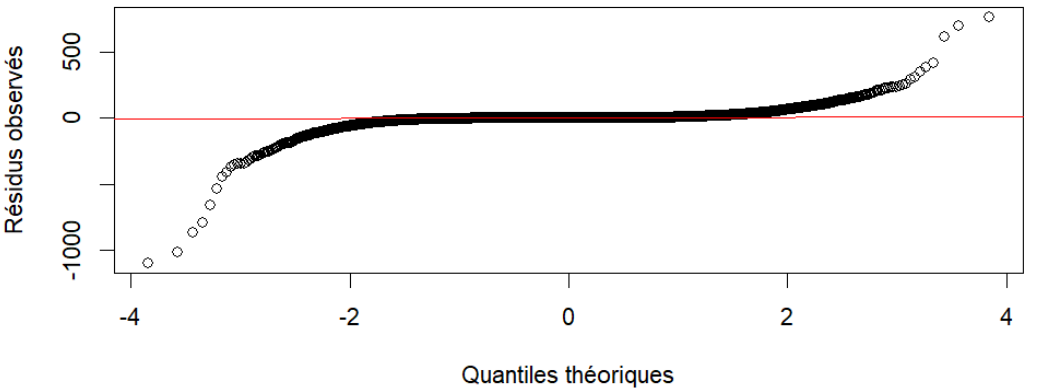
Figure 56 : Résidus entre la population désagrégée et celle estimée par le modèle ascendant



Sources : Calculs de l'auteur à partir des données issues du programme TeleCense

Nous concluons l'analyse des résidus en utilisant un graphique QQ-plot (Quantile-Quantile Plot), qui permet de comparer les résidus observés issus du modèle ascendant aux quantiles théoriques d'une distribution normale.

Figure 57 : Graphique QQ-plot



Sources : Calculs de l'auteur à partir des données issues du programme TeleCense

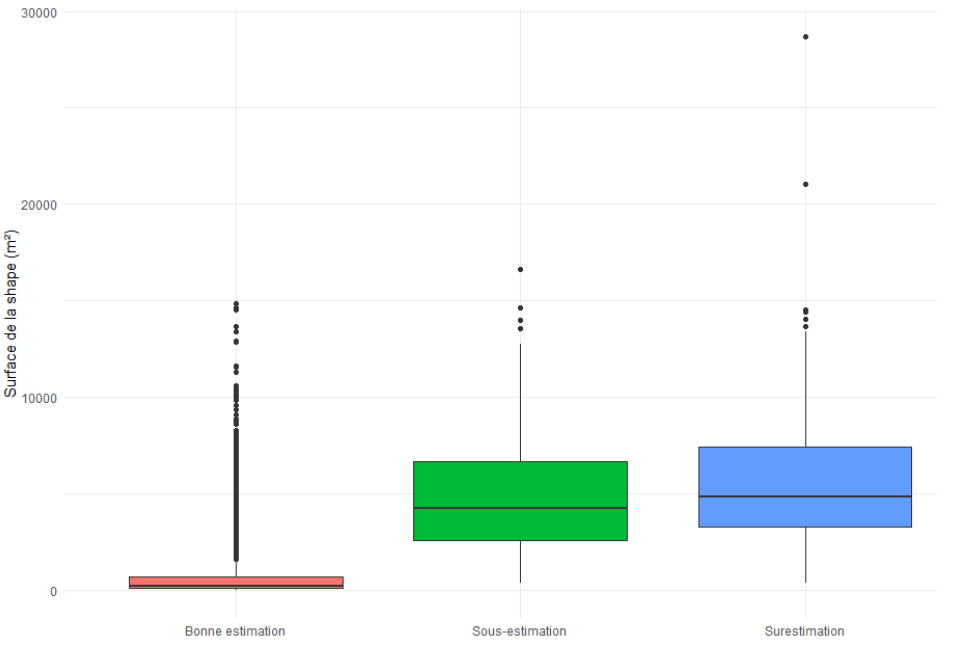
Si les résidus suivaient parfaitement une distribution normale, les points seraient alignés sur la ligne rouge. Bien qu'une grande majorité soit alignée, il y a une déviation importante dans les queues de distribution indiquant que les résidus extrêmes ne suivent pas une distribution normale et sont particulièrement surestimés (jusqu'à une différence maximale de 1000 habitants) ou sous-estimés (différence maximale de 500 habitants). Ce graphique indique clairement que malgré une réussite importante pour la majorité des shapes, le modèle ascendant a des difficultés à prédire correctement certaines valeurs extrêmes de la population désagrégée.

Afin de poursuivre l'analyse des résidus, et plus particulièrement des shapes fortement sous- et surestimées, nous définissons d'abord un seuil basé sur les résidus observés. Ainsi, les shapes dont les résidus observés dépassent l'intervalle [-2, 2] des quantiles théoriques de la figure 57 sont classées comme ayant des estimations extrêmes. Ces shapes aux estimations extrêmes sont au nombre de 191 des deux côtés (sous- et surestimées). Cela nous permet de les diviser en trois groupes :

- Les shapes correctement estimées : ce sont celles pour lesquelles les résidus observés sont proches de zéro, c'est-à-dire que la population estimée par le modèle ascendant est proche de la population attendue selon la distribution théorique, c'est-à-dire proche de la population attendue, celle obtenue par désagrégation par interpolation surfacique.
- Les shapes sous-estimées : ces shapes ont des résidus négatifs importants, ce qui signifie que la population estimée est inférieure à la population théorique attendue.
- Les shapes surestimées : ces shapes ont des résidus positifs importants, ce qui signifie que la population estimée est supérieure à la population théorique attendue.

À la recherche de motifs expliquant ces valeurs extrêmes, deux variables attirent notre attention et s'avèrent significatives. La première est la surface des shapes, comme le montre la figure 58 qui confirme que les shapes de grande surface présentent des écarts importants dans les estimations de population, tant en sous-estimation qu'en surestimation.

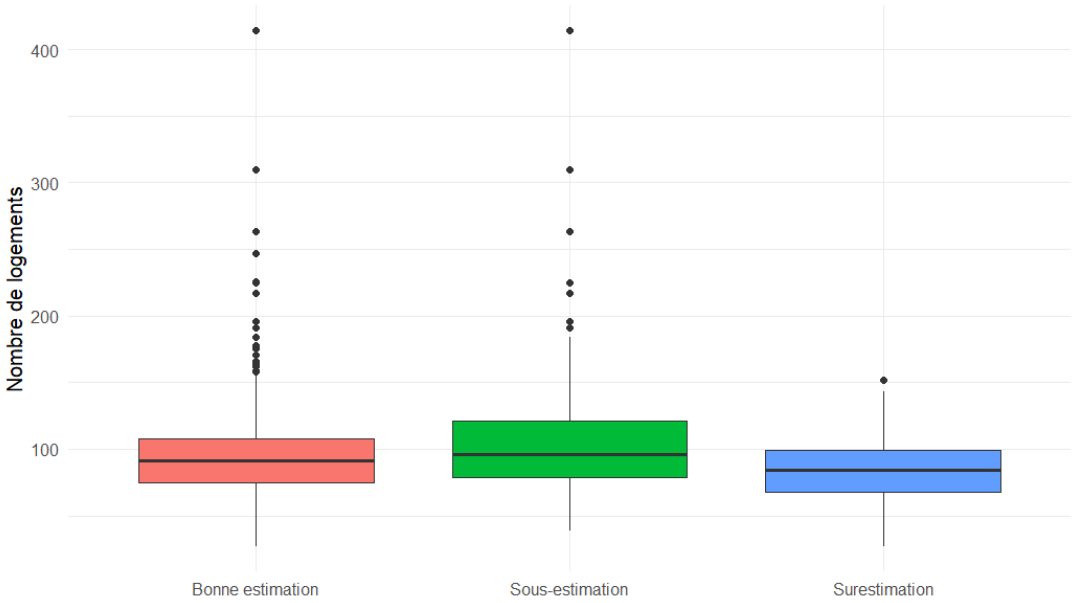
Figure 58 : Comparaison de la surface des shapes selon les trois catégories d'estimation



Sources : Calculs de l'auteur à partir des données issues du programme TeleCense

La seconde variable est le nombre de logements par îlot, issue de la base de données de l'Insee et spécifique à Mayotte. Bien que les différences ne semblent pas très marquées visuellement (figure 59), l'analyse statistique par ANOVA a révélé une différence significative dans le nombre de logements entre les trois groupes (shapes sous-estimées, correctement estimées et surestimées), avec une p-valeur de 1.54e-11. Ces résultats indiquent que la variation du nombre de logements entre les groupes est statistiquement significative et plus spécifiquement que les shapes sous-estimées sont associées à un nombre moyen plus élevé de bâtiments par îlot, tandis que les shapes surestimées sont associées à un nombre moyen plus faible de bâtiments par îlot.

Figure 59 : Comparaison du nombre de logements par ilots selon les trois catégories d'estimation



Sources : Calculs de l'auteur à partir des données de recensement de population de l'Insee (2017) et des données issues du programme TeleCense

Pour les shapes sous-estimées, il est probable que nous ayons eu des difficultés à détecter l'ensemble des bâtiments, ce qui a conduit à une sous-évaluation de la surface bâtie, et, par conséquent, de la population dans ces grandes zones. Cela est particulièrement lié à la structure de notre modèle, qui repose fortement sur la surface bâtie. À l'inverse, pour les shapes surestimées, les graphiques suggèrent que nous avons surévalué la surface bâtie, ce qui a entraîné une surestimation du nombre de logements et, en fin de compte, de la population dans ces zones.

2.4.3 Une estimation de la population prometteuse mais qui présente toujours de nombreuses limites

Mayotte présente de nombreuses similarités avec Madagascar, ce qui a probablement favorisé la performance de notre modèle. Sur le plan contextuel, nous avons établi notre modèle à Madagascar en 2018, et ici, nous estimons la population de Mayotte pour 2017, deux périodes très proches. D'autre part, en ce qui concerne le bâti, bien que Mayotte ait une plus grande proportion d'habitat précaire que Madagascar, ces structures restent largement composées de maisons individuelles et linéaires, sans grands immeubles à plusieurs étages, comme c'est le cas dans la majorité du département (INSEE, 2023). Ce type de bâti est donc similaire à celui que nous avons observé à Madagascar, où la majorité des constructions sont des maisons individuelles ou des concessions : dans les régions d'Atsinanana, Melaky, Androy et Diana, moins de 5 % des habitations sont des appartements ou maisons collectives ; le reste des habitations sont des maisons individuelles pour la majorité et quelques concessions, comme vu dans le chapitre sur Madagascar (INSTAT, 2021b).

Dans ce cadre, et pour un modèle ascendant entraîné sur des données désagrégées par interpolation surfacique, Mayotte apparaît clairement comme l'un des territoires les plus favorables pour tester cette approche. Et qu'observe-t-on ? Que le modèle fonctionne très bien à l'échelle départementale. Atteindre une marge d'erreur de seulement 3 % par rapport à la population totale, en n'utilisant qu'une dizaine de variables, est une véritable réussite. Le fait d'obtenir un nombre si proche de la réalité, en ne se basant que sur des caractéristiques liées aux shapes, démontre que l'approche globale développée dans cette thèse peut fonctionner. En effet, cela suggère que le programme TeleCense parvient à capturer correctement les zones bâties à cette échelle (ou même à l'échelle nationale, selon le contexte ; ici, nous nous concentrons sur un département) et que bien que des erreurs de sous- ou surestimation existent, elles semblent globalement se compenser à des échelles plus larges, ce qui constitue un autre point positif pour le modèle.

Effectivement, si l'on descend à des échelles plus détaillées, des erreurs apparaissent. C'est particulièrement notable à partir du niveau des îlots, où les écarts entre estimation et réalité deviennent plus marqués. Autrement dit, il existe nécessairement des erreurs d'estimation au niveau des shapes ; on l'a vu, environ 5 % des shapes étaient soit très fortement sous- ou surestimées, avec une plus grande marge d'erreur pour les shapes surestimées. Bien que ce ne soit qu'environ 5 % des zones bâties totales, ces shapes, en raison de leur surface significative, englobent une part importante de la population à estimer. Dans le cas précis de Mayotte, les erreurs se sont compensées, mais cela ne sera pas toujours le cas. Si ces compensations ne se

produisent pas, nous pourrions aboutir à des sous-estimations ou surestimations majeures, ce qui pose évidemment un problème pour la fiabilité du modèle.

Dès lors, quelle implication pour l'utilisation des estimations à un niveau local ? Il est nécessaire de comprendre la source de ces erreurs et leur impact sur l'utilisation locale des estimations. Après tout, si l'on sait déjà estimer la population départementale avec d'autres techniques, comme les projections démographiques, la véritable valeur ajoutée de notre approche repose sur sa capacité à fournir des estimations précises à des niveaux plus fins.

Nos résultats semblent confirmer ce que nous avions déjà observé au Bénin : les erreurs d'estimation concernent majoritairement les shapes de grande surface. Ces dernières posent un problème de caractérisation, car elles englobent un grand nombre de bâtiments possiblement hétérogènes, incluant potentiellement des structures non résidentielles (entrepôts, bâtiments administratifs, etc.). Plus une shape est grande, plus le risque d'erreur est élevé si cette hétérogénéité n'est pas correctement prise en compte. À l'inverse, les résultats de Mayotte confirment que la population des shapes de petite et moyenne taille est nettement plus facile à estimer, ces dernières étant plus homogènes et uniformes.

Ces résultats suggèrent plusieurs axes d'amélioration. Tout d'abord, au niveau du programme d'identification du bâti de TeleCense, il est essentiel d'améliorer la segmentation des zones bâties en séparant mieux les zones bâties entre elles, réduisant de fait leur surface totale. Le programme TeleCense devra intégrer des informations complémentaires, notamment sur le type de bâti et la hauteur des bâtiments, afin d'affiner les estimations de population. Pour l'instant, la variable de densité du bâti repose principalement sur une approche horizontale, ce qui est insuffisant pour capturer la complexité de certaines zones.

Également, il sera définitivement important d'incorporer davantage de variables, que ce soit dans l'approche descendante pour ne pas s'appuyer que sur la surface comme facteur de répartition de la population, ou dans l'approche ascendante afin de mieux rendre compte des variations locales.

Pour le moment, nous passons à l'évaluation du modèle ascendant dans la sous-préfecture d'Abidjan mais ces considérations seront discutées et approfondies en fin de ce chapitre et surtout dans le dernier chapitre de la thèse, où nous reprendrons notre approche globale avec une détection du bâti réalisée à partir d'images satellites plus précises.

3. Application du modèle ascendant dans la sous-préfecture d'Abidjan

Mayotte demeure un territoire singulier en Afrique : c'est une île, un département français, et une région de petite taille, relativement homogène en termes de climat, d'environnement et de type de bâti. Il n'est donc pas possible d'évaluer pleinement le modèle ascendant en le testant uniquement sur Mayotte.

Nous avions donc besoin d'une autre zone géographique pour évaluer le modèle, et en particulier d'une région urbaine. Compte tenu de l'urbanisation croissante en Afrique (OCDE, Club du Sahel et de l'Afrique de l'Ouest, 2020; Salenson, 2020), il est essentiel que notre modèle soit en capacité d'estimer la population des villes et agglomérations urbaines d'Afrique. C'est ainsi que nous avons choisi de travailler sur la sous-préfecture d'Abidjan, composée de 10 communes, qui concentre à elle seule 36 % de la population urbaine de la Côte d'Ivoire (INS, 2022). Bien que l'idéal aurait été de cartographier également plusieurs autres régions, cela n'a pas été possible dans le temps imparti, c'est pourquoi nous présentons ici uniquement l'analyse de la sous-préfecture d'Abidjan (localisée dans le sud-est de la côte d'Ivoire, figure 60) pour l'année 2021.

Figure 60 : Carte de situation de la Côte d'Ivoire

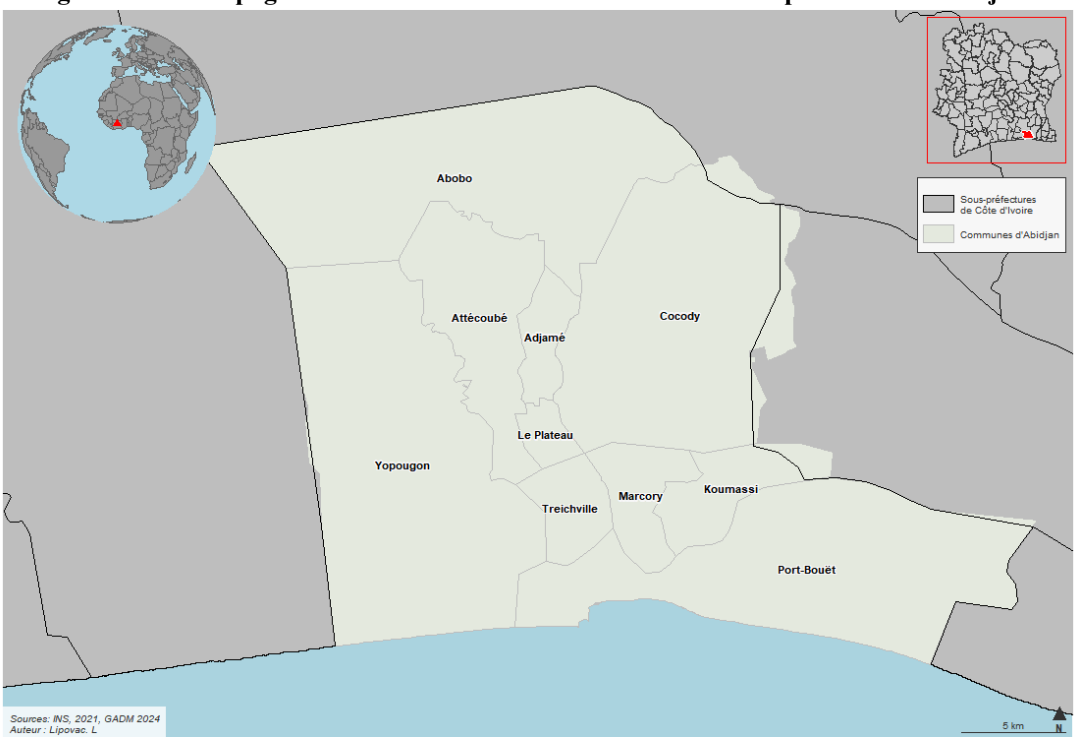
Crédits : Encyclopædia Universalis France

C’est en 2021 que s’est déroulé le dernier recensement de la population et de l’habitat en Côte d’ivoire qui a décompté 29 389 150 habitants (INS, 2022). Administrativement, la côte d’Ivoire est découpée en districts autonomes, en régions, en départements puis en sous-préfectures. La sous-préfecture d’Abidjan, dont le découpage en dix communes est illustré en figure 61, est la ville la plus peuplée de Côte d’Ivoire, anciennement capitale administrative et politique du pays.

3.1 La population et les zones administratives d’Abidjan

Les résultats globaux du 5ème recensement général de la population sont disponibles sur la page d’accueil de l’Institut National de la Statistique de la Côte d’Ivoire (INS) et sont distribués par départements puis par sous-préfectures/communes (INS, 2021)[59]. Les limites administratives sont en accès libre jusqu’au niveau le plus fin des sous-préfectures via des sources comme GADM[60]. Cependant, afin d’avoir accès aux tracés des dix communes de la sous-préfecture, nous avons dû en faire la demande à l’INS. Comme on peut le constater dans la figure 61, les fichiers géographiques des sous-préfectures et des communes ne se superposent pas parfaitement.

Figure 61 : Découpage administratif des dix communes de la sous-préfecture d’Abidjan :



Sources : Carte réalisée par l’auteur à partir de la couche des sous-préfectures de Côte d’Ivoire issue de GADM et de la couche des communes d’Abidjan issue de l’Institut National de la Statistique (INS)

[59] https://plan.gouv.ci/assets/fichier/RGPH2021-RESULTATS-GLOBAUX-VF.pdf

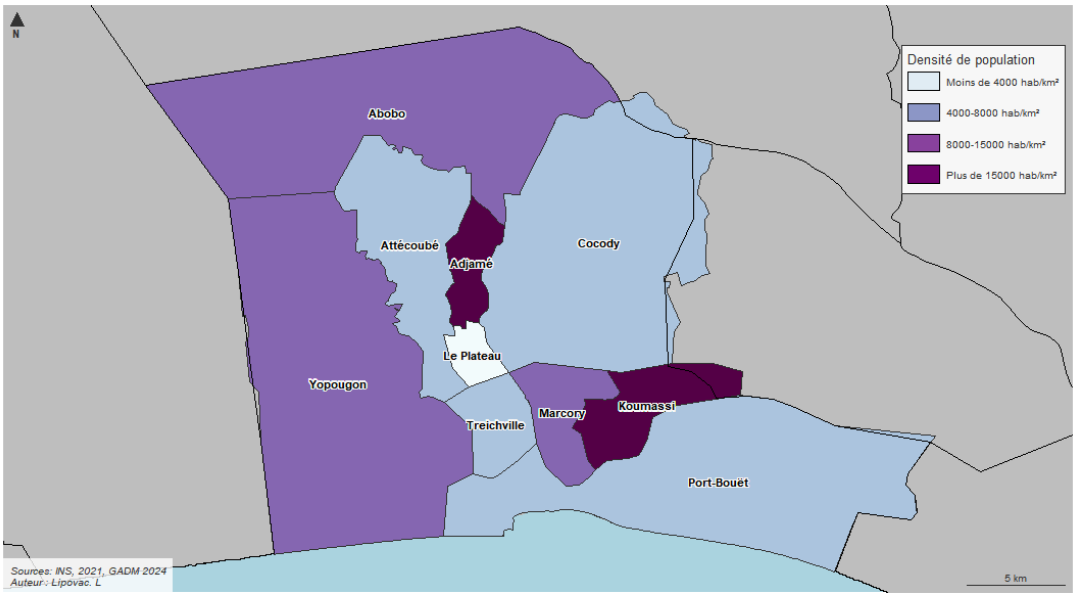
[60] https://gadm.org/download_country.html

Parmi ces communes, cinq couvrent une superficie inférieure à 22 km² (Le Plateau, Adjamé, Treichville, Marcory et Koumassi), tandis qu'Attécoubé s'étend sur 48 km² et Port-Bouët, Cocody, Abobo et Yopougon occupent entre 117 et 150 km². À titre de comparaison, la superficie moyenne d'une commune à Madagascar est de 292 km², mais les six arrondissements du district de la capitale – la zone la plus urbanisée du pays – , ont une superficie moyenne de seulement 14 km².

La Côte d'Ivoire affiche une densité de population moyenne de 91 habitants par km², mais la densité est considérablement plus élevée dans le district autonome d'Abidjan, où elle atteint 2 994 habitants par km² (INS, 2022). Au sein même de la sous-préfecture d'Abidjan, qui regroupe les dix communes présentées, la densité est encore plus élevée, atteignant 8 787 habitants par km².

Bien que toutes ces communes soient urbaines, elles présentent des profils démographiques contrastés, notamment en termes de densité de population, comme l'illustre la figure 62.

Figure 62 : Densité de population des dix communes de la sous-préfecture d'Abidjan



Sources : Carte réalisée par l'auteur à partir de la couche des sous-préfectures de Côte d'Ivoire issue de GADM et de la couche des communes d'Abidjan et décomptes de population issues de l'Institut National de la Statistique (INS)

Yopougon et Abobo sont les deux communes les plus étendues et les plus peuplées d'Abidjan, jouant un rôle commun important dans l'absorption de la croissance démographique de la ville (Steck, 2008). En 2021, elles affichent des densités de population similaires, autour de 10 300 habitants/km². Toutefois, cette valeur est calculée sur l'ensemble de leur superficie, alors que le sud de Yopougon et

l'ouest d'Abobo sont très peu urbanisés. La densité effective dans les zones réellement construites y est donc encore plus élevée.

Ces deux communes sont en grande partie de classe moyenne et populaire, avec de nombreux quartiers populaires, notamment à Abobo, où environ 60 % de la population vivait dans des quartiers précaires en 2008 (ONU-Habitat, 2012a). À Abobo, Yopougon (et Koumassi), plusieurs quartiers se sont développés sans planification urbaine formelle et ont vu leur population croître de manière importante au fil des années (CICG, 2023). C'est pourquoi ces zones ont été intégrées au Projet d'Aménagement des Quartiers Restructurés d'Abidjan (PAQRA), qui vise à restructurer certains quartiers précaires afin d'améliorer les conditions de vie des habitants et d'avoir « des conditions de vie décentes » pour chaque Ivoirien et Ivoirienne (CICG, 2023).

Les communes du Plateau, d'Adjamé et de Treichville sont situées au cœur d'Abidjan. Treichville reste un quartier résidentiel, en témoigne le fait que 66 % de sa population vit dans un habitat évolutif ou « cours commune » (ONU-Habitat, 2012b). La cour commune, symbole du vivre-ensemble et modèle d'habitation favorisant l'intégration par le partage d'espaces collectifs, est un type de bâti structuré autour d'une cour centrale, où se concentrent la plupart des activités quotidiennes (Manou-Savina, 1989; *NCI*, 2020). Treichville accueille également une activité commerciale dynamique, bien que moins importante qu'à Adjamé, qui abrite le plus grand centre commercial de la ville. De nombreuses personnes y travaillent, notamment dans les marchés et les gares routières, contribuant à la forte densité d'Adjamé, la plus élevée d'Abidjan avec près de 29 000 habitants par km².

Enfin, à l'échelle d'Abidjan, Le Plateau est très peu densément peuplé, avec seulement 1 139 habitants par km², ce qui s'explique par son statut de quartier d'affaires regroupant de nombreuses institutions gouvernementales et entreprises. Le Plateau est notamment caractérisé par ses nombreuses tours et grands immeubles.

A l'est d'Abidjan, on trouve la commune de Cocody, unique de par ses quartiers huppés et ambassades. Sa population est aisée et adopte pour une grande partie un mode vie occidental (Kouakou, 2019). On y trouve notamment de nombreux quartiers résidentiels.

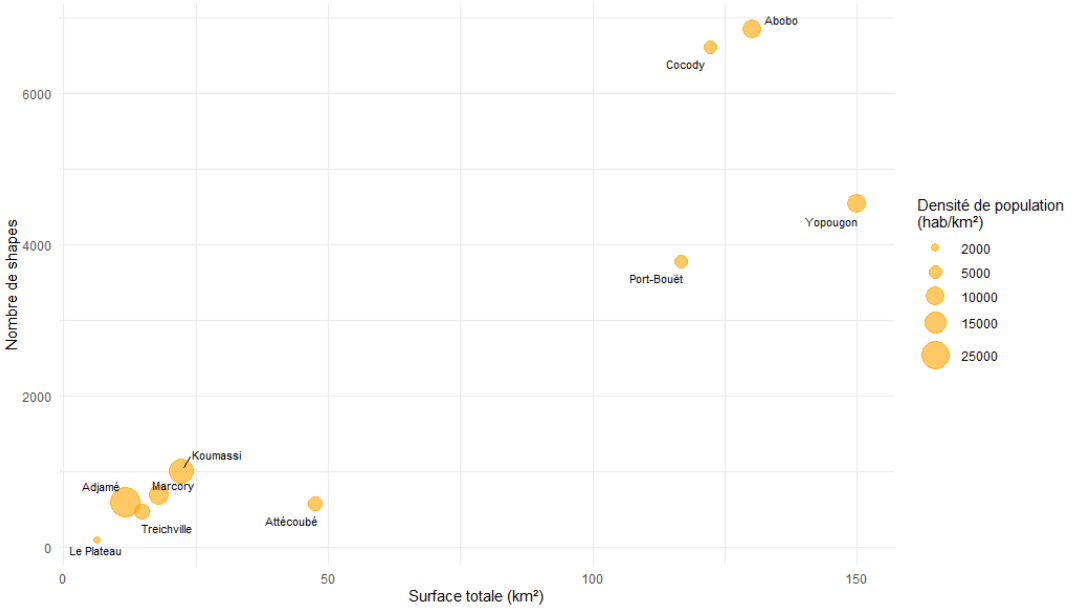
Les communes de Marcory, Koumassi et Port-Bouët au sud-est, accueillent en même temps des populations de classe moyenne et populaire. Il existe tout de même des zones résidentielles à Marcory comme on peut en trouver dans la commune de Cocody, alors que Koumassi et Port-Bouët ont plus des caractères industriels et populaires.

Pour finir, Attécoubé se distingue par sa configuration géographique unique, avec le parc national du Banco qui domine la commune, entouré de zones d'habitation réparties de part et d'autre de la baie. Ces quartiers sont situés sur un terrain accidenté, avec des ravins et des zones inondables, souvent occupés de manière non régulée et environ la moitié de l'espace bâti est caractérisée par une occupation informelle (Mairie Attecoube, 2024).

3.2 Cartographie des zones bâties d'Abidjan

Afin d'estimer la population des dix communes de la sous-préfecture d'Abidjan, nous les avons cartographiées et TeleCense a identifié 34 806 shapes. Les zones administratives étant relativement grandes, nous ajoutons seulement 187 shapes issues du découpage de shapes situées sur deux communes ou plus. Ensuite nous utilisons la détection issue de Google Open Buildings afin de modifier la surface des zones bâties identifiées par TeleCense. La majorité des shapes se trouvent dans les régions d'Abobo, de Cocody, de Port-Bouët et de Yopougon et on peut retrouver certaines différences entre les communes, par exemple Cocody et Abobo (en haut à droite de la figure 63) ont des surfaces et un nombre de de shapes similaires mais une densité de population différente. En effet, Abobo est bien plus densément peuplée, ce qui est confirmé par la taille moyenne des ménages de 4,8 à Abobo contrairement à 4,1 à Cocody (INS, 2021).

Figure 63 : Répartition des shapes en fonction de la surface dans les 10 communes d'Abidjan



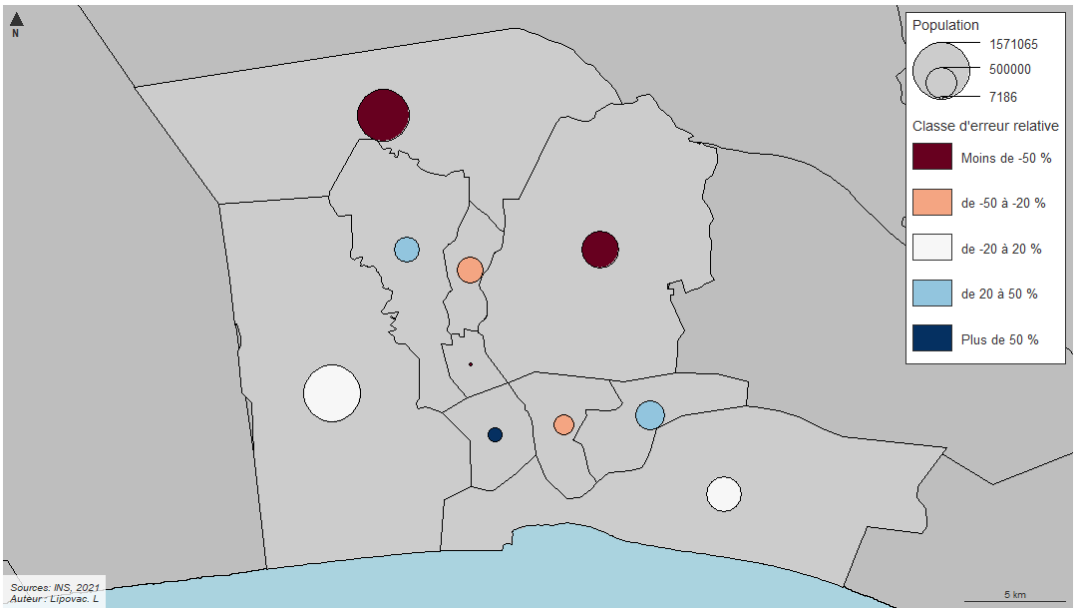
Sources : Calculs réalisés par l'auteur à partir de la cartographie du bâti TeleCense et des tracés et des décomptes de population issus de l'Institut National de la Statistique (INS, 2021)

Nous allons désormais appliquer le modèle ascendant aux shapes identifiées dans la sous-préfecture d'Abidjan. La méthode de vérification employée est la même que la première utilisée à Mayotte, à savoir que l'on agrège les estimations réalisées au niveau des shapes jusqu'au niveau des communes, puis on compare la somme des estimations aux décomptes officiels issus du RGPH 2021.

3.3 Résultats de l'approche ascendante à Abidjan

Les résultats du modèle ascendant appliqué à la sous-préfecture d'Abidjan montrent une population estimée agrégée de 7 574 447 habitants. En comparaison, les chiffres officiels du dernier RGPH indiquent une population de 5 616 634 habitants. Cette différence représente une surestimation d'environ 35 % par rapport aux données officielles. Seules les deux communes de Yopougon et de Port-Bouët sont considérées comme correctement estimées avec respectivement -4,6 % et 9,4 % d'erreur relative. Cela correspond à une surestimation d'environ 73 000 habitants et une sous estimation de 58 000 habitants. Les communes d'Adjamé et de Marcory sont surestimées et les communes d'Attécoubé et de Koumassi sont sous-estimées, avec une erreur relative de l'ordre de 40 %. Enfin, la commune de Treichville est fortement sous-estimée et communes du Plateau, d'Abobo et de Cocody sont très fortement surestimées avec des erreurs relatives respectives de -75 %, de -82 % et de -134 %. Abobo et Cocody sont particulièrement mal estimées avec une surestimation d'environ 1 million d'habitants. Les résultats sont succinctement présentés en carte (Figure 64) et dans le tableau 26.

Figure 64 : Erreur d'estimation au niveau communal de la sous-préfecture d'Abidjan



Sources : Carte réalisée par l'auteur à partir de la couche des sous-préfectures de Côte d'Ivoire issue de GADM et de la couche des communes d'Abidjan issue de l'INS.
Décomptes de population issus de l'Institut National de la Statistique (INS) et cartographie du bâti TeleCense

Tableau 26 : Estimation de la population et erreur relative des 10 communes d'Abidjan

Nom de la commune	Population RGPH-2021	Population estimée Modèle ascendant	Erreur relative (%)
Abobo	1 340 083	2 440 627	-82,1
Adjamé	340 892	492 365	-44,4
Attécoubé	313 135	205 927	34,2
Cocody	692 583	1 622 613	-134,3
Koumassi	412 282	240 001	41,8
Le Plateau	7 186	12 583	-75,1
Marcory	214 061	308 996	-44,3
Port-Bouët	618 795	560 796	9,4
Treichville	106 552	46 501	56,4
Yopougon	1 571 065	1 644 039	-4,6

Sources : Calculs réalisés par l'auteur à partir des décomptes de population issus de l'Institut National de la Statistique (INS) et de la cartographie du bâti TeleCense

3.4 Un modèle ascendant pas encore universel

Dans la section précédente, nous avions souligné la performance du modèle à Mayotte, où l'erreur d'estimation au niveau départemental n'atteint que 3 % pour une population de 320 000 habitants répartis sur 374 km². Si l'on applique la même logique à la commune de Yopougon – qui compte près de 1,6 million d'habitants sur une superficie de 150 km², caractérisée par un habitat populaire en perpétuelle évolution – une erreur de 4,6 % à cette échelle pourrait également être perçue comme un résultat satisfaisant. En effet, la capacité du modèle ascendant à fournir des estimations précises à l'échelle de certaines communes ou départements confirme que l'approche globale développée dans cette thèse peut fonctionner.

Par ailleurs, il n'est pas surprenant que Port-Bouët figure parmi les communes les mieux estimées. Cette commune présente des similarités avec Madagascar en termes de typologie d'habitat, notamment une forte prévalence de logements modestes d'une à deux pièces : en 2012, environ 65 % des ménages y vivaient dans ce type de logement (ONU-Habitat, 2012c). Ainsi, le modèle y affiche une légère sous-estimation, avec une erreur de 9,4 %.

Toutefois, bien que l'estimation agrégée au niveau communal soit proche des décomptes du RGPH 2021, il est probable qu'avec une granularité équivalente à celle de Mayotte, permettant des vérifications au niveau des shapes, nous observerions des erreurs similaires. Comme à Mayotte, ces écarts pourraient en partie se compenser à un niveau agrégé, masquant ainsi des imprécisions plus localisées.

Justement, dans les autres communes de la sous-préfecture d'Abidjan, cette compensation ne semble pas se produire. En témoigne l'importance des erreurs relatives dans huit des dix communes étudiées. Ce constat soulève une limite majeure

du modèle ascendant : dans son état actuel, il ne semble pas généralisable en dehors du contexte malgache.

Les cas d'Abobo et de Cocody sont particulièrement révélateurs. Dans ces deux communes, les erreurs d'estimation s'accumulent principalement dans le sens d'une surestimation, et lorsqu'elles sont répétées plusieurs milliers de fois, l'erreur finale peut être importante

Dans le cas de Cocody, il est probable que le modèle ait surestimé le nombre d'habitants par habitation. En effet, du fait du niveau de vie élevé dans cette commune, les logements y sont généralement moins densément peuplés. La taille moyenne des ménages y est de 4,1 personnes, soit la plus faible de la sous-préfecture. De plus, la présence d'un nombre plus élevé de logements secondaires et/ou vacants pourrait accentuer cette surestimation.

En revanche, les résultats obtenus pour Abobo sont plus surprenants. Cette commune présente un habitat qui, à première vue, aurait dû être mieux estimé par le modèle, notamment en raison de sa similitude avec certaines zones de Madagascar. Une hypothèse pourrait être une mauvaise identification du bâti spécifique à cette année-là, mais cela reste à approfondir.

Un autre cas intéressant est celui de Treichville. Il est possible que l'évaluation des cours communes y pose un problème, entraînant une sous-estimation du nombre d'habitants par shape. Il faudrait approfondir afin de comprendre si le modèle attribue trop peu de population à ces types d'habitats collectifs.

Enfin, la commune du Plateau illustre bien les limites actuelles du modèle face aux environnements urbains verticaux. Bien que l'erreur relative y soit importante, elle ne représente en réalité qu'un écart d'environ 5 000 habitants. Avec un faible nombre de shapes, il suffit de mal estimer quelques immeubles – une difficulté récurrente dans notre approche qui ne prend pas encore en compte la hauteur des bâtiments – ou encore de ne pas identifier correctement des édifices non résidentiels (bureaux, hôtels, ambassades, etc.) pour entraîner une surestimation rapide de la population.

Ces résultats mettent en évidence plusieurs limites du modèle ascendant dans son application à la sous-préfecture d'Abidjan. L'une des principales raisons réside dans les différences significatives entre le contexte d'Abidjan et celui sur lequel le modèle a été entraîné. La structure du bâti y est très hétérogène, et l'estimation de la population représente un véritable défi pour le modèle ascendant. Le tableau 27 illustre ces écarts : alors qu'à Mayotte et Madagascar, les médianes des populations estimées par shape sont respectivement de 5 et 6 habitants, avec des surfaces médianes

de 200 et 300 m², celles d'Abidjan atteignent 148 habitants pour 2 900 m². Ces disparités se renforcent lorsque l'on analyse les distributions complètes, Abidjan présentant à la fois des shapes de surfaces beaucoup plus étendues et un nombre de shapes nettement plus élevé à ces niveaux de surface.

Même en comparant ces résultats aux six arrondissements d'Antananarivo, où environ 3 000 shapes ont été cartographiées, l'écart demeure important. Certes, environ 15 % des shapes y sont comparables à celles d'Abidjan en termes de surface et de population, mais elles ne représentent qu'environ 500 unités, un effectif insuffisant pour assurer la robustesse et l'exportabilité du modèle à un contexte de très forte urbanisation comme celui d'Abidjan. Il apparaît donc que l'entraînement du modèle sur des données issues majoritairement de Madagascar a une influence significative sur les résultats obtenus.

Tableau 27 :
Comparaison des distributions de population et de surface des shapes selon les régions étudiées

Région		1er quantile	Médiane	3ème quantile	85 %	95 %	99 %
Abidjan	Population estimée	22	148	358	512	1 008	2 583
	Surface (m^2)	400	2 900	6 300	8 900	18 700	48 000
Madagascar	Population désagrégée	3	6	15	26	82	333
	Surface (m^2)	100	300	500	900	3 100	12 100
Mayotte	Population estimée	2	5	23	51	145	350
	Surface (m^2)	100	200	900	1 900	4 400	8 800
Antananarivo (3 000 Shapes)	Population désagrégée	6	23	176	450	2 200	6 400
	Surface (m^2)	100	400	3 600	9 300	42 000	125 000

Sources : Calculs réalisés par l'auteur à partir des données issues du programme TeleCense

Afin de valider cette hypothèse, nous avons tenté de construire un modèle basé uniquement sur les shapes des six arrondissements d'Antananarivo. Cependant, cette approche a conduit à une surestimation encore plus marquée de la population d'Abidjan, confirmant ainsi que les différences structurelles du bâti constituent un facteur déterminant des écarts observés.

Dès lors, comment expliquer précisément ces écarts ? Comme observé à Mayotte, un des problèmes majeurs réside sans doute dans l'identification et la segmentation du bâti que nous réalisons.

L'entraînement d'un modèle exclusivement sur les données d'Antananarivo pour estimer la population urbaine d'Abidjan pouvait sembler pertinent. Cependant, le fait que cette approche ait conduit à une surestimation encore plus marquée de la population est révélateur des limites de cette méthodologie. À Madagascar, comme

illustré dans le chapitre 3 (figure 25), ou encore comme au Bénin (chapitre 4), les six arrondissements de la capitale présentent eux aussi de très grandes shapes, qui englobent en réalité plusieurs zones bâties aux caractéristiques différentes. Comme l'identification des shapes, via l'estimation de sa surface, est un paramètre fondamental, il apparaît désormais clairement que baser l'entraînement du modèle sur cette identification constitue une contrainte majeure. La segmentation imprécise des shapes nous empêche de les caractériser davantage.

Évidemment, les erreurs d'estimation peuvent être attribuées à de nombreux autres facteurs. Cependant, il est regrettable qu'une segmentation initiale imprécise du bâti limite notre capacité à les analyser en détail. Prenons, par exemple, la situation de Cocody, qui illustre bien cette complexité : certaines zones comptent un grand nombre de logements secondaires, ce qui fausse les estimations (comment les identifier et, même si cela était possible, quel poids leur accorder ?). De même, la présence de logements vacants, notamment dans des zones en construction en périphérie, complique l'évaluation précise de la population.

Un autre phénomène à considérer est celui de la décohabitation. À Mayotte, lors de notre collaboration avec l'Insee entre 2017 et 2024, nous avons observé une augmentation du nombre d'habitations, accompagnée d'une diminution de la population dans certains villages (ex. : Bouéni, Moinatrindri, Chembenyoumba). Cette évolution peut s'expliquer par une réduction de la taille des ménages, mais aussi par des dynamiques résidentielles spécifiques : les habitants quittent temporairement une zone pendant la construction de nouveaux logements, avant d'y revenir plus tard. Ce sont autant de facteurs contextuels qu'il serait nécessaire d'étudier plus en profondeur.

Bien sûr, ces problèmes ne pourront pas être résolus du jour au lendemain avec des techniques de télédétection, d'autant plus avec des images d'une résolution aussi limitée que celle des satellites Sentinel (10 mètres). Toutefois, une segmentation plus précise et une base solide permettraient d'affiner ces analyses, dans l'objectif d'améliorer nos estimations et de développer des modèles cohérents pour estimer la population dans le temps, sans dépendre systématiquement de données démographiques préexistantes.

C'est précisément dans cette perspective que nous allons consacrer le dernier chapitre de cette thèse à une nouvelle approche. Dans les faits, nous reprendrons la méthodologie développée jusqu'ici, à savoir la désagrégation appliquée à Madagascar suivie de l'entraînement d'un modèle basé sur les estimations réalisées au niveau des shapes. Cependant, cette fois-ci, nous utiliserons un jeu de données inédit :

l'identification des zones bâties par le projet *Google Open Buildings 2.5D*[61], publié en septembre 2024.

Ce nouvel ensemble de données présente plusieurs avantages majeurs. Tout d'abord, il intègre des informations sur la hauteur des bâtiments, ouvrant ainsi la possibilité d'explorer leur contribution, notamment dans des contextes comme Abidjan, où de nombreux bâtiments, comme des immeubles ont une hauteur significative. Ensuite, bien que reposant en partie sur des images Sentinel, il bénéficie d'un entraînement sur des images Google de très haute résolution (0,5 à 1 mètre), aboutissant à une résolution effective d'environ 4 mètres, soit une amélioration significative par rapport aux 10 mètres de Sentinel utilisés dans TeleCense.

Enfin, ces nouvelles données nous permettront de comparer nos résultats avec ceux obtenus via TeleCense et d'évaluer si une segmentation améliorée du bâti constitue un facteur d'amélioration déterminant, ce qui viendrait confirmer ou affiner nos hypothèses.

4. Conclusion

Ce chapitre a présenté le dernier volet de notre approche globale, à savoir le développement de l'approche ascendante basée sur les résultats de la désagrégation de la population dans les shapes de Madagascar.

Nous avons d'abord construit le modèle en testant une régression linéaire et un algorithme de type forêt aléatoire, avec validation croisée sur un échantillon d'entraînement et un échantillon de validation. Bien que la forêt aléatoire ait montré des performances statistiques significativement supérieures, le modèle linéaire a, lors des applications sur les deux régions, systématiquement obtenu de meilleurs résultats en moyenne. Cela s'explique probablement par l'importance trop forte accordée à la variable de surface des shapes lors de la forêt aléatoire. C'est donc cette approche par modèle linéaire qui a été retenue pour les applications sur Mayotte et Abidjan.

À Mayotte, notre collaboration avec l'Insee nous a permis d'accéder directement aux îlots administratifs et à leurs populations. Grâce à cela, nous avons pu appliquer le modèle linéaire sans étapes intermédiaires complexes. Les premiers résultats étaient encourageants : en agrégeant les estimations obtenues au niveau des shapes, nous avons estimé une population de 246 470 habitants en 2017, soit une sous-

[61] https://sites.research.google/gr/open-buildings/temporal/

estimation d'environ 3 % par rapport au décompte officiel. Néanmoins, l'erreur augmente à mesure que l'on descend en précision, notamment au niveau des villages et des îlots.

La richesse des données disponibles à Mayotte a été un atout majeur pour évaluer la précision des estimations. En désagrégeant à partir des îlots, nous avons pu comparer directement les résultats du modèle ascendant avec des données désagrégées issues de très petites zones administratives, ce qui a limité les erreurs de répartition dans les shapes. Nous avons notamment observé que l'erreur était particulièrement marquée pour les shapes de grande surface, souvent composées de nombreux bâtiments aux caractéristiques variées. Ces zones entraînent fréquemment des sous-estimations ou surestimations importantes, confirmant la nécessité d'améliorer la segmentation du bâti pour affiner les estimations de population.

Fort de ces constats, nous avons appliqué le modèle ascendant à la sous-préfecture d'Abidjan, en estimant la population des shapes des dix communes qui la composent. Cependant, les résultats ont montré que, dans sa forme actuelle, le modèle ascendant n'est pas encore pleinement opérationnel pour estimer la population de zones géographiques en dehors de Madagascar, en particulier dans un contexte aussi densément peuplé qu'Abidjan. Cela confirme par ailleurs qu'il aurait été plus qu'intéressant de tester le modèle dans une autre région de Côte d'Ivoire afin de pouvoir comparer les résultats. Plusieurs raisons expliquent ces limites, mais la principale rejoint nos observations précédentes : la difficulté à caractériser les shapes de grande surface.

Ces constats soulignent la nécessité d'une caractérisation plus fine du bâti pour améliorer les estimations. C'est précisément l'objectif du prochain chapitre, qui explorera l'apport d'images satellitaires à plus haute résolution. En affinant la détection du bâti, nous espérons améliorer la précision du modèle ascendant et, à terme, proposer une approche englobant plus de variables socio-démographiques et permettant d'estimer la population sans données démographiques en entrée.

Chapitre 6 – Estimation de population à partir des nouvelles données de bâti à très haute résolution issues de la détection Google 2.5D

Le dernier chapitre de cette thèse est consacré à la base de données *Open Buildings 2.5D*, développée par Google Research, publiée en septembre 2024[62]. Ce jeu de données, en plus d'offrir une représentation détaillée et à haute résolution des surfaces bâties, introduit deux informations clés jusqu'alors absentes dans nos approches : la hauteur des bâtiments et le nombre de bâtiments dans un voisinage proche. Ces avancées, combinées au fait que ces données sont en accès libre et disponibles pour une grande partie de la planète (Afrique, Asie du Sud, Asie du Sud-Est, Amérique latine et Caraïbes), pour une période allant de 2016 à 2023, à un pas annuel, ouvrent des perspectives nouvelles pour la modélisation de la population.

Ce chapitre porte donc sur l'intégration des données Google 2.5D dans les modèles descendants et ascendants vus aux parties 2 et 3. Ce jeu de données est utilisé pour évaluer son potentiel pour améliorer les résultats obtenus à partir de TeleCense, et pour répondre à certaines des limites identifiées dans les chapitres précédents. Il s'agit d'une première exploration, permettant de proposer de nouveaux résultats tout en ouvrant la voie à de nombreux autres développements.

Dans un premier temps, les caractéristiques des données Google 2.5D sont présentées, ainsi que leur méthode d'acquisition et les ajustements éventuels visant à affiner davantage la segmentation du bâti, permettant une meilleure séparation des bâtiments entre eux. Ensuite, les résultats obtenus à Madagascar serviront de base à un nouveau modèle d'estimation de population, qui sera testé dans le contexte ivoirien. Les performances de ce modèle seront analysées et comparées à celles des approches précédentes, permettant d'évaluer la valeur ajoutée réelle de ces données. En résumé, nous reprenons notre approche globale réalisée dans les chapitres 3 et 5 de cette thèse mais à partir d'une autre détection du bâti.

Enfin, les limites, tant des données en elles-mêmes que de la démarche méthodologique, seront discutées. Ces points souligneront les axes potentiels d'amélioration, tout en ouvrant sur de nouvelles perspectives, particulièrement prometteuses grâce à ces données innovantes.

[62] https://sites.research.google/gr/open-buildings/temporal/

1. Présentation et objectif des données Google 2.5D

L'identification de bâti à très haute résolution opérée par Google Research, nommée *Open Buildings*[63], est déjà utilisée dans les premières parties de cette thèse. Elle permettait de calculer la surface modifiée par Google dans chacune des shapes identifiées par TeleCense. Toutefois, un problème majeur résidait dans l'absence de datation précise : ces données résultaient d'une mosaïque d'images acquises sur plusieurs années, sans qu'il soit possible de déterminer précisément l'année de chaque observation.

Elle a fait l'objet d'une récente extension publiée en septembre 2024. Désormais, Google Research propose une base de données temporelle couvrant les années 2016 à 2023, offrant non seulement la présence des bâtiments, mais également une estimation de leur hauteur. Ce nouveau projet, intitulé *Open Buildings 2.5D – Temporal Dataset*[64], ou Google 2.5D, s'appuie sur deux sources de données complémentaires. Tout comme le programme TeleCense, Google 2.5D mobilise les images satellites Sentinel-2 à une résolution de 10 mètres. C'est une ressource précieuse car gratuite et disponible à haute fréquence (tous les deux à cinq jours) à travers le monde (Sirko et al., 2024). Cependant, contrairement à TeleCense, ce programme repose sur des données supplémentaires : des images à très haute résolution (de 50 cm à 1 m) déjà utilisées dans le produit initial, *Open Buildings*.

Grâce à l'association de ces deux types de données, *Google Research* a développé une méthode qui permet, à partir d'un grand volume d'images Sentinel-2 à faible résolution, de produire des résultats à haute résolution. Cette méthode repose sur deux modèles principaux :

- Le modèle enseignant (*teacher model*) : formé sur des images à très haute résolution (0,5 à 1 mètre), il génère des cartes précises de bâti et de routes.
- Le modèle étudiant (*student model*) : entraîné à partir des images Sentinel-2 à faible résolution, il apprend à imiter les prédictions du modèle enseignant. Ainsi, même sans avoir accès aux images à très haute résolution, le modèle étudiant parvient progressivement à reproduire une segmentation de haute qualité à partir des images d'entrée de résolution plus basse.

Bien que ce processus entraîne une légère perte de précision, la performance globale est maintenue, et la résolution effective obtenue est de 4 mètres. Pour chaque

[63] https://sites.research.google/gr/open-buildings/

[64] https://sites.research.google/gr/open-buildings/temporal/

année de la période 2016-2023, les 32 meilleures images Sentinel disponibles (dans certains cas, par exemple pour les zones très nuageuses, il peut y en avoir moins) sont sélectionnées. Ensuite, une image à très haute résolution est choisie, idéalement de manière à ce qu'elle corresponde au centre de la séquence, c'est-à-dire entre la 16e et la 17e image des 32 images Sentinel sélectionnées (Sirko et al., 2024). Cette approche permet d'estimer, pour chaque année, la présence des bâtiments autour du 30 juin, et ce, pour la quasi-totalité des pays du Sud. Les sorties fournies par ce modèle comprennent pour tout point :

- La présence d'un bâtiment, avec un indice de confiance (variant de 0 à 1).
- Le nombre de bâtiments dans une certaine zone (en estimant le centroïde des bâtiments)
- La hauteur estimée du bâti, en mètres.

Bien que les données Google 2.5D apportent des avancées significatives, elles présentent a priori des limites importantes à prendre en compte. Premièrement, l'estimation de la hauteur des bâtiments repose sur un modèle entraîné exclusivement avec des données provenant de pays du Nord, où des références fiables étaient disponibles. Cette caractéristique des données d'entraînement pourrait induire un biais et/ou limiter la précision des prédictions dans les pays du Sud, où les contextes architecturaux et urbains diffèrent souvent. Bien que *Google Research* indique une erreur absolue moyenne de 1,5 mètres dans la prédiction des hauteurs, on peut s'attendre à des imprécisions, notamment dans les premières étapes d'utilisation dans des contextes peu similaires à ceux de l'entraînement initial.

Une autre limite, cette fois liée aux satellites Sentinel, est celle de la couverture nuageuse, particulièrement problématique dans les zones tropicales très humides. Dans ces régions, il arrive que très peu d'images exploitables soient disponibles sur une période donnée, ce qui peut fortement affecter la qualité des estimations du bâti et rendre l'utilisation des données Google 2.5D fortement limitée pour certaines régions.

Enfin, ces données sont aujourd'hui en accès libre et comportent des informations très récentes (2023) mais leur accès futur est conditionné par la volonté de *Google Research* de continuer à les publier et de les mettre à jour de manière régulière. Cette dépendance au bon vouloir d'une société privée, une GAFA, dont ce service n'est pas l'intérêt primaire, peut poser des questions sur la pérennité et la stabilité des analyses basées sur ces données dans le futur.

1.1 Téléchargement des données

Contrairement au produit original de Google, *Open Buildings*, qui proposait des sorties au format polygone directement exploitables dans les logiciels SIG, les données Google 2.5D présentent un format de sortie plus complexe et volumineux. En effet, bien que la résolution effective des bâtiments soit de 4 mètres, les données disponibles au format GeoTIFF (raster) sont d'une résolution de 50 cm. Elles sont composées de trois bandes, chacune représentant une information spécifique (voir figure 65, images b), c) et d) :

- Bande 1 - building_fractional_count [0, 0.0216] : Donnée calculée à partir de la détection des centroïdes des bâtiments (Image b) de la figure 1). Elle permet de connaître le nombre de bâtiments dans une zone en additionnant la somme de la valeur des pixels ;

- Bande 2 - building_height [0, 100] : Hauteur en mètres comptée de 0.5 en 0.5 mètre illustrée dans l'image c) de la figure 1 ; plus c'est blanc plus le bâtiment a de la hauteur. Cette valeur doit être corrélée avec la présence de bâtiments pour être pertinente ;

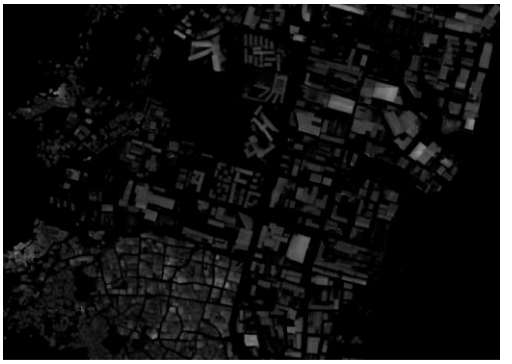
- Bande 3 - building_presence [0, 1] : Confiance du modèle sur la présence d'un bâtiment. L'ordre des valeurs peut varier en fonction des régions. Le seuil conseillé en Afrique est de 0.34 (Sirko et al., 2024), et de la même manière plus la couleur est blanche dans l'image d) plus le seuil est élevé.

Figure 65 : Présentation des données Google 2.5D dans nord est de Mamoudzou à Mayotte - 2023 Mamoudzou

a) Image satellite 2023

b) Bande 1 – Détection des centroïdes des bâtiments

c) Bande 2 – Détection de la hauteur des bâtiments



d) Bande 3 – Présence bâtiments

Sources : Identification du bâti issue du programme Open Buildings 2.5D Temporal Dataset

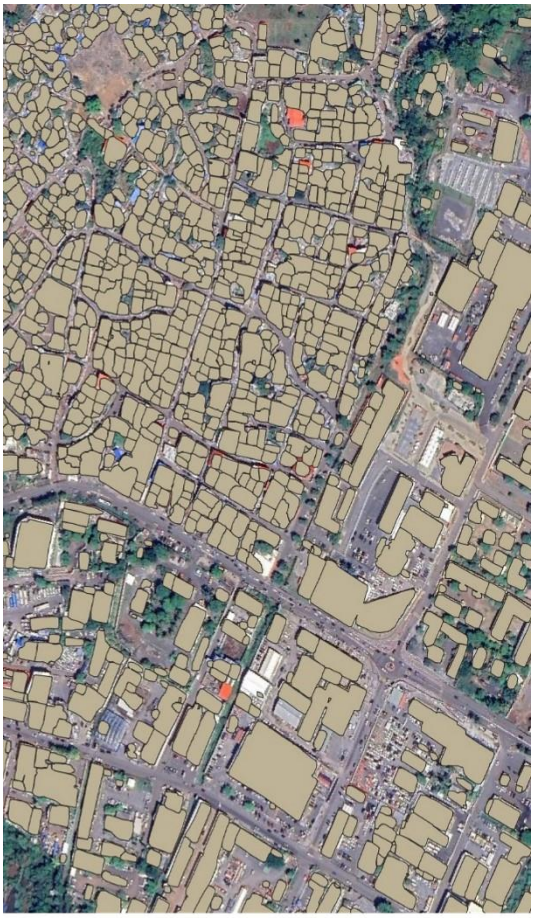
1.2 Utilisation et préparation des données

Pour identifier le bâti, nous appliquons le seuil recommandé de 0,34 pour la présence des bâtiments (Sirko et al., 2024). Les données issues de Google 2.5D (Figure 66), déjà d'une grande précision, pourraient être utilisées directement.

Nous choisissons néanmoins de retravailler légèrement ces données afin d'améliorer la segmentation entre les bâtiments et d'obtenir une délimitation plus fine, bâtiment par bâtiment. Ce traitement supplémentaire est réalisé en collaboration avec l'équipe TeleCense, mais les détails précis de cette étape ne peuvent être développés ici pour des raisons de confidentialité. La différence est visible sur les figures 66 et 67 : à gauche, l'identification du bâti avec le seuillage à 0,34 recommandé par Google 2.5D, et à droite, la version retravaillée offrant une meilleure individualisation des bâtiments.

Figure 66 : Identification du bâti Google 2.5D avec le seuil conseillé de 0,34

Figure 67 : Identification du bâti Google 2.5D – toujours avec un seuil de 0,34 - retravaillée pour une meilleure segmentation des bâtiments entre



Sources : Identification du bâti issue du programme Open Buildings 2.5D Temporal Dataset

Ainsi, pour la suite de ce chapitre c'est la détection illustrée dans la figure 3 que nous allons utiliser ; elle est bien issue de Google 2.5D mais elle est légèrement retravaillée.

Ensuite, nous devons déterminer si les bâtiments sont habitables ou non et pour cela, nous utilisons de nouveau OpenStreetMap (OSM). L'objectif est d'exclure de l'estimation de population les bâtiments non résidentiels (bâtiments administratifs, entrepôts, magasins, hôpitaux, écoles, etc.). Pour rappel, avec l'identification TeleCense l'approche consistait à croiser les shapes avec les données OSM pour repérer et éliminer les zones non résidentielles, en calculant un pourcentage de la zone habitée (chapitre 2 – Figure 13). Autrement dit, on supprimait dans la shape, la surface des bâtiments considérés comme non habités. Compte tenu des surfaces parfois importantes des shapes, nous estimions que cette approche suffisait, et c'était de toute manière l'unique manière de faire.

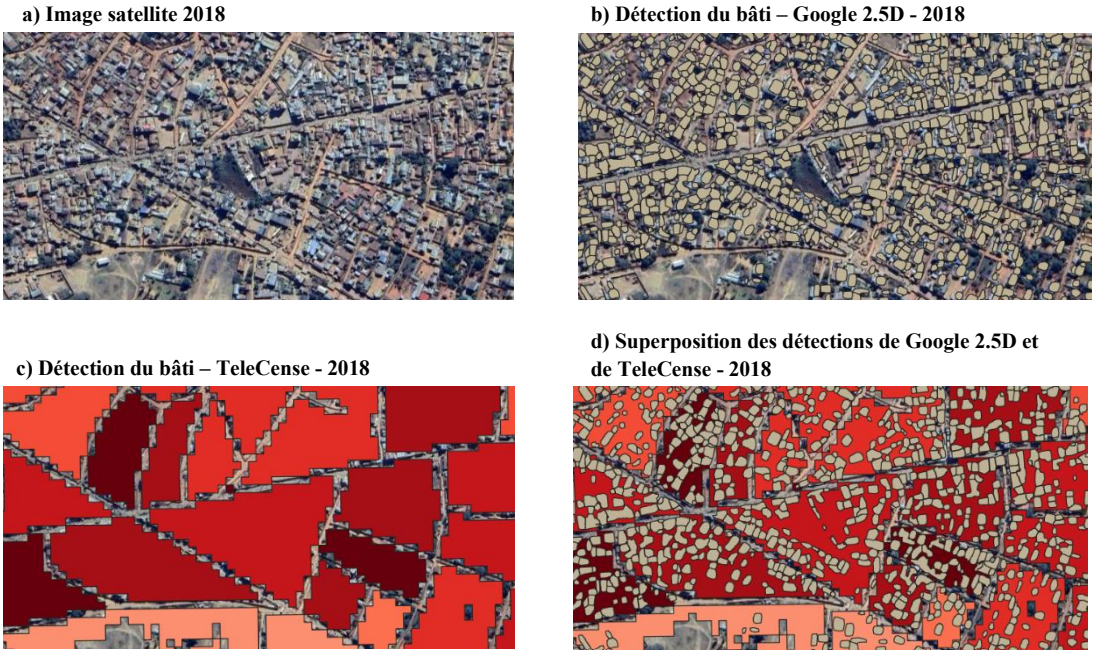
Cependant, l'identification du bâti par Google 2.5D présente une différence majeure par rapport à TeleCense : les shapes sont significativement plus petites et représentent, dans la plupart des cas, des bâtiments individuels. Bien qu'OSM soit un outil puissant, notamment dans les grandes villes, son étiquetage repose sur une contribution bénévole et il est de fait sujet à des disparités de précision selon les pays et les régions étudiées.

Malgré cela, OSM reste la seule source disponible pour attribuer des catégories aux shapes de manière automatique et efficace afin d'éviter d'estimer une population sur des bâtiments non résidentiels. C'est donc ce qu'on va continuer à utiliser mais il faut néanmoins être conscient des limites, notamment si nous commençons à utiliser la hauteur dans nos modélisations. On peut par exemple penser à un entrepôt non identifié comme tel qui pourrait se voir attribuer une population élevée simplement en raison de sa grande hauteur.

Une fois ces traitements réalisés, les shapes sont identifiées et enrichies avec de nouvelles variables. La surface est calculée à partir de la géométrie propre à chaque shape. La hauteur est obtenue à partir de la couche « buildings_height » et le volume de la shape est calculé en multipliant la surface par la hauteur. Pour finir le nombre de bâtiments par hectare est calculé en utilisant la couche « building_fractional_count », ce qui permet d'évaluer la densité des bâtiments dans le voisinage. Cette nouvelle variable de densité de bâtiment est propre à chaque shape car on calcule dans un hectare autour de chaque shape le nombre de bâtiments. Chaque shape a donc une valeur propre à elle, contrairement à la variable relativement similaire de proportion de surface bâtie calculée avec l'identification TeleCense qui était une variable calculée au niveau de la commune et donc chacune des shapes d'une même commune avait la même valeur.

Afin de conclure la présentation des shapes issues de la nouvelle détection Google 2.5D, nous la comparons à celle effectuée par TeleCense. Les illustrations de la figure 68 montrent la ville d'Imerintsiatosika dans la région d'Itasy à Madagascar, et présentent successivement une image à haute résolution des bâtiments, suivie des résultats de la détection Google 2.5D, puis de la détection TeleCense, et enfin des deux détections superposées pour mieux visualiser les différences.

Figure 68 : Comparaison des détections Google 2.5D et TeleCense dans la ville d'Imerintsiatosika – Région d'Itasy – 2018



Sources : Identification du bâti issue des programmes Open Buildings 2.5D Temporal Dataset et TeleCense

En résumé, la différence entre la détection Google 2.5D et celle de TeleCense réside principalement dans la taille des shapes en sortie. TeleCense fournit une détection plus générale, limitée par les données Sentinel à 10 mètres, ce qui empêche d'atteindre un niveau de précision supérieur. En revanche, avec Google 2.5D, la détection plus fine et précise, bâtiment par bâtiment, permet de changer d'ambition et de passer à une estimation de la population au niveau du bâtiment. Cela est également rendu possible par le fait que – contrairement à la première édition d'Open Buildings – la détection du bâti est datée. On peut donc utiliser directement la sortie 2.5D et non pas une mosaïque de bâtis sur plusieurs années, qui amenait forcément sont lot d'incertitudes.

1.3 Réinterpréter les modèles avec une identification de bâti plus fine

L'objectif de ce chapitre est de confirmer le potentiel des données Google 2.5D et d'en évaluer l'impact sur les approches de modélisation démographique. Pour cela, nous reprenons le travail effectué dans les parties deux et trois de cette thèse. Comme nous avons de nouvelles données de détection de bâti, nous allons les utiliser et mettre en place les approches descendante et ascendante vues précédemment. Cela va nous

permettre de comparer les résultats entre détection TeleCense et détection Google 2.5D.

Ainsi, dans un premier temps, nous réalisons une désagrégation à Madagascar, dont l'objectif est de répartir la population des 430 communes dans les shapes identifiées par Google 2.5D. Un premier objectif sera d'évaluer l'apport de la nouvelle surface des shapes : est-ce que l'interpolation surfacique donne des résultats avec moins d'erreurs qu'avec les shapes de TeleCense ? Pour cela nous réalisons une interpolation surfacique comme suit :

$$(1)\ Population_{shape} = \frac{Population_{ZA} * Aire_{shape}}{\Sigma(Aire_{shape})}$$

Si la désagrégation s'avère plus précise, cela pourrait valider l'apport d'une détection du bâti plus fine et continue dans le temps.

Un second objectif sera d'évaluer si l'utilisation de la hauteur des bâtiments permet une amélioration significative des résultats par rapport aux méthodes précédentes. Pour cela, on effectue une interpolation par le volume des bâtiments, rendue possible en intégrant la variable de hauteur, ainsi :

$$(2) Population_{shape} = \frac{Population_{ZA} * Aire_{shape} * Hauteur_{shape}}{\Sigma(Aire_{shape} * Hauteur_{shape})}$$

On pourrait par exemple s'attendre à ce que la désagrégation utilisant le volume des bâtiments (surface * hauteur) soit plus efficace que celle fondée seulement sur la surface.

Enfin, nous statuerons de la méthode qui offre les meilleurs résultats et les utiliserons afin de construire un nouveau modèle ascendant. Ce modèle sera entraîné sur un échantillon de 70 % des données malgaches, validé statistiquement sur les 30 % restants, et enfin testé sur les sous-préfectures d'Abidjan.

Ce dernier chapitre se veut avant tout exploratoire. Il constitue un premier pas vers une validation des apports des données Google 2.5D. Les résultats obtenus serviront de base pour envisager des recherches plus ambitieuses, dont les perspectives seront discutées en conclusion de ce chapitre et de cette thèse.

2. Répartition de la population dans les shapes Google 2.5D : une nouvelle désagrégation à Madagascar

Afin d'effectuer la désagrégation, nous nous appuyons sur exactement les mêmes données que dans la partie 2, à savoir les 430 communes couvrant les six régions de Madagascar. La différence se situe dans l'identification des zones bâties, qui est désormais issue de Google 2.5D. Au total dans les six régions, 1 343 443 shapes sont identifiées (dont 600 529 à Analamanga, 62 745 à Androy, 237 066 à Atsinanana, 201 516 à Diana, 177 616 à Itasy et 63 971 à Melaky) ; soit un peu plus de quatre fois plus de shapes qu'avec l'identification TeleCense.

Pourtant, malgré ce nombre bien plus élevé, la surface totale de la détection Google 2.5D est environ deux fois plus petite que celle obtenue avec TeleCense. Cela confirme que les shapes sont, en moyenne, de plus petite taille et mieux séparées entre elles ; tout en gardant en tête que cette surface bien moins élevée pourrait aussi venir d'une sous-estimation de la surface bâtie par Google 2.5D.

Les shapes de Madagascar se caractérisent par leur hauteur relativement modeste. En effet, la distribution des hauteurs révèle une médiane à 3,4 mètres, puis 90 % des bâtiments mesurent moins de 5,6 mètres et 99 % des bâtiments ont une hauteur inférieure à 10 mètres. Finalement, seuls quelques bâtiments dépassent les 20 mètres de hauteur sur l'ensemble du territoire malgache.

Les deux méthodes de désagrégation sont mises en œuvre (interpolation par la surface, puis par le volume des shapes) et nous décidons d'évaluer les résultats, puis de les comparer avec ceux obtenus avec TeleCense en utilisant les mêmes critères d'erreurs relatives utilisés auparavant. Pour cela on désagrège à partir des 32 districts puis on mesure le pourcentage d'erreur relative pour chacune des communes, selon la formule suivante :

$$Pct.Erreur_{relative} = \frac{Decomtpes_{RGPH} - Predictions}{Decomtpes_{RGPH}} * 100$$

Les résultats, présentés dans le tableau 28, montrent que l'interpolation surfacique produit une erreur moyenne nettement inférieure par rapport à l'interpolation volumique. Cela peut sembler contre-intuitif, mais plusieurs facteurs peuvent l'expliquer. Tout d'abord, un relief relativement plat de Madagascar, qui limite l'impact des variations de hauteur dans la répartition démographique. De plus, les données de hauteur sont encore peu précises, du moins peu validées, ce qui peut introduire une incertitude accrue dans l'interpolation volumique. Un autre point

d'attention est que l'intégration de la hauteur peut parfois surévaluer fortement certaines populations, surtout lorsque des bâtiments avec un fort volume sont identifiés mais ne sont pas correctement caractérisés par OSM.

Tableau 28 : Distribution de l'erreur relative selon les deux méthodes de pondération

Interpolation	Erreur moyenne	25 %	50 %	75 %	90 %
Surface	21,2	8,1	16,9	29,4	42,5
Volume	25,9	10,5	21,8	34,1	49,5

Sources : Calculs de l'auteur à partir de l'identification de bâti issue de Google 2.5D

Comme dans les chapitres précédents, la répartition de la population basée uniquement sur la surface des bâtiments donne de nouveau l'erreur moyenne la plus faible, avec 21,2 pour l'interpolation surfacique contre 25,9 pour l'interpolation utilisant le volume des shapes. De plus, la dispersion des erreurs est globalement plus faible à chaque niveau d'analyse. En pratique, l'utilisation de la hauteur des bâtiments semble, pour le moment et à Madagascar, ajouter plus d'incertitude qu'elle n'apporte d'améliorations.

En comparant nos résultats avec ceux obtenus à partir de l'identification du bâti TeleCense dans le chapitre 3, l'utilisation des surfaces issues de Google 2.5D apporte un gain significatif : l'erreur moyenne passe de 23,2 à 21,2, soit une amélioration de 2 points en moyenne. Cette amélioration se traduit également dans la répartition des communes selon leur classe d'erreur relative (Tableau 29) : on observe une augmentation du nombre de communes considérées comme correctement estimées (erreur entre -20 et 20 %) et une forte diminution des cas de sous- ou surestimation extrême. En effet, seules 23 communes présentent une erreur supérieure à ±50 % avec Google 2.5D, contre 48 avec TeleCense.

Tableau 29 : Répartition des erreurs relatives selon la source des données de bâti

Classe d'erreur relative	Google 2.5D (n)	Google 2.5D (%)	TeleCense (n)	TeleCense (%)
Moins de -50 %	16	3.7	33	7.7
de -50 à -20 %	53	12.3	72	16.7
de -20 à 20 %	243	56.5	220	51.2
de 20 à 50 %	111	25.8	90	20.9
Plus de 50 %	7	1.6	15	3.5

Sources : Calculs de l'auteur à partir de l'identification de bâti issue de Google 2.5D et de TeleCense

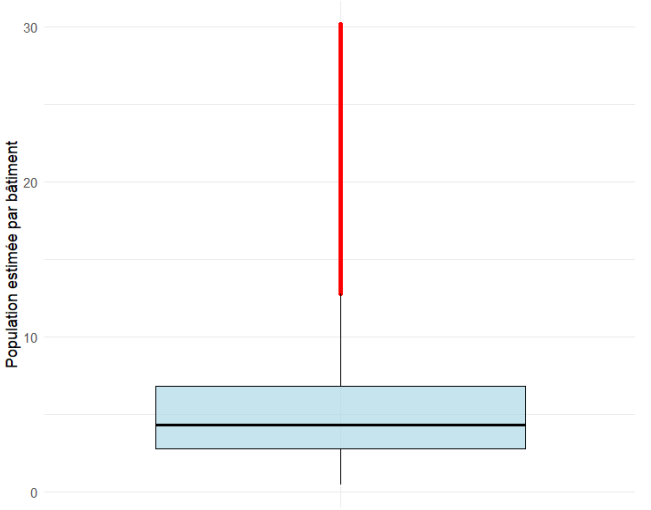
Ces résultats soulignent, d'une part, qu'une répartition fondée uniquement sur la surface des bâtiments semble à nouveau suffisante. D'autre part, on observe une amélioration notable grâce à la détection plus fine du bâti issue de Google 2.5D. En résumé, la désagrégation présente en moyenne moins d'erreurs, ce qui permet une

répartition de la population plus précise. Nous nous interrogeons maintenant sur l'impact de cette amélioration sur l'approche ascendante. Plus précisément, une désagrégation comportant moins d'erreurs en moyenne conduit-elle systématiquement à un modèle ascendant plus adapté ? C'est ce que nous examinons dans la suite de ce chapitre.

3. Développement du modèle ascendant avec les shapes Google 2.5D de Madagascar

Par rapport à la population d'entrainement utilisée dans le chapitre 5 (celle désagrégée à partir des données TeleCense), la nouvelle population désagrégée à partir des données Google 2.5D présente en moyenne des valeurs de population plus faibles, ainsi que des populations maximales nettement réduites, en raison de shapes en moyenne plus petites. Comme le montre la figure 69 ci-dessous, 99 % des shapes sont estimées avec moins de 30 habitants, 90 % avec moins de 11 habitants, et 75 % avec moins de 7 habitants. La population maximale attribuée est de 916 habitants, et moins de 0,1 % des shapes ont une population estimée supérieure à 100 habitants. Dans l'ensemble, puisque chaque shape correspond presque à un bâtiment unique, il semble logique que les valeurs médianes de population reflètent les tailles moyennes des ménages. Ici, nous trouvons une médiane a 4,3 habitants par shape et on rappelle qu'à l'échelle de Madagascar, un ménage est composé en moyenne de 4,1 habitants.

Figure 69 : Répartition de la population estimée par désagrégation, excluant les 1 % des shapes avec la plus grande population.



Sources : Calcul de l'auteur à partir de l'identification de bâti issue de Google 2.5D

3.1 Création du modèle et analyse des coefficients

Avec la population désagrégée, nous construisons un nouveau modèle prédictif ascendant. Comme dans la partie précédente, un échantillonnage aléatoire est tiré en utilisant la même valeur de graine pour garantir la reproductibilité des résultats : 70 % des communes (soit 298 communes et cette fois-ci 859 861 shapes) sont utilisées pour l'entraînement, et 30 % (soit 132 communes et 483 582 shapes) réservées pour la validation. Le modèle prédictif repose sur une régression linéaire dont la formule mathématique est :

$$Y_i = \beta_0 + \beta_1 \log(surface) + \beta_2 \log(distanceToCity) + \beta_3 altitude + \beta_4 densite_{bati_{hectare}} + \varepsilon_i$$

Avec Y_i le logarithme de la population de la shape i et ε_i un terme résiduel aléatoire supposé indépendant et distribué selon une distribution $N(0, \sigma^2)$. Le tableau 30, récapitulatif des résultats, montre un R^2 égal à 0,6459 calculé sur l'ensemble de test, indiquant qu'environ 65 % de la variance de la population désagrégée (log-transformée) est expliquée par les variables du modèle.

Tableau 30 : Résultats du modèle ascendant par modélisation linéaire pour les données 2.5D

Variable	Estimation du coefficient	Coefficient standardisé
Constante = Population désagrégée (au log)	- 1,48***	
Surface (en m^2 et au log)	0,74***	0,83
Distance à la grande ville la plus proche (en mètres)	0,000015***	0,31
Altitude (en mètres)	-0,00012***	-0,12
Densité de bâtiment (par hectare)	0,0018***	0,08
R^2 sur l'ensemble de test	0,6459	

Sources : Calculs réalisés à partir de la cartographie du bâti Google 2.5D

Les coefficients standardisés (3ème colonne) permettent de comparer directement l'importance relative de chaque variable car ils sont exprimés sur la même échelle. Avec un coefficient standardisé de +0,83, la variable de surface est la plus influente, ce qui est logique puisque la désagrégation est fondée sur cette métrique. Vient ensuite la distance à la grande ville, dont le coefficient standardisé de +0,31 indique qu'une augmentation de la distance par rapport à une grande ville est associée à une augmentation modérée de la population estimée.

Comme dans le chapitre précédent, cette variable est plus intuitive à interpréter dans le sens inverse : lorsqu'une shape appartient à une grande ville ou à une

agglomération urbaine (ou qu'elle en est proche), elle tend à avoir en moyenne une population légèrement plus faible qu'une shape située en dehors de l'agglomération. Cela confirme une nouvelle fois que cette variable joue un rôle de proxy rural/urbain, reflétant la tendance générale selon laquelle les ménages urbains comptent en moyenne moins d'habitants que les ménages ruraux.

Pour autant, ce n'est pas parce qu'une shape est rurale qu'elle aura une population très élevée. La surface joue un rôle déterminant en premier lieu, et le coefficient positif de la densité de bâtiments à l'hectare (+0,08) indique qu'une shape isolée aura en moyenne une population légèrement plus faible. À l'inverse, des shapes regroupées dans une zone densément bâtie auront tendance à être un peu plus peuplées. Cela semble cohérent, par exemple, avec le cas d'un groupe d'habitats précaires en périphérie d'une agglomération urbaine. Enfin, l'altitude, avec un coefficient négatif (-0,12), suggère que les shapes situées en altitude tendent à être moins peuplées.

Pour rappel, la variable de densité de bâtiment est bien une variable liée au niveau de la shape, dans le sens où l'on calcule le nombre de bâtiments autour de chaque shape dans un hectare autour d'elle. Par rapport au modèle créé avec TeleCense, cela constitue une avancée, car la variable équivalente utilisée auparavant – la proportion de surface bâtie – était calculée au niveau communal. L'introduction de cette nouvelle variable a pour conséquence une diminution de l'effet explicatif des variables d'intensité de la lumière la nuit, de distance à la route, et de distance au point de santé le plus proche : leurs contributions deviennent négligeables, avec des p-valeurs moins significatives qu'auparavant. Afin d'éviter une redondance des effets dans le modèle, nous avons choisi de les retirer, d'où l'aspect d'une régression linéaire simplifiée par rapport à celui développé dans le chapitre 5.

Cette impression de simplification est renforcée par le fait que nous n'avons pas pu intégrer les variables climatiques (précipitations et température), en raison de contraintes de temps de calcul. Bien que leur effet soit modéré dans la modélisation de la population, ces variables étaient néanmoins pertinentes et mériteraient probablement d'être intégrées si elles sont disponibles.

Puisque les effets captés par la lumière nocturne et les distances semblent désormais capturés par d'autres variables, notamment la distance à la grande ville et la densité du bâti à l'hectare, réduire le nombre de variables permet non seulement de renforcer la robustesse et la stabilité des résultats, mais aussi de favoriser une meilleure généralisabilité du modèle lorsqu'il est appliqué à d'autres contextes.

Pour finir, la variable de hauteur des shapes a également été testé. Bien qu'intuitivement pertinente dans un modèle démographique basé sur les

caractéristiques du bâti, elle n'apporte ici pas d'amélioration significative, en raison de la désagrégation initiale (seulement avec la surface et sans la hauteur prise en compte) déjà effectuée.

3.2 Prédictions et erreurs sur les données de validation

Les prédictions effectuées sur les données de validation permettent de comparer les valeurs prédites aux valeurs observées, en calculant l'erreur comme la différence entre ces deux mesures. Ces mêmes métriques sont ensuite utilisées pour confronter les performances du modèle actuel, basé sur les shapes identifiées à partir des données Google 2.5D, avec celles obtenues dans la partie précédente à partir des Shapes identifiées par TeleCense.

Tableau 31 : Comparaison de la performance des modèles ascendants

Modèle linéaire	R^2	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)
Shapes Google 2.5D	0,73	0,30	0,22
Shapes TeleCense (avec surface Open Buildings)	0,89	0,39	0,28

Sources : Calculs de l'auteur à partir de l'identification de bâti issue de Google 2.5D et de TeleCense

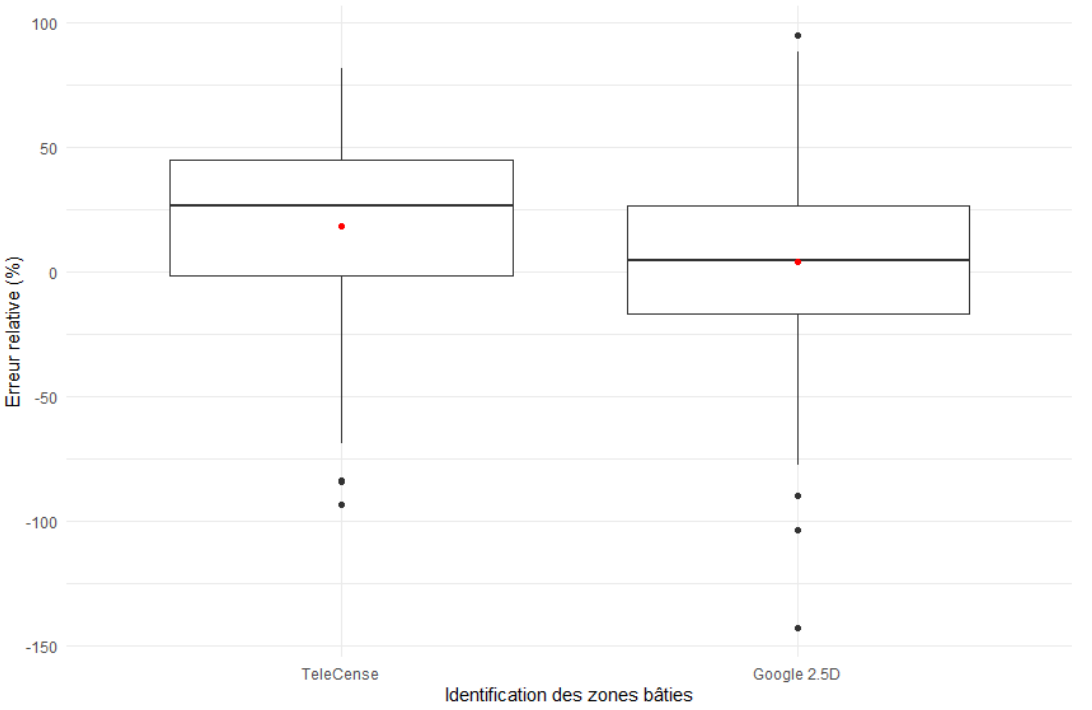
Les résultats révèlent des différences significatives dans les performances des deux approches. Le modèle basé sur les shapes identifiées avec Google 2.5D affiche un RMSE plus faible (0,29) et une erreur absolue moyenne (MAE) également réduite (0,22). Cependant, son coefficient de détermination (R^2) est inférieur à celui du modèle précédent basé sur les shapes de TeleCense, qui atteignait un R^2 de 0,89. Cette différence peut être attribuée aux variations dans l'identification des shapes. Bien que les communes utilisées pour l'entraînement et la validation soient identiques dans les deux modèles, les shapes elles-mêmes diffèrent en nombre et en délimitation (les shapes issues de Google 2.5D sont bien plus nombreuses et aussi plus petites en termes de surfaces, car identifiées avec plus de précision).

Les critères d'erreur comme le RMSE et le MAE sont des indicateurs pertinents, notamment pour évaluer la capacité prédictive future du modèle. En effet, ces performances pourraient suggérer une meilleure robustesse du modèle dans un contexte de prédiction étendu, même au-delà de Madagascar. Cette hypothèse sera directement testée dans la suite de l'analyse, à travers l'identification du bâti des dix communes de la sous-préfecture d'Abidjan. L'échantillon de dix communes reste relativement faible mais il permet néanmoins d'effectuer des comparaisons avec nos résultats initiaux du chapitre 5.

À l'échelle de Madagascar, pour l'échantillon de 30 % de communes retenues pour la validation, nous estimons une population totale de 2 792 892 habitants, contre 3 098 837 habitants recensés dans le RGPH-3, soit un différentiel négatif d'environ 10 %. Au niveau communal, nous avions avec la détection TeleCense une forte tendance à sous-estimer la population, en témoigne la boîte à moustaches de gauche (figure 70), dont la distribution est largement positive. Désormais, avec les shapes issues de Google 2.5D, la distribution des écarts est bien plus centrée autour de zéro (boîte à moustaches de droite) tout en restant globalement comparable à celle obtenue lors de la désagrégation.

On observe néanmoins une dispersion plus marquée des erreurs dans les communes les plus grandes, ce qui indique, d'une part, la possibilité d'erreurs plus importantes pour certaines d'entre elles, et, d'autre part, que le modèle reste perfectible. Afin d'évaluer sa robustesse, nous souhaitons désormais tester ces résultats sur des zones géographiques en dehors de Madagascar.

Figure 70 : Distribution des erreurs selon la source des données de bâti



Sources : Calculs de l'auteur à partir de l'identification de bâti issue de Google 2.5D et de TeleCense

4. Exportation du modèle dans les communes d'Abidjan

Après traitement des données Google 2.5D pour l'année 2021, 538 519 shapes ont été identifiées dans la sous-préfecture d'Abidjan. En appliquant le modèle ascendant, nous estimons la population de ces shapes dans les dix communes, en utilisant les coefficients issus de la modélisation linéaire sur Madagascar et basés sur l'identification Google 2.5D.

L'évaluation des performances du modèle se fait selon deux approches : premièrement nous désagrégeons la population des dix communes dans les shapes identifiées en appliquant la même méthode d'interpolation surfacique. Cette approche permet d'évaluer les performances du modèle et de calculer les métriques d'erreur classiques (RMSE, MAE) en comparant les estimations obtenues par désagrégation avec celles du modèle ascendant.

Deuxièmement**,** nous agrégeons les estimations des shapes au niveau des dix communes, puis comparons ces résultats avec le décompte officiel de 2021 afin d'analyser les écarts.

Le tableau 32 ci après résume tous ces résultats, réalisés avec l'identification Google 2.5D. Le tableau 33 est celui réalisé avec l'identification TeleCense et permet de comparer les résultats effectués avec les deux programmes.

Tableau 32 : Evaluation des performances du modèle ascendant à partir des shapes Google 2.5D dans la sous-préfecture d'Abidjan

Commune	RMSE	MAE	Population RGPH-2021	Population estimée Modèle ascendant	Erreur relative (%)
Abobo	0,14	0,11	1 340 083	1 304 452	2,7
Adjamé	0,35	0,31	340 892	213 831	37.3
Attécoubé	0,22	0,17	313 135	234 709	25
Cocody	0,57	0,57	692 583	1 258 902	-81,8
Koumassi	0,41	0,38	412 282	238 030	42,3
Le Plateau	0,74	0,72	7 186	15 143	-110,7
Marcory	0,21	0,19	214 061	228 954	-7
Port-Bouët	0,20	0,18	618 795	680 425	-10
Treichville	0,28	0,23	106 552	71 641	32.8
Yopougon	0,14	0,10	1 571 065	1 332 842	15.2
Abidjan total	0,325	0,248	5 616 634	5 578 931	0.7

Sources : Calculs de l'auteur à partir de l'identification de bâti issue de Google 2.5D et des décomptes issus de l'INS (2021)

Tableau 33 : Evaluation des performances du modèle ascendant à partir des shapes TeleCense dans la sous-préfecture d'Abidjan

Commune	RMSE	MAE	Population RGPH-2021	Population estimée Modèle ascendant	Erreur relative (%)
Abobo	0,60	0,59	1 340 083	2 440 627	-82,1
Adjamé	0,37	0,37	340 892	492 365	-44,4
Attécoubé	0,50	0,45	313 135	205 927	34,2
Cocody	0,79	0,78	692 583	1 622 613	-134,3
Koumassi	0,55	0,55	412 282	240 001	41,8
Le Plateau	0,46	0,43	7 186	12 583	-75,1
Marcory	0,37	0,36	214 061	308 996	-44,3
Port-Bouët	0,17	0,14	618 795	560 796	9,4
Treichville	0,84	0,84	106 552	46 501	56,4
Yopougon	0,12	0,08	1 571 065	1 644 039	-4,6
Abidjan total	0,55	0,47	5 616 634	7 574 447	-35

Sources : Calculs de l'auteur à partir de l'identification de bâti issue de TeleCense et des décomptes issus de l'INS (2021)

4.1 Une nette amélioration par rapport à l'identification du bâti TeleCense

D'après le tableau 32, les métriques moyennes, avec un RMSE de 0,325 et un MAE de 0,25, montrent une bonne adaptation du modèle au contexte d'Abidjan, en moyenne, et des performances globales comparables à celles obtenues à Madagascar. L'erreur relative globale, particulièrement faible (0,7 %), pourrait à première vue indiquer une excellente précision à l'échelle totale d'Abidjan. Toutefois, cette valeur masque d'importantes disparités entre les communes, révélant une performance non uniforme du modèle.

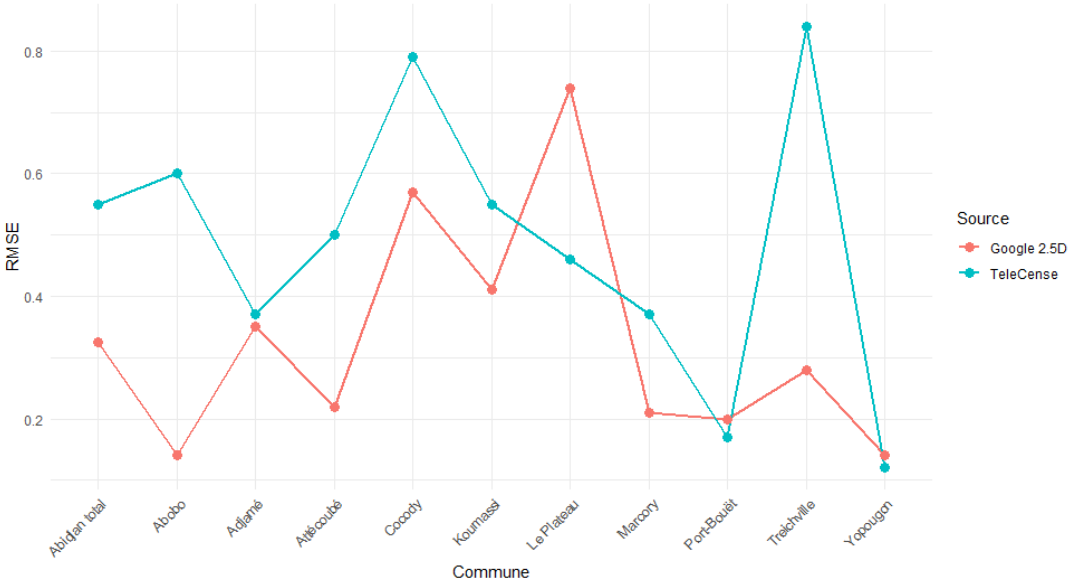
Avant d'examiner ces écarts au niveau communal, il est essentiel de souligner que l'approche ascendante appliquée aux shapes Google 2.5D produit en moyenne de meilleures estimations que celle fondée sur l'identification TeleCense. En effet, toutes les communes hormis Le Plateau, ont une population estimée par le modèle ascendant plus proche (ou très proche pour Yopougon et Port-Bouët par exemple) que celle estimée par le modèle ascendant TeleCense et les métriques d'erreur moyennes sont nettement réduites, avec un RMSE passant de 0,55 à 0,32 et un MAE de 0,47 à 0,25.

L'analyse des résultats à l'échelle communale permet de mieux apprécier les différences entre les deux approches. La figure 71 illustre les variations d'erreur en comparant les RMSE par commune pour les deux méthodes. Hormis Port-Bouët et Yopougon – qui présentent des RMSE légèrement plus élevés mais relativement proches, ce qui suggère des estimations toujours assez précises – ainsi que Le Plateau, dont le RMSE augmente de manière marquée, toutes les autres communes présentent une amélioration, parfois très importante. C'est notamment le cas de la commune

d'Abobo, dont le RMSE est passé de 0,60 à 0,14, ou de celle de Treichville, où il diminue de 0,84 à 0,28.

Néanmoins, un RMSE plus faible ne garantit pas nécessairement une meilleure estimation globale au sein d'une commune. En effet, on rappelle que le RMSE est calculé à partir de la désagrégation initiale effectuée, et non par rapport à des chiffres de population officiels, venant par exemple d'une enquête. Autrement dit, pour calculer cette mesure statistique, nous n'avons pas comparé nos estimations à des valeurs réelles. Cependant, cet indicateur reste utile et permet de comparer les résultats obtenus dans différents contextes.

Figure 71 : Comparaison du RMSE par commune selon la source des données de bâti



Sources : Calcul de l'auteur à partir de l'identification de bâti issue de Google 2.5D et TeleCense

Lorsqu'il est mis en collaboration avec l'erreur relative – qui, elle, est calculée par rapport aux décomptes du RGPH lors de l'agrégation des estimations, c'est-à-dire des valeurs réelles – on peut évaluer la performance de la désagrégation initiale.

Prenons le cas de la commune d'Abobo : avec l'identification TeleCense, nous avions obtenu un RMSE de 0,60, accompagné d'une erreur relative très élevée de -82,1 %, ce qui traduisait une surestimation significative de la population, avec plus d'un million d'habitants en trop. En revanche, après l'application du modèle ascendant basé sur les shapes Google 2.5D, le RMSE diminue considérablement à 0,14, couplé à une estimation désormais remarquable avec seulement 2,7 % d'erreur.

Un RMSE faible indique que le modèle ascendant est capable de prédire avec précision les valeurs observées. Dans le cas d'Abobo cela signifie que le modèle ascendant – créé et entraîné sur les shapes de Madagascar – est capable de prédire avec précision la population désagrégée de la commune d'Abobo.

Une erreur relative faible, de l'ordre de 3 % pour Abobo, montre par ailleurs que la population estimée des shapes agrégée au niveau communal est très proche des décomptes issus du recensement.

Le croisement de ces deux indicateurs – un RMSE faible et une erreur relative faible – suggère que la méthode d'interpolation surfacique utilisée ici était bien adaptée pour estimer la population d'une commune comme Abobo, confirmant la capacité du modèle ascendant à s'exporter en dehors de Madagascar (le pays d'entrainement).

Ce double résultat – RMSE faible et erreur relative également faible – est également observé dans d'autres communes d'Abidjan, notamment à Yopougon, Port-Bouët et Marcory. Cela confirme que l'interpolation initiale par la surface était bien adaptée dans ces contextes, ce qui n'est pas surprenant pour Yopougon et Port-Bouët, en raison de la forte présence de structures bâties horizontales, déjà soulignée au chapitre 5.

Les résultats obtenus dans ces communes renforcent l'idée que l'approche développée dans cette thèse (résumée graphiquement dans l'introduction en figure 7) fonctionne : en partant d'une désagrégation de la population dans les zones bâties, il est possible de construire un modèle ascendant capable de prédire la population dans des zones extérieures à la zone d'entraînement.

4.2 Des problèmes d'estimations qui persistent

Pour autant, plusieurs communes présentent toujours des sous- ou surestimations parfois importantes. Le cas de Treichville est particulièrement intéressant : si le RMSE a fortement diminué – passant de 0,84 avec TeleCense à 0,28 avec le modèle ascendant –, l'erreur relative reste élevée, autour de 32,8 %. Cela signifie que, malgré une meilleure capacité du modèle à reproduire la distribution interne des valeurs (celle issue de la désagrégation), la population totale reste largement sous-estimée.

Ce décalage suggère que des informations clés manquent encore au modèle ascendant pour améliorer ses performances dans ce type de contexte. Ces informations pourraient être intégrées à différents niveaux : soit dès la phase de désagrégation, en

affinant les hypothèses de répartition, soit directement dans l'entraînement du modèle ascendant, en y ajoutant des variables explicatives supplémentaires.

Ce constat général se vérifie encore davantage dans deux communes où les résultats du modèle ascendant restent très éloignés des décomptes observés : Le Plateau et Cocody, qui affichent des valeurs de RMSE et d'erreur relative particulièrement importants.

Concernant Le Plateau, la situation peut être relativisée : bien que l'erreur dépasse 100 %, cela ne représente qu'un surplus de 8 000 habitants, ce qui reste négligeable dans le contexte global. Avec ses grands immeubles et la difficulté de caractériser l'ensemble du bâti résidentiel à partir des données OSM, cette zone reste particulière. Par ailleurs, c'est sans doute une zone où l'intégration des données de hauteur des bâtiments, avec une méthodologie appropriée, offrirait probablement une meilleure estimation. En revanche, Cocody constitue et demeure un réel problème d'estimation. Bien que l'erreur soit significativement moindre qu'avec le modèle précédent – où la population était surestimée de près d'un million d'habitants –, elle reste très importante avec une surestimation de 570 000 habitants.

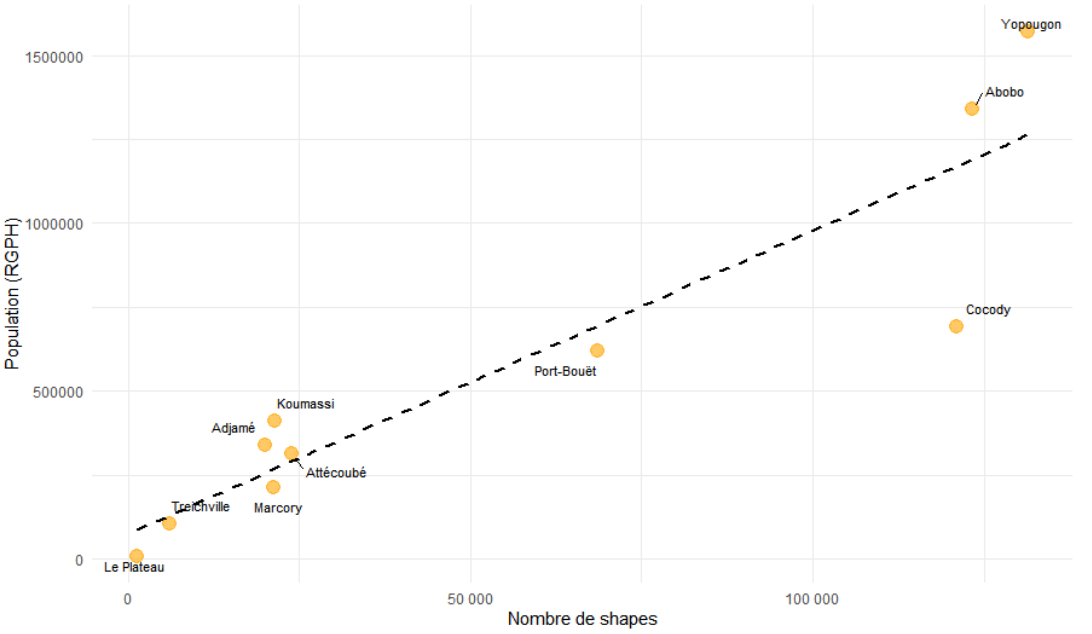
La figure 72 qui montre une relation linéaire entre le nombre de shapes par communes et les décomptes de population issus du RGPH permet d'apporter quelques éléments de compréhension. Si l'on se concentre sur les trois plus grandes communes d'Abidjan – Abobo, Yopougon et Cocody – on observe qu'elles présentent des surfaces comparables (130, 149 et 122 km² respectivement) ainsi qu'un nombre de shapes similaire, aux alentours de 125 000 (en haut à droite de la figure 72). Pourtant, Cocody, malgré ces caractéristiques similaires, se distingue par une population bien plus faible et une estimation fortement erronée dans le modèle ascendant.

La figure suggère ainsi que le nombre de zones bâties dans Cocody serait « trop » élevé au regard de sa population totale. Plusieurs explications peuvent être avancées : tout d'abord, comme évoqué dans le chapitre 5, Cocody se caractérise par la taille moyenne des ménages la plus faible de la sous-préfecture (4,1 habitants par ménage), ainsi que par un statut socio-économique relativement élevé, pouvant favoriser la présence d'un plus grand nombre de logements secondaires. Ces éléments permettent d'expliquer une partie de l'écart observé mais l'ampleur de l'erreur semble trop importante pour qu'ils en soient les seules causes. Deux hypothèses supplémentaires peuvent être formulées : soit une sous-estimation de la population par le RGPH, ce que la figure 72 semble suggérer, soit une surestimation du nombre de zones bâties par le programme Google 2.5D. La première hypothèse pourrait être étayée par les rapports de post-recensement – qui abordent généralement ce type de

biais de décompte – lorsqu'ils seront disponibles. Quant à la seconde, elle reste difficile à vérifier à ce stade, c'est pourquoi elle ne sera pas davantage explorée dans ce chapitre.

On peut toutefois noter qu'en décidant de répartir uniformément la population dans chaque shape, Cocody n'afficherait en moyenne que 5,7 habitants par bâti, contre 10,8 à Abobo et 11,9 à Yopougon. Ces écarts sont d'autant plus notables que la surface moyenne des shapes est relativement similaire dans les trois communes (196 m² à Cocody et Abobo, 183 m² à Yopougon). Ce constat renforce l'idée que des caractéristiques sont manquantes à notre modélisation de la population des shapes.

Figure 72 : Relation entre le nombre de shapes par communes et les décomptes de population du RGPH 2021 pour la sous-préfecture d'Abidjan



Sources : Cartographie du bâti issue du programme Google 2.5D et des décomptes issus de l'INS (2021)

À l'inverse, un phénomène comparable peut être observé pour les communes de Koumassi et Adjamé, toutes deux sous-estimées par le modèle ascendant. Dans la figure 72, elles apparaissent au-dessus de la ligne de régression, ce qui suggère que le nombre de zones bâties détectées y est légèrement insuffisant pour rendre compte de la population observée.

5. Bilan et perspectives finales

Si l'on fait le bilan, le modèle fournit désormais des estimations correctes pour quatre communes, sous-estime la population dans quatre autres, et en surestime fortement deux – dont Cocody. À titre de comparaison, avec le modèle issu de l'identification TeleCense, seules deux communes étaient correctement estimées, tandis que toutes les autres présentaient des écarts importants, en particulier avec un nombre plus élevé de cas de fortes sur- ou sous-estimations. Avec le nouveau modèle ascendant, sept communes sur dix affichent une amélioration des performances, deux autres voient une légère dégradation mais restent malgré tout bien estimées (Yopougon et Port-Bouët), et enfin Le Plateau reste un cas à part, difficile à modéliser.

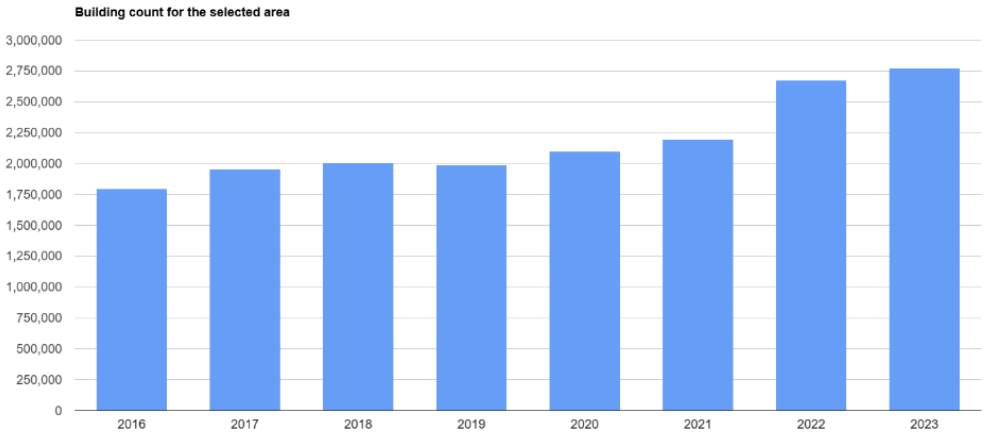
Sans omettre qu'il puisse toujours exister des erreurs de détection, on peut conclure que le passage à la détection du bâti par Google 2.5D – qui permet de d'homogénéiser la surface des shapes – constitue une avancée plus que significative qui a permis une nette amélioration des estimations.

Comment expliquer que le nouveau modèle montre lui-aussi des performances limitées dans certaines communes, et notamment une surestimation très marquée à Cocody ? Est-ce dû à une mauvaise identification du bâti de Google 2.5D? A priori ce n'est pas le cas.

Une exploration rapide du cas du Bénin, pour lequel nous avions noté des irrégularités dans la détection du bâti de TeleCense (très grandes shapes, décroissance de la surface totale bâtie détectée entre 2018 et 2019, puis entre 2020 et 2021) nous confirme cela. Sans analyser en profondeur les données google 2.5D pour ce pays, nous avons utilisé l'application « Earth Engine Dataset Explorer »[65] pour compter, année par année de 2016 à 2023, le nombre de bâtiments détectés par Google 2.5D. La figure 73, directement issue et créée par cette application, montre une progression presque linéaire du nombre de zones bâties, à l'exception de l'année 2019, où un léger recul est observé par rapport à 2018 (équivalent à ce qu'on avait trouvé avec TeleCense mais avec une diminution moindre de la part de la détection Google 2.5D).

[65] https://mmeka-ee.projects.earthengine.app/view/open-buildings-temporal-dataset

Figure 73 :
Évolution du nombre de bâtiments identifiés par Google 2.5D au Bénin de 2016 à 2023



Sources : Calcul effectué à partir de l'application Earth Engine Dataset Explorer et à partir des données des données d'identification de bâti de Google 2.5D

Au vu de l'uniformité de la détection des surfaces des bâtiments par Google 2.5D, on peut supposer qu'une augmentation du nombre des bâtiments implique une augmentation de la surface bâtie totale. Ainsi, Google 2.5D, en plus de résoudre le problème d'hétérogénéité dans la surface des shapes, ouvre des perspectives nouvelles pour des analyses longitudinales précises, qui étaient jusqu'ici difficiles à mettre en œuvre.

Alors, est-ce plutôt un problème lié à notre plan de recherche général ? On a montré que l'approche globale pouvait fonctionner, notamment pour les communes de Marcory, Port-Bouët, Yopougon et Abobo. Autrement dit, on a montré que construire un modèle ascendant à partir de la désagrégation fondée sur les régions retenues à Madagascar était une approche pertinente et viable, en témoignent les écarts observés entre nos estimations et les décomptes officiels qui sont relativement faibles.

Néanmoins, on peut effectivement supposer que les données d'entrainement ne sont pas idéales. D'abord la désagrégation par la seule variable de surface ne permet évidemment pas de refléter toutes les caractéristiques liées à la population des zones bâties. De plus, sans doute que toute la diversité des zones bâties d'Abidjan n'a pas été capturée avec les six régions retenues à Madagascar. On ne peut pas s'étonner que le modèle ascendant ne soit pas en capacité d'estimer la population de zones bâties sur lesquelles le modèle ne s'est pas entrainé, comme par exemple des quartiers avec des bâtiments hauts.

Dès lors, aurait-il fallu, ou faudrait-il à l'avenir, entraîner le modèle ascendant sur des communes présentant des caractéristiques similaires à celles de Cocody ? La réponse est à la fois non… et oui.

Non, dans le cadre actuel. Compte tenu de la méthode de désagrégation actuelle, cela n'aurait sans doute pas permis d'améliorer significativement les résultats. Prenons l'exemple de Cocody : lorsque nous faisons le test de désagréger la population à partir du niveau de la sous-préfecture d'Abidjan – soit 5 616 634 habitants répartis entre les shapes des dix communes –, on observe déjà une forte surreprésentation de la population à Cocody. Cela signifie que les données issues de la désagrégation sont biaisées dès le départ. Entraîner un modèle sur ces données reviendrait donc à apprendre à reproduire une estimation initialement erronée.

Autrement dit, même si Cocody avait été incluse comme zone d'entraînement, le modèle ascendant aurait appris à estimer une population mal répartie entre les shapes. Il est vrai que, lors de l'entraînement du modèle ascendant, on utilise les données de désagrégation qui s'appuient sur le niveau administratif le plus fin disponible – ici, celui de la commune –, ce qui fait que l'erreur relative au niveau communal serait mécaniquement nulle. Pour autant, si la désagrégation à partir d'un niveau plus grossier (sous-préfecture d'Abidjan ici) génère déjà une forte surestimation intra-communale, on peut raisonnablement supposer que la distribution au niveau des shapes reste problématique, même avec un niveau administratif plus fin.

Ainsi, le modèle ascendant se trouve entraîné sur une répartition spatiale peu représentative de la réalité. Il ne dispose donc pas des informations nécessaires pour ajuster correctement ses prédictions, et risque au contraire de reproduire la même erreur que lorsque l'on désagrège avec un niveau plus grossier.

Nous sommes donc confrontés à un problème fondamental dans nos données d'entraînement fournies au modèle ascendant. Pour autant, ce problème n'est pas insurmontable et il peut être corrigé, ou du moins atténué, selon plusieurs hypothèses et pistes de réflexion que l'on envisage au fil des différentes étapes présentées dans cette thèse. Comme il n'est pas possible d'espérer estimer la population de n'importe quelle zone géographique d'Afrique à partir d'un modèle ascendant basé sur des données d'entrainement inadéquates, nous commençons par donner des pistes concernant la désagrégation.

5.1 Perspectives concernant la désagrégation

Les différents chapitres de cette thèse auront montré que, bien qu'imparfaite, la désagrégation fondée uniquement sur la surface bâtie était la méthode la plus efficace. Dans tous les essais menés, elle s'est révélée systématiquement plus performante que les autres approches testées, offrant une erreur relative plus faible et

par conséquent une meilleure précision pour répartir la population dans une large partie des communes.

Cependant, cette méthode présente la limite évidente qu'elle repose exclusivement sur la variable de surface des shapes, sans tenir compte d'autres facteurs importants qui pourraient influencent la répartition de la population. Autrement dit, même si la surface bâtie reste une variable primordiale elle ne suffit pas, à elle seule, à capter toute la complexité des dynamiques de peuplement.

Il serait donc pertinent d'enrichir cette approche en y intégrant progressivement d'autres variables explicatives. Dans le cadre de la désagrégation, la seule manière de le faire est de chercher des relations statistiques entre population (ou densité de population) et caractéristiques spatiales, au niveau administratif, puis de supposer que ces relations peuvent s'appliquer à une échelle plus fine — celle des shapes. C'est précisément le principe qui sous-tend la méthode popularisée par WorldPop, et qui a montré de très bonnes performances pour la visualisation carroyée avec des cellules de 100 × 100 mètres (Stevens et al., 2015). Toutefois, dans le cadre de cette thèse dont la volonté est de distribuer la population directement dans les zones bâties et non dans des cellules carrées, cette approche réalisée dans le chapitre 3 n'a jamais surpassé les résultats obtenus par la simple interpolation surfacique.

Cela ne signifie pas qu'il faut abandonner l'idée de construire de telles relations, mais plutôt chercher à les améliorer, en particulier en exploitant les nouvelles variables rendues disponibles par Google 2.5D. La variable de hauteur, bien que sans doute peu pertinente calculée à une échelle agrégée comme celle de la commune, mériterait d'être testée. Surtout, la variable de densité du bâti autour de chaque shape offre des perspectives intéressantes. Le fait de pouvoir compter précisément le nombre de bâtiments autour de chaque shape nous a permis de calculer une variable de densité de bâti à l'hectare qui pourrait constituer une variable explicative pertinente dans une modélisation de la densité de population à l'échelle communale. Ce type de relation pourrait être formulé de manière simple, par exemple sous la forme suivante :

$$Y_i = \beta_0 + \beta_1 * densite_{bati_{hectare},i} + \varepsilon_i$$

Où Y_i représente la densité de population de la commune i, et la densité de bâti par hectare autour des shapes serait utilisée comme variable explicative, seule ou en combinaison avec d'autres facteurs.

Par exemple, à Abidjan, les deux communes présentant les plus faibles densités de bâti moyenne à hectare, Le Plateau et Cocody (respectivement 41 bâtiments à l'hectare pour Cocody, contre 50 et 57 pour Abobo et Yopougon), illustrent une corrélation possible entre une densité plus faible et une population moindre. Couplée à d'autres variables – climatiques (température et précipitations), topographiques (altitude), ou encore d'intensité de la lumière la nuit, voire de hauteur (avec la moyenne des bâtiments de la zone administrative) – cette approche pourrait permettre une répartition bien plus logique et, potentiellement, plus précise.

5.2 Perspectives concernant l'approche ascendante

L'approche ascendante étant étroitement liée à l'approche descendante, les perspectives spécifiques qui lui sont associées restent, à ce stade, relativement limitées. Néanmoins, un enjeu central réside dans le choix de la méthode de modélisation la plus appropriée pour relier les caractéristiques du bâti à la population observée.

Plusieurs familles de méthodes pourront être envisagées : des modèles linéaires classiques, des techniques issues de l'apprentissage automatique (comme les forêts aléatoires ou les réseaux de neurones), ou encore des approches probabilistes, notamment bayésiennes. Le choix final devra reposer sur la capacité de la méthode retenue à produire des estimations précises tout en étant adaptable à différents contextes.

5.3 Perspectives concernant l'approche globale

Jusqu'à présent, notre modèle ascendant a été construit à partir de la population désagrégée de 70 % des communes de Madagascar, puis évalué sur les 30 % restants, avant d'être testé sur les dix communes d'Abidjan. Cette approche a permis une première validation dans un contexte extérieur à Madagascar, mais le périmètre reste limité. Pour aller plus loin, il nous faut mobiliser les différentes perspectives citées précédemment dans un projet encore plus ambitieux en mobilisant un volume de données plus important, à la fois pour améliorer l'apprentissage des modèles et pour les tester dans des environnements variés.

L'un des autres principaux apports des données Google 2.5D réside dans leur accessibilité et leur couverture spatiale étendue. En dehors de la contrainte de stockage, qui reste importante, ces données sont relativement simples à télécharger et sont donc mobilisables pour de nombreux pays, sur plusieurs années. L'élargissement du champ d'étude à d'autres pays devient aujourd'hui non seulement envisageable, mais également réaliste.

Dans ce cadre, l'objectif serait de construire un modèle plus global en agrégeant les données issues de plusieurs pays ayant récemment réalisé un recensement détaillé. Plusieurs pays pourraient d'ores et déjà être envisagés : la Côte d'Ivoire, au-delà de la seule zone d'Abidjan, mais aussi le Burkina Faso (recensement de 2020), le Ghana (2021), la Tanzanie et la Zambie (2022), ou encore le Sénégal (2023).

Il faudrait néanmoins pour cela réaliser un travail important de récupération et d'homogénéisation des fichiers géographiques des zones administratives de ces pays, en espérant qu'ils soient suffisamment à jour et compatibles avec les résultats des derniers recensements. Comme cela a été montré dans le chapitre 3 pour Madagascar, cette tâche n'est pas toujours triviale car les découpages administratifs ne sont pas actualisés de manière systématique rendant les correspondances avec les données du dernier recensement pas toujours assurées.

Une telle généralisation constituerait un prolongement naturel du travail engagé dans cette thèse. Elle permettrait de concevoir un modèle de désagrégation multi-pays, mobilisant à la fois des variables classiques (surface bâtie et variables issues de la télédétection) et des données plus récentes, comme la hauteur ou la densité de bâtiments autour de chaque shape. Alors, l'objectif de produire un modèle généralisable capable d'estimer de manière fiable la répartition de la population dans des contextes variés, même en l'absence de données détaillées, serait sans doute encore plus réaliste.

Une meilleure prise en compte de la hauteur et plus généralement, de la caractérisation des bâtiments, permettrait certainement d'obtenir de meilleurs résultats. Jusqu'à présent, l'intégration de la hauteur des bâtiments dans nos analyses a généré plus d'incertitudes que d'améliorations, comme cela a été constaté lors de son intégration pour la désagrégation à Madagascar. Toutefois, cette variable représente une avancée prometteuse qu'il serait pertinent de tester et d'intégrer dans les modèles futurs. Un premier constat intéressant est que, pour l'ensemble des shapes d'Abidjan, la moyenne et la médiane de la hauteur des bâtiments sont supérieures de 0,80 mètre à celles des six régions étudiées à Madagascar. Cette différence, loin d'être négligeable, pourrait indiquer un rôle différencié à explorer dans des contextes extérieurs à Madagascar.

Conjointement à l'utilisation de la hauteur, il faudra aussi améliorer notre capacité à distinguer tous les bâtiments résidentiels des autres types de structures,

notamment des entrepôts ou centres commerciaux par exemple. Cette limitation augmente le risque d'erreur lorsque l'on utilise le volume des bâtiments comme variable, comparé à une analyse uniquement basée sur leur surface. Une mauvaise caractérisation des bâtiments pourrait ainsi fortement biaiser les estimations. Ce point soulève donc la question de comment mieux caractériser les bâtiments que l'on détecte et de si l'on peut les qualifier de non-résidentiel sans l'utilisation d'OSM ?

Deux approches peuvent être envisagées afin de mieux caractériser le bâti. Tout d'abord, une approche centrée sur la technique d'analyse d'images pourrait aider à distinguer le bâti résidentiel. Par exemple, l'analyse de la texture des toits des bâtiments pourrait permettre de mieux distinguer différents types de structures. Une telle méthode offrirait des perspectives intéressantes pour les modèles démographiques.

Une autre voie réside dans l'utilisation de données statistiques pour créer des classes de bâtiments. En combinant plusieurs variables, il serait possible d'établir des typologies de bâti. Une fois ces classes définies, on pourrait les associer à des distributions de population pour affiner la répartition, comme cela a été fait dans des travaux précédents, comme par exemple les travaux d'Afripop qui au début des années 2010 associait des densités de population à des types de couverture du sol (Linard et al., 2012) ou encore Hallot, Grippa, Stephenne, Wolff (2019) qui proposaient des densités de population liées à des classes basées sur le taux d'imperméabilisation du sol. Ces approches permettent de pondérer la répartition de la population selon les classes de bâti. Dans notre étude, une zone d'habitat précaire pourrait être caractérisée par une taille de bâtiment relativement faible (valeur de la variable de hauteur faible donc) et une forte densité de bâti à l'hectare (valeur de la variable de densité de bâti forte donc), et sans doute avoir une population en moyenne plus nombreuse. A l'inverse, un bâtiment identifié comme une villa aurait probablement moins d'habitants.

6. Conclusion

Ce dernier chapitre a été consacré à l'exploration d'une nouvelle itération de l'approche globale développée dans cette thèse. Cette fois, l'analyse a été conduite à partir d'un nouveau jeu de données de bâti, produit par la détection Google 2.5D. Cette base de données se distingue par une identification plus fine et plus homogène du bâti, avec une détection quasi bâtiment par bâtiment.

La détection Google 2.5D présente plusieurs avantages. Tout d'abord, elle réduit le temps de travail de traitement préalable par rapport aux données issues du projet TeleCense. En effet, grâce à leur surface réduite, il n'est dans la majorité des cas plus nécessaire de redécouper les shapes pour qu'elles n'appartiennent bien qu'à une seule zone administrative. De plus, il n'est plus nécessaire de recalculer la variable de surface en fonction de sources auxiliaires comme c'était le cas avec l'intégration des données provenant du premier programme Google Open Buildings.

Surtout, Google 2.5D offre de nouvelles variables descriptives potentiellement précieuses pour l'estimation de la population : l'estimation de la hauteur des bâtiments et la localisation précise de chaque structure, via l'identification de leur centroïde. Cette dernière permet de calculer des indicateurs tels que la densité de bâti à l'hectare autour de chaque shape, qui s'est révélée prometteuse pour modéliser la densité de population.

Enfin, ces données sont disponibles en accès libre pour plusieurs années, de 2016 à 2023, ce qui ouvre la voie à des analyses longitudinales basées sur des images satellites à haute résolution. Il convient néanmoins de rester prudent car à ce jour, aucune mise à jour n'a encore été réalisée pour l'année 2024, et la dépendance à un programme tiers tel que Google constitue une incertitude. En cas d'arrêt du service, la reproductibilité et la continuité des études en seraient clairement affectées.

Les résultats ont montré que, pour chacune des étapes de l'approche générale, l'utilisation des données Google 2.5D a permis d'améliorer la qualité des estimations par rapport à l'identification du bâti réalisée dans le cadre du projet TeleCense. Bien que l'approche par interpolation surfacique reste, dans notre cas, la méthode de désagrégation la plus performante, les erreurs moyennes ont diminué de deux points avec les nouvelles données. Cette amélioration se répercute également sur les résultats du modèle ascendant : la majorité des communes d'Abidjan dont la population des shapes a été estimée avec ce nouveau modèle, affichent désormais des erreurs moindres et une estimation de la population plus proche de celle décomptée par le RGPH.

On note en particulier l'amélioration importante obtenue pour la commune d'Abobo. Alors que l'estimation précédente surestimait la population de plus d'un million d'habitants, l'erreur n'est plus aujourd'hui que de 36 000 habitants, soit seulement 2,7 % d'erreur relative. Ces résultats confirment le potentiel des données Google 2.5D et l'apport d'une identification du bâti plus fine et homogène pour l'estimation de la population des zones bâties.

Malgré ces avancées, certaines communes restent difficiles à modéliser avec précision. C'est le cas notamment de la commune de Cocody où la complexité de la structure résidentielle, avec probablement une part importante de logements vacants ou secondaires, rend la modélisation de la population plus incertaine.

Les perspectives d'amélioration sont nombreuses, en particulier grâce à l'introduction de ces nouvelles données prometteuses. Le travail à venir consistera désormais à mettre en œuvre ces développements, dans l'objectif de produire un modèle capable d'estimer la population à partir d'images satellites, sans dépendre directement de données de recensement en entrée.

Conclusion générale

Est-ce qu'on peut compter les habitants à partir d'images prises depuis le ciel ? Est-il possible, en s'appuyant principalement sur des images à une résolution qui ne permet pas de voir les individus mais uniquement les groupements de bâtiments, d'obtenir une estimation fiable du nombre d'habitants dans un espace donné ? C'est cette interrogation, formulée de manière très directe et intuitive, qui a guidé le début de ce travail de recherche.

Au fil du temps, à travers lectures, premières expérimentations et rencontres avec d'autres chercheurs, ces premières intuitions ont été affinées, précisées et encadrées pour donner lieu à une démarche scientifique cohérente. Nous nous sommes notamment rendu compte que, malgré des progrès significatifs et importants dans la cartographie des bâtiments à partir d'imagerie satellitaire, les programmes de cartographie de population n'estimaient pas encore la population directement à l'échelle des zones bâties détectées. C'est l'objectif de cette thèse. A partir de cette ambition, le travail de thèse s'est structuré et a été pensé comme un tout, avec une progression logique où chaque étape a été construite en lien avec la précédente dans le but d'apporter une méthode pouvant produire des résultats suffisamment précis pour qu'ils soient utiles et mobilisables à différentes échelles, dans différents contextes et à différentes dates, dans les pays du continent africain où les données démographiques sont plus rares qu'ailleurs et les transformations actuelles parfois très rapides.

Pour cela, nous avons développé une démarche articulant approches de cartographie de population descendante et ascendante dont l'ambition finale est d'estimer la population sans données de recensement au préalable. Cette méthode s'appuie sur deux programmes d'identification du bâti différents (TeleCense et Google 2.5D) et sur les décomptes de recensement de quatre pays : Madagascar, le Bénin, la France (via le département de Mayotte) et la Côte d'Ivoire.

Le programme TeleCense, entièrement produit à partir de la combinaison des données Sentinel-1 et Sentinel-2 du programme européen Copernicus, en accès libre, a servi de base au cœur du travail de thèse (chapitres 2 à 5). La cartographie des zones bâties des régions de Madagascar en 2018, année de son dernier recensement (RGPH-3), nous a permis d'ajuster les données de recensement à l'échelle de ces zones bâties (nos « shapes »). La surface des shapes est une première fois ajustée à l'aide de la détection Google Open Buildings, ce qui permet d'affiner la surface des grandes shapes et de corriger d'éventuelles zones non bâties qui auraient pu être intégrées lors de leur détection initiale. Elle est ajustée de nouveau à l'aide de la détection

OpenStreetMap afin de ne pas inclure de bâtiments considérés comme non-résidentiels dans les shapes identifiées. Ensuite, plusieurs caractéristiques géo-climatiques ont été associées aux niveaux des shapes et des communes. Il s'agit de variables climatiques (précipitations, température), environnementales (altitude), de distance à différentes entités (routes, points d'eau, centres de santé, grandes villes), d'intensité lumineuse nocturne, ainsi que de variables liées au bâti (surface et densité).

Nous avons montré que, malgré l'existence de relations significatives entre certaines de ces variables et la densité de population à l'échelle communale, ces relations ne s'avéraient pas aussi pertinentes lors de l'exercice de désagrégation de la population des communes vers les shapes. Nous avons même observé que la méthode la plus simple à mettre en œuvre – une interpolation surfacique répartissant la population uniquement en fonction de la surface des zones bâties détectées au sein d'une même commune – était celle qui produisait les résultats les plus précis (ou ceux avec les erreurs les plus faibles).

Nous avons également mis en évidence, à travers la comparaison avec les estimations carroyées du programme WorldPop, que les résultats obtenus à Madagascar étaient plus précis que les données d'estimation de population disponibles en libre accès pour l'année 2018 dans ces régions (chapitre 3). En moyenne, le programme TeleCense entraîne non seulement moins d'erreurs globales, mais aussi une réduction des écarts les plus importants. Là où WorldPop a progressivement intégré des contraintes liées à la présence du bâti dans ses modèles (estimations contraintes), notre méthode repose dès le départ sur une délimitation des espaces construits qui permet une répartition de la population plus logique et renforce la qualité des estimations produites.

Ces résultats confirment notre hypothèse initiale : à partir du moment où l'on dispose de données de population à un niveau suffisamment détaillé – comme celui des communes à Madagascar – il est non seulement possible d'estimer la population au niveau des surfaces bâties détectées mais c'est aussi plus efficace que de le faire dans un carroyage. Une méthode simple, reposant uniquement sur ces surfaces, produit des estimations plus cohérentes que des approches plus complexes intégrant de nombreuses variables explicatives. Autrement dit, dans ce contexte, la qualité de la délimitation des espaces construits suffit à elle seule pour assurer une désagrégation fiable de la population, ce qui renforce l'intérêt de développer des approches centrées sur le bâti comme unité principale d'analyse.

Cependant, le cas d'application sur le Bénin, et plus précisément dans la zone du Grand Nokoué, a montré que l'approche descendante n'était plus adaptée dès lors que l'on s'éloigne trop de la date de référence du recensement (chapitre 4). Le programme TeleCense rencontre en effet des difficultés à maintenir une cohérence

temporelle dans la détection de l'évolution du bâti, notamment en raison de données auxiliaires incomplètes sur la délimitation des routes pour certaines années, issues d'OpenStreetMap. Par ailleurs, les incertitudes inhérentes aux projections démographiques s'ajoutent à celles liées à la détection du bâti. Ensemble, ces deux sources d'imprécision rendent les désagrégations de moins en moins fiables au fil du temps : les écarts entre les estimations désagrégées et les projections communales s'accentuent, en particulier dans certaines communes où apparaissent des erreurs importantes.

Cela nous a permis de faire le constat de la nécessité d'une autre approche, notamment pour les régions ou pays ne disposant pas de données de population récentes. Nous avons donc développé une approche ascendante, en construisant un modèle linéaire à partir de la population désagrégée des communes dans les shapes de Madagascar, afin de suivre l'évolution de la population de manière précise même en dehors des périodes de recensement (chapitre 5). En dépit d'une prédiction de la population reposant fortement sur la surface des shapes, d'autres caractéristiques comme la distance à la grande ville la plus proche, l'altitude et la densité de bâti communale jouent alors un rôle dans la modélisation.

Ce modèle ascendant a montré une très bonne capacité d'adaptation lorsqu'il a été appliqué au département de Mayotte, avec une erreur de seulement 3 % à l'échelle départementale, ainsi qu'à certaines communes de la sous-préfecture d'Abidjan, notamment celle de Yopougon, où la population totale a été prédite avec seulement 5 % d'erreur. L'application à Mayotte a permis de confirmer que le modèle ascendant avait beaucoup moins de difficultés à estimer la population des shapes de petites et moyennes surfaces. En revanche, les shapes de grande surface, qui, du fait de leur grande taille peuvent englober une variété importante de bâtiments, y compris non-résidentiels, posent un problème de modélisation qui complique l'estimation. Cela a été confirmé à Abidjan, où deux communes en particulier, Abobo et Cocody, ont présenté de nombreuses shapes surestimées, ce qui a conduit à une erreur de surestimation importante.

Bien que ces résultats montrent des disparités aux échelles les plus fines et qu'il est encore compliqué d'estimer avec précision la population de toutes les shapes, le fait d'avoir construit un modèle uniquement fondé sur les caractéristiques des zones bâties, capable d'estimer correctement la population totale d'un territoire sur lequel il n'a pas été entraîné, constitue déjà un résultat important en soi. Plus concrètement, ce travail a montré que le modèle ascendant, entrainé sur l'identification TeleCense à Madagascar, est capable de produire des estimations précises à l'échelle des zones administratives pour des zones géographiques au contexte de bâti similaire à celui de Madagascar, caractérisées par de nombreux bâtiments horizontaux et peu de hauteur.

Ce premier travail exploitant images à relative faible résolution montre que l'approche globale développée dans cette thèse fonctionne et que nous sommes sur la bonne voie.

Réitérer l'approche globale à partir de l'identification des zones bâties issues du programme Google 2.5D – qui combine des images Sentinel et des images à très haute résolution (de 0,5 à 1 mètre), pour une résolution effective de 4 mètres – a permis d'améliorer les résultats à toutes les étapes (chapitre 6). Tout d'abord, la distribution de la population désagrégée comporte avec ces données des estimations beaucoup plus faibles – principalement du fait que les shapes détectées sont souvent de l'ordre du bâtiment unique – et est très proche en valeur médiane du nombre d'habitants moyen par ménage à Madagascar. Cette nouvelle répartition contribue à une diminution notable de l'erreur moyenne calculée au niveau communal, confirmant ainsi des améliorations par rapport au bâti issu de TeleCense. Ces gains se répercutent également sur le modèle ascendant : appliqué à Abidjan, ce nouveau modèle offre des estimations bien plus précises dans presque toutes les communes. L'exemple d'Abobo est encore une fois particulièrement marquant : alors que la population était surestimée de plus d'un million d'habitants, l'estimation est désormais très proche des chiffres officiels du dernier recensement de 2021.

Ces résultats confirment que plus nous améliorons la détection du bâti, plus les estimations de population sont fiables. Nous l'avions déjà observé dans le chapitre 3, lorsque nous avons observé que la désagrégation était plus précise avec la surface pondérée par la détection du programme Google Open Buildings qu'avec celle de TeleCense ; aujourd'hui, avec l'utilisation de la toute dernière version du programme Google 2.5D, qui a l'avantage supplémentaire d'être précisément datée, les performances des modèles sont encore renforcées.

Désormais, et avec l'ensemble des éléments ici présentés, nous pouvons revenir à la question centrale qui a guidé cette thèse : est-il possible d'estimer la population d'un pays sans recensement récent, à partir des caractéristiques du bâti détecté par satellite ? Pour reprendre le cas de la République démocratique du Congo, évoqué en introduction et dont le dernier recensement général de la population remonte à 1984, plusieurs indices nous laissent penser que nous sommes désormais en capacité d'estimer la population nationale avec un bon niveau de précision. Avec la première version du modèle, reposant sur les shapes issues du programme TeleCense, nous avions déjà obtenu une erreur de seulement 3 % pour Mayotte en 2017. Avec la seconde itération, utilisant les données du bâti Google 2.5D, l'estimation de la population d'Abidjan atteint un niveau de précision remarquable, avec moins de 1 % d'écart par rapport aux chiffres officiels du recensement de 2021 (5 578 931 habitants estimés contre 5 616 634 recensés). Bien qu'un échantillon de validation plus large serait nécessaire pour généraliser pleinement ces résultats, ils

montrent tout de même que l'on peut exporter le modèle créé en dehors de Madagascar et produire des estimations globales robustes.

La question demeure toutefois plus complexe dès lors que l'on s'intéresse aux échelles les plus fines, en particulier à celle des zones bâties. Comme cela a été observé dans la commune de Cocody, le contexte socio-économique spécifique – une population plus aisée, une taille moyenne des ménages plus faible, et la présence probable de bâtiments secondaires ou vacants – a conduit à des difficultés dans l'estimation de la population, entraînant, même avec l'utilisation de Google 2.5D, une surestimation persistante de la population des zones bâties. Ainsi, malgré les améliorations notables apportées par cette nouvelle identification, le modèle continue de présenter des disparités locales, avec des sous- et surestimations selon les contextes. À ce stade, il apparaît encore difficile de s'appuyer uniquement sur ce type de modélisation pour estimer la population d'une petite aire géographique, comme cela pourrait être requis dans le cadre d'un projet de collecte ciblée. Une validation locale, même partielle, reste donc indispensable pour garantir la fiabilité des estimations à une échelle opérationnelle fine.

Bien que cette thèse propose une contribution originale à l'estimation de la population directement à partir des zones bâties, elle n'échappe pas à certaines limites et sources de regret. Tout d'abord, dans les chapitres 5 et 6, il aurait été particulièrement pertinent de disposer d'une cartographie du bâti pour une autre région de Côte d'Ivoire. Cela aurait permis d'évaluer la robustesse du modèle ascendant en dehors du contexte d'Abidjan, agglomération urbaine très densément peuplée. Bien que cette zone constitue un cas d'étude essentiel pour tester la méthode, elle est finalement peu représentative de l'ensemble du territoire ivoirien et encore moins de l'ensemble de la diversité des contextes démographiques d'Afrique. L'ajout d'une région moins densément peuplée, aurait permis d'obtenir des indices supplémentaires sur la capacité que l'on a d'exporter le modèle dans d'autres contextes, notamment celui de la République démocratique du Congo précédemment cité.

Dans cette thèse, nous avons travaillé sur plusieurs pays – Madagascar (chapitres 3, 5 et 6), le Bénin (chapitre 4), Mayotte (chapitre 5) et la Côte d'Ivoire (chapitres 5 et 6) – chacun avec des critères et découpages administratifs différents. Dans chaque contexte, il a été difficile d'obtenir les tracés administratifs correspondant aux dates des données démographiques utilisées, issues des derniers recensements. Pourtant, ces données administratives sont essentielles dans le cadre d'études spatiales contemporaines et notamment pour assurer le suivi de l'évolution

des données de recensement en général. À Madagascar, par exemple, nous avons dû effectuer un travail important pour reconstituer la base de données géographique liant décomptes de population du recensement de 2018 avec les tracés administratifs correspondant. Des hypothèses supplémentaires ont été nécessaires pour certaines communes, dont la localisation était difficile à déterminer, ce qui a forcément introduit des biais susceptibles de même affecter l'étape finale de l'approche globale développée.

Par ailleurs et de manière générale, nous avons travaillé dans cette thèse avec les résultats de TeleCense et l'identification de Google 2.5D, deux programmes issus d'entreprises, qui si elles ont des statuts différents, demeurent privées et à but commercial : cela soulève donc évidemment la question de notre dépendance à ces structures privées à qui ces données appartiennent. WorldPop a commencé à utiliser les données de Google 2.5D pour les estimations de population. Cependant, rien ne garantit que Google poursuive le développement et la mise à disposition de ces données après 2023, en particulier dans des zones sensibles et/ou en conflit où les enjeux géopolitiques et commerciaux peuvent être importants. Cette situation soulève donc la question de la fragilité de nos ressources et met en lumière la nécessité de fonder des modèles d'estimations de population pérennes reposant sur des initiatives de détection indépendantes et publiques afin de garantir leur accessibilité et leur autonomie.

Enfin, un regret majeur, tant dans l'approche développée avec TeleCense qu'avec Google 2.5D, a été l'incapacité à développer une désagrégation prenant en compte davantage de variables que la simple surface pour répartir la population. Bien que cette méthode ait permis d'obtenir la répartition la plus précise en termes d'erreurs, le fait de ne pas intégrer d'autres variables dans cette répartition a nettement limité notre capacité d'adaptation par la suite, notamment lors de la construction du modèle ascendant. Ainsi, malgré des structures bâties variées issues de plusieurs régions de Madagascar, il restait difficile de construire un modèle ascendant capable de s'adapter à n'importe quelle shape de territoire en dehors de Madagascar.

Dans ce sens, de nombreuses pistes restent à explorer, d'autant plus que les progrès récents, illustrés par l'arrivée des données Google 2.5D, ouvrent la voie à des travaux supplémentaires prometteurs, ainsi qu'à l'espoir de modèles encore plus précis. L'un des principaux apports de cette nouvelle détection réside dans sa disponibilité sur plusieurs années et sur la majorité du continent africain. Cela signifie que, moyennant les contraintes techniques de stockage et la disponibilité des données

de recensement et des découpages administratifs (ce qui, comme on l'a vu, reste un réel défi), il est désormais envisageable de construire une désagrégation, non plus seulement avec des régions différentes d'un même pays, mais directement à partir de régions appartenant à plusieurs pays différents. Cela permettrait d'enrichir le modèle de désagrégation puis ascendant avec une diversité bien plus importante de contextes, concernant le bâti et la répartition de la population. Ce serait également l'occasion de mieux mobiliser la variable de hauteur, autre apport clé des données 2.5D, mais dont la pertinence s'est révélée limitée à Madagascar en raison de la faible hauteur moyenne des bâtiments.

Une fois cette désagrégation élargie à plusieurs pays menée à bien – idéalement avec des variables discriminantes permettant de mieux différencier les shapes entre elles – un nouveau modèle ascendant pourra être construit, reposant sur une plus grande différence de répartition de la population dans les shapes. Ce sera également le moment opportun pour expérimenter d'autres approches de modélisation, au-delà des modèles linéaires ou des forêts aléatoires. On peut notamment envisager de recourir à des réseaux de neurones, à des méthodes bayésiennes ou d'analyse spatiale. Puis, une fois convaincu des nouvelles modélisations il sera important de commencer à caractériser la population en sortie. En effet, jusqu'ici, cette thèse s'est concentrée sur le dénombrement de population, mais il serait pertinent de chercher à caractériser la population, notamment en âge et de sexe.

Enfin, à défaut de disposer dans la sphère publique de données suffisamment précises pour désagréger dans les shapes, pourquoi ne pas imaginer la démarche menée par les instituts statistiques eux-mêmes ? Lors des recensements, ceux-ci disposent de données géoréférencées de très haute précision, ou du moins des zones de dénombrement dans lesquelles se situent la population. Plutôt que de désagréger les totaux dans les shapes, une piste serait de s'appuyer directement sur ces données, et de suivre la population entre deux recensements à l'aide d'un modèle construit à partir de ces décomptes. Si les résultats s'avèrent robustes, cela pourrait ensuite être généralisé à d'autres régions connaissant des lacunes de collecte.

Pour conclure, on ne remplacera jamais un recensement national ni les enquêtes de terrain, qui restent – comme l'a montré cette thèse – des sources primordiales. Nous restons fondamentalement tributaires de données de terrain de qualité car ce sont elles qui permettent de connaitre avec précision la population et ses caractéristiques et sans elles, il est tout simplement impossible de mettre en relation la répartition de la population avec ses autres caractéristiques socio-démographiques. Sans elles, il est également difficile de construire des modèles fiables. Cela étant dit,

les approches par satellite et la modélisation peuvent jouer un rôle d'appui important, notamment durant les périodes intercensitaires ou dans des contextes où la collecte sur le terrain est difficile, voire impossible. Il est donc essentiel de poursuivre ces travaux, en mettant notamment l'accent sur l'actualisation des tracés géographiques, condition indispensable à toute analyse spatiale contemporaine.

Liste des tableaux

Tableau 1 : Récapitulatif des projets de cartographie de bâti 54
Tableau 2 : récapitulatif des variables 73
Tableau 3 : Récapitulatif du nombre de communes recensées en 2018 et du nombre de communes disponibles dans le fichier géoréférencé (OCHA), pour en déduire le nombre de communes manquantes pour chaque région 91
Tableau 4 : Déterminer la commune mère d'Antsahabe 99
Tableau 5 : Déterminer la commune mère de Bemara Atsinanana 100
Tableau 6 : Déterminer la commune mère de Maromitety I..... 101
Tableau 7 : Déterminer la commune mère d'Andragnanivo 101
Tableau 8 : Déterminer la commune mère d'Ankilivalo 101
Tableau 9 : Récapitulatif et vérification des localisations de chacune des nouvelles communes 102
Tableau 10 : Nombre de shapes par régions 106
Tableau 11 : Distribution de la population des 430 communes..... 109
Tableau 12 : Récapitulatif et distribution des variables pour les deux couches de données 112
Tableau 13 : Résultats et coefficients de la modélisation de la densité de population des communes de Madagascar..... 117
Tableau 14 : Comparaison des performances des modèles de régression linéaire et de forêt aléatoire pour la prédiction de la densité de population..... 119
Tableau 15 : Distribution de la population selon les différentes méthodes de pondération et à partir de la désagrégation de la population des districts 121
Tableau 16 : Distribution de l'erreur relative selon les différentes méthodes de pondération avec la surface TeleCense..... 121
Tableau 17 : Distribution de l'erreur relative selon les différentes méthodes de pondération avec la surface TeleCense ajustée à partir de la détection Google 122
Tableau 18 : Distribution de l'erreur au niveau des communes pour la désagrégation TeleCense..... 127
Tableau 19 : Distribution de l'erreur au niveau des communes à partir des estimations non contraintes de WorldPop en 2018..... 127
Tableau 20 : Projections des décomptes de population des arrondissements de la commune d'Abomey-Calavi au Bénin pour les années 2014 à 2023 151
Tableau 21 : Comparaison de la distribution de l'erreur entre le Bénin et Madagascar 153
Tableau 22 : Coefficients associés à la modélisation linéaire de la population des shapes de Madagascar..... 165
Tableau 23 : Comparaison des performances des modèles ascendants pour la prédiction de la population..... 167

Tableau 24 : Erreurs d'estimation selon le niveau administratif d'agrégation des résultats 178
Tableau 25 : Comparaison des distributions de population entre l'estimation par désagrégation et par modèle ascendant 179
Tableau 26 : Estimation de la population et erreur relative des 10 communes d'Abidjan 191
Tableau 27 : 193
Tableau 28 : Distribution de l'erreur relative selon les deux méthodes de pondération 207
Tableau 29 : Répartition des erreurs relatives selon la source des données de bâti 207
Tableau 30 : Résultats du modèle ascendant par modélisation linéaire pour les données 2.5D 209
Tableau 31 : Comparaison de la performance des modèles ascendants 211
Tableau 32 : Evaluation des performances du modèle ascendant à partir des shapes Google 2.5D dans la sous-préfecture d'Abidjan 213
Tableau 33 : Evaluation des performances du modèle ascendant à partir des shapes TeleCense dans la sous-préfecture d'Abidjan 214
Tableau 34 : Projections des décomptes de population des arrondissements du Bénin pour les années 2014 à 2023 258

Liste des figures

Figure 1 : Période du dernier recensement pour l'ensemble des pays africains 21
Figure 2 : Approches descendantes et ascendantes de l'estimation de population carroyée à partir d'images satellites..... 24
Figure 3 : Satellite Sentinel-1 28
Figure 4 : Satellite Sentinel-2 28
Figure 5 : Illustration des algorithmes de détection TeleCense – identification de zones bâties dans la région d'Abuja, Nigéria - 2018. 29
Figure 6 : Identification des zones bâties Google 2.5D dans la région d'Abuja, Nigéria - 2018 31
Figure 7 : Développement étapes par étapes des approches descendantes et ascendantes de la thèse..... 37
Figure 8 : Architecture du programme TeleCense..... 61
Figure 9 : Nord de Cotonou (Bénin) vu de l'espace par Sentinel 2 – 2023..... 62
Figure 10 : Identification du bâti au 1er janvier 2023 au nord de Cotonou, Bénin ... 62
Figure 11 : Variation de la densité de bâti des shapes 67
Figure 12 : Identification de l'utilisation du sol via OSM pour calculer la surface habitable des shapes..... 68
Figure 13 : Occupation de l'hôpital dans la shape totale 69
Figure 14 : Problème d'identification pour certaines grandes shapes – Région de Diana – Madagascar - 2018 71
Figure 15 : Exemple schématique de désagrégation avec pondération 75
Figure 16 : Exemple schématique de la désagrégation par interpolation surfacique 77
Figure 17 : Carte de situation et présentation des six régions retenues à Madagascar 85
Figure 18 : Illustration de la désagrégation à partir d'une commune de la région d'Itasy 86
Figure 19 : Densité de population par district des six régions retenues à Madagascar 87
Figure 20 : Scenario idéal représentant la probable division de la commune A..... 92
Figure 21 : Proposition de solution : la commune B étant localisée dans la commune A, on ajoute à la commune A les 6000 habitants de la commune B..... 93
Figure 22 : Localisation des nouvelles communes de la région d'Itasy 95
Figure 23 : Localisation des nouvelles communes de la région de Diana 96
Figure 24 : Distribution des effectifs des communes avant et après modification, et des communes ne nécessitant pas de modifications 104
Figure 25 : Exemple d'une shape intersectant les communes de Bemaneviky Haut Sambirano et Ambohimarina deux communes – Région de Diana, Madagascar... 108

Figure 26 : Distribution de la densité de population avant et après transformation logarithmique 110
Figure 27 : Relation entre la surface totale des régions, le nombre de shapes et la densité de population par région 111
Figure 28 : Analyse en Composante Principales des variables des communes de Madagascar 114
Figure 29 : Évaluation de la normalité et de l'homogénéité des résidus de la modélisation linéaire 118
Figure 30 : Importance des variables de l'algorithme de forêt aléatoire 119
Figure 31 : Comparaison des erreurs au niveau régional en fonction des deux premières méthodes de désagrégation 123
Figure 32 : Carte des erreurs relative d'Itasy avec la méthode d'interpolation surfacique 123
Figure 33 : Carte des erreurs relative d'Itasy avec la méthode utilisant la densité de bâti 124
Figure 34 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – exemple dans la région de Diana - 2018 126
Figure 35 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop pour toutes les régions – 2018 127
Figure 36 : 129
Figure 37 : Carte de situation du Bénin et localisation de ses départements 138
Figure 38 : Zone du Grand Nokoué – image sentinel-S2 optique 2023, résolution 10m 139
Figure 39 : Densité de population des communes (niveau 2) du Bénin – décompte de population du RGPH4 - 2013 141
Figure 40 : Densité de population des communes de la zone élargie du Grand Nokoué – décompte de population du RGPH4 - 2013 142
Figure 41 : Évolution du nombre de shapes et de leur surface totale au fil des années au Bénin 145
Figure 42 : Différence d'identification des shapes entre 2017 et 2023 – Arrondissement d'Hêvié 146
Figure 43 : Identification satisfaisante des shapes dans la région sud du Grand Nokoué – 2017 147
Figure 44 : Identification satisfaisante des shapes dans la commune de Porto-Novo – 2017 147
Figure 45 : Pourcentage d'arrondissements appartenant à chaque 154
Figure 46 : Histogramme des résidus du modèle ascendant 166
Figure 47 : distribution des Résidus en fonction des valeurs prédites 166

Figure 48 : Importance des variables de l'algorithme de forêt aléatoire pour le modèle ascendant 167
Figure 49 : Carte de situation du département de Mayotte 169
Figure 50 : Les 17 communes de Mayotte 173
Figure 51 : Répartition des villages et ilots de Mayotte 173
Figure 52 : Vue satellitaire de la répartition de la population dans les ilots – Commune de Mamoudzou – 2017 174
Figure 53 : Erreur relative lors de l'agrégation des estimations pour les 17 communes de Mayotte 176
Figure 54 : Erreur relative lors de l'agrégation des estimations pour les 72 villages de Mayotte 177
Figure 55 : Relation entre la population désagrégée et celle estimée par le modèle ascendant 179
Figure 56 : Résidus entre la population désagrégée et celle estimée par le modèle ascendant 180
Figure 57 : Graphique QQ-plot 180
Figure 58 : Comparaison de la surface des shapes selon les trois catégories d'estimation 181
Figure 59 : Comparaison du nombre de logements par ilots selon les trois catégories d'estimation 182
Figure 60 : Carte de situation de la Côte d'Ivoire 185
Figure 61 : Découpage administratif des dix communes de la sous-préfecture d'Abidjan : 186
Figure 62 : Densité de population des dix communes de la sous-préfecture d'Abidjan 187
Figure 63 : Répartition des shapes en fonction de la surface dans les 10 communes d'Abidjan 189
Figure 64 : Erreur d'estimation au niveau communal de la sous-préfecture d'Abidjan 190
Figure 65 : Présentation des données Google 2.5D dans nord est de Mamoudzou à Mayotte - 2023 Mamoudzou 201
Figure 66 : Identification du bâti Google 2.5D avec le seuil conseillé de 0,34 202
Figure 67 : Identification du bâti Google 2.5D – toujours avec un seuil de 0,34 - retravaillée pour une meilleure segmentation des bâtiments entre eux 202
Figure 68 : Comparaison des détections Google 2.5D et TeleCense dans la ville d'Imerintsiatosika – Région d'Itasy – 2018 204
Figure 69 : Répartition de la population estimée par désagrégation, excluant les 1 % des shapes avec la plus grande population. 208
Figure 70 : Distribution des erreurs selon la source des données de bâti 212

Figure 71 : Comparaison du RMSE par commune selon la source des données de bâti 215
Figure 72 : Relation entre le nombre de shapes par communes et les décomptes de population du RGPH 2021 pour la sous-préfecture d'Abidjan..... 218
Figure 73 : 220
Figure 74 : Localisation des nouvelles communes de la région de Melaky 248
Figure 75 : Localisation des nouvelles communes de la région d'Atsinanana 249
Figure 76 : Localisation des nouvelles communes de la région d'Analamanga..... 250
Figure 77 : Localisation des nouvelles communes de la région d'Androy..... 251
Figure 78 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Itasy - 2018 253
Figure 79 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Atsinanana - 2018..... 254
Figure 80 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Androy - 2018..... 254
Figure 81 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région de Melaky - 2018 255
Figure 82 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Analamanga - 2018..... 255
Figure 83 : Superposition de la détection TeleCense (en rouge et en 2018) et de la détection Meta (en noir et en 2020) dans la région de Diana 256
Figure 84 : Image satellite montrant les zones bâties et la détection de Meta (en noir et en 2020)..... 256
Figure 85 : Évolution du nombre de shapes et de leur surface totale au fil des années au Bénin avec la détection TeleCense 257

Annexes

Annexe 1 : Requête OSM

```
nwr[landuse=religious];
nwr[landuse=industrial];
nwr[landuse=cemetery];
nwr[landuse=commercial];
nwr[landuse=retail];
nwr[landuse=garages];
nwr[landuse=military];
nwr[landuse=school];
nwr[landuse=construction];
nwr[landuse=education];
nwr[landuse=fairground];
nwr[landuse=depot];
nwr[landuse=garages];
nwr[landuse=port];
nwr[landuse=railway];
nwr[landuse=quarry];
nwr[landuse=brownfield];
nwr[landuse=landfill];
nwr[landuse=farmland];
nwr[landuse=farmyard];
nwr[landuse=animal_keeping];
nwr[landuse=logging];
nwr[landuse=orchard];
nwr[leisure];
nwr[amenity=police];
nwr[amenity=public_building];
nwr[amenity=college];
nwr[amenity=prison];
nwr[amenity=school];
nwr[amenity=hospital];
nwr[amenity=nursing_home];
nwr[amenity=clinic];
nwr[amenity=parking];
nwr[amenity=university];
nwr[amenity=arts_center];
nwr[amenity=casino];
```

```
nwr[amenity=cinema];
nwr[amenity=theatre];
nwr[amenity=conference_centre];
nwr[amenity=courthouse];
nwr[amenity=embassy];
nwr[amenity=townhall];
nwr[amenity=community_centre];
nwr[amenity=conference_centre];
nwr[amenity=place_of_worship];
nwr[amenity=marketplace];
nwr[amenity=post_office];
nwr[amenity=recycling];
nwr[amenity=restaurant];
nwr[amenity=fast_food];
nwr[amenity=driving_school];
nwr[amenity=bar];
nwr[amenity=cafe];
nwr[amenity=food_court];
nwr[amenity=ice_cream];
nwr[amenity=pub];
nwr[amenity=dancing_school];
nwr[amenity=first_aid_school];
nwr[amenity=kindergarten];
nwr[amenity=language_school];
nwr[amenity=library];
nwr[amenity=surf_school];
nwr[amenity=toy_library];
nwr[amenity=research_institute];
nwr[amenity=training];
nwr[amenity=music_school];
nwr[amenity=traffic_park];
nwr[amenity=car_rental];
nwr[amenity=car_wash];
nwr[amenity=vehicle_inspection];
nwr[amenity=driver_training];
nwr[amenity=fuel];
nwr[amenity=bank];
nwr[amenity=bureau_de_change];
nwr[amenity=money_transfer];
nwr[amenity=payment_centre];
```

```
nwr[amenity=dentist];
nwr[amenity=doctors];
nwr[amenity=pharmacy];
nwr[amenity=social_facility];
nwr[amenity=veterinary];
nwr[amenity=brothel];
nwr[amenity=events_venue];
nwr[amenity=exhibition_centre];
nwr[amenity=gambling];
nwr[amenity=love_hotel];
nwr[amenity=music_venue];
nwr[amenity=nightclub];
nwr[amenity=planetarium];
nwr[amenity=social_centre];
nwr[amenity=stage];
nwr[amenity=stripclub];
nwr[amenity=studio];
nwr[amenity=swingerclub];
nwr[amenity=fire_station];
nwr[amenity=post_depot];
nwr[amenity=ranger_station];
nwr[amenity=lounge];
nwr[amenity=shelter];
nwr[amenity=animal_boarding];
nwr[amenity=animal_breeding];
nwr[amenity=animal_shelter];
nwr[amenity=baking_oven];
nwr[amenity=clock];
nwr[amenity=crematorium];
nwr[amenity=dive_center];
nwr[amenity=funeral_hall];
nwr[amenity=grave_yard];
nwr[amenity=hunting_stand];
nwr[amenity=internet_cafe];
nwr[amenity=monastery];
nwr[amenity=mortuary];
nwr[amenity=place_of_mourning];
nwr[amenity=place_of_worship];
nwr[amenity=public_bath];
nwr[shop];
```

```
nwr[shop=supermarket];
nwr[military];
nwr[office];
nwr[tourism=alpine_hut];
nwr[tourism=apartment];
nwr[tourism=aquarium];
nwr[tourism=artwork];
nwr[tourism=attraction];
nwr[tourism=camp_site];
nwr[tourism=camp_pitch];
nwr[tourism=caravan_site];
nwr[tourism=chalet];
nwr[tourism=gallery];
nwr[tourism=guest_house];
nwr[tourism=hostel];
nwr[tourism=hotel];
nwr[tourism=information];
nwr[tourism=motel];
nwr[tourism=museum];
nwr[tourism=theme_park];
nwr[tourism=wilderness_hut];
nwr[tourism=zoo];
nwr[religion];
nwr[healthcare];
nwr[bridge];
nwr[emergency];
nwr[historic];
nwr[sport];
nwr[telecom=data_center];
nwr[aeroway=aerodrome];
nwr[aeroway=hangar];
nwr[building=industrial];
nwr[building=hangar];
nwr[building=warehouse];
nwr[building=college];
nwr[building=school];
nwr[building=university];
nwr[building=prison];
nwr[building=hospital];
nwr[building=nursing_home];
```

```
nwr[building=clinic];
nwr[building=church];
nwr[building=cathedral];
nwr[building=chapel];
nwr[building=mosque];
nwr[building=synagogue];
nwr[building=temple];
nwr[building=shrine];
nwr[building=stadium];
nwr[building=office];
nwr[building=service];
nwr[building=parking];
nwr[building=arts_center];
nwr[building=casino];
nwr[building=cinema];
nwr[building=theatre];
nwr[building=conference_centre];
nwr[building=courthouse];
nwr[building=embassy];
nwr[building=townhall];
nwr[building=community_centre];
nwr[building=place_of_worship];
nwr[building=marketplace];
);
```

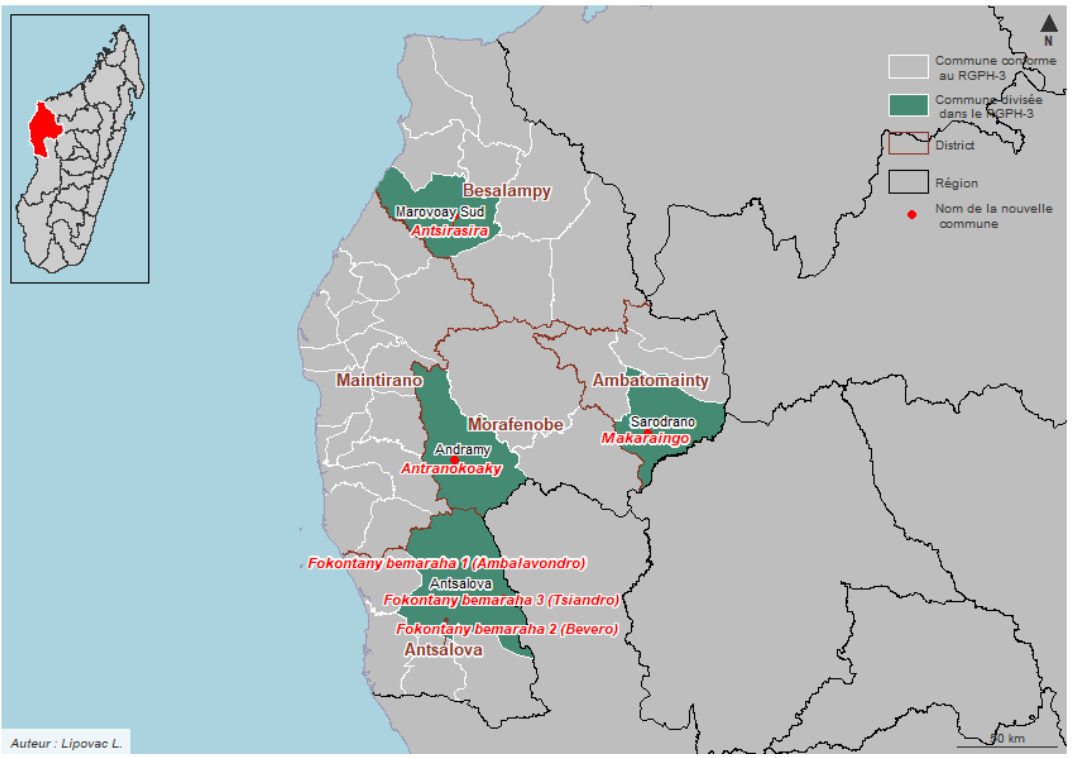
Annexe 2 : Mise en place de la méthode permettant de localiser les nouvelles communes dans les régions de Melaky, Atsinanana, Analamanga et Androy

1) Melaky

a) Modifications démographiques

Dans la région de Melaky, quatre nouvelles communes ont été créées (noms en rouge sur la figure 74). C'est le cas de Makaraingo (district de Sarodrano) localisée à Sarodrano (Wikipedia, 2021b), de Bemara Atsinanana* (district d'Antsalova) localisée à Antsalova, d'Antsirasira (district de Marovoay Sud) localisée à Marovoay Sud (Mindat, 2024h) et d'Antranokoaky (district d'Andramy) qui se trouve quant à elle dans la commune d'Andramy (Wikipedia, 2022b).

Figure 74 : Localisation des nouvelles communes de la région de Melaky



Sources : Tracés des limites administratives fournis par OCHA et localisation des nouvelles communes par Mindat et Wikipedia

De la même manière, nous ajoutons 5915 habitants à Sarodrano, 2610 habitants à Antsalova, 1835 habitants à Marovoay Sud et 5236 à Antranokoaky.

b) Modifications apportées aux limites administratives

La commune nommée Berevo/Ranobe est corrigée en « Berevo/ranobe dans la table du RGPH-3.

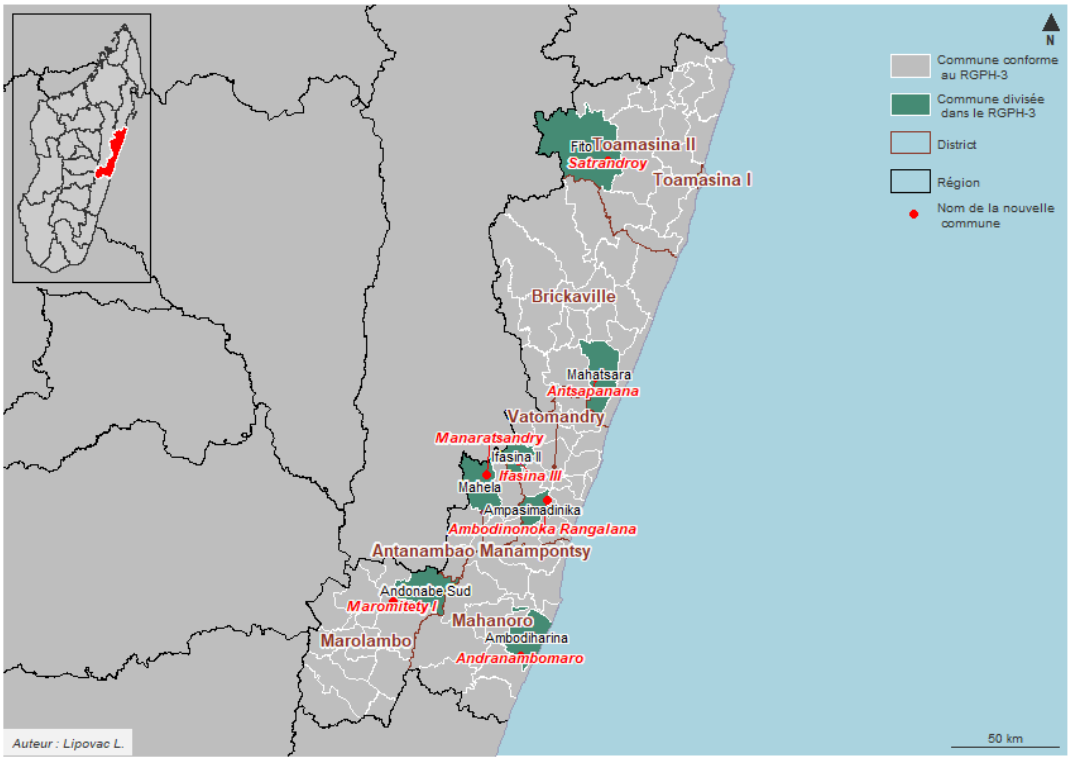
2) Atsinanana

a) Modifications démographiques / sur les données de population :

Sept nouvelles communes ont été créées dans la région d'Atsinanana (figure 75), il s'agit de Manaratsandry (district d'Antanambao Manampontsy), d'Antsapanana (district de Brickaville), d'Andranambomaro (district de Mahanoro), de Maromitety I* (district de Marolambo), de Satrandroy (district de Toamasina II), d'Ambodinonoka Rangalana et de Ifasina Iii (district de Vatomandry).

Nous avons respectivement localisé ces communes au sein de Mahela (Mindat, 2024i), Mahatsara (Mindat, 2024j), Ambodiharina (Mindat, 2024k), Andonabe Sud, Fito (Wikipedia, 2023a), Ampasimadinika (Mindat, 2024l) et de Ifasina II (Wikipedia, 2021c). Ainsi, pour chacune de ces communes mères, nous ajoutons respectivement : 5 833, 13 538, 14 510, 10 757, 6 665, 6 455 et 2 910 habitants en plus.

Figure 75 : Localisation des nouvelles communes de la région d'Atsinanana



Sources : Tracés des limites administratives fournis par OCHA et localisation des nouvelles communes par Mindat et Wikipedia

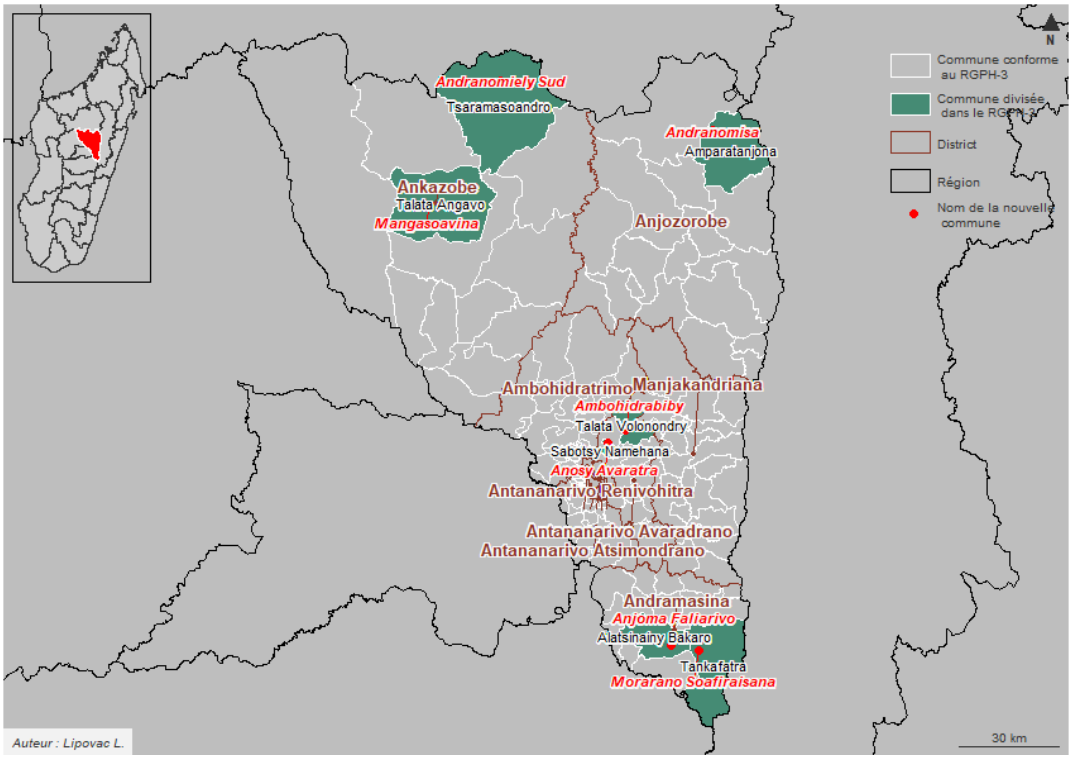
3) Analamanga

a) Modifications démographiques

Sept communes ont également été créées dans la région d'Analamanga (figure 76), il s'agit d'Anjoma Faliarivo, de Morarano Soafiraisana (district d'Andramasina), d'Andranomisa (district d'Anjozorobe), d'Andranomiely Sud et de Mangasoavina (district d'Ankazobe), d'Ambohidrabiby et d'Anosy Avaratra (district d'Antananarivo Avaradrano).

Ces communes se trouvent respectivement dans les communes d'Alatsinainy Bakaro (Mindat, 2024m), Tankafatra (Mindat, 2024n), Amparatanjona (Wikipedia, 2021a), Talata Angavo (Mindat, 2024o), Tsaramasoandro (Mindat, 2024p), Talata Volonondry (Wikipedia, 2022c) et de Sabotsy Namehana (Wikipedia, 2022a). Nous ajoutons respectivement à chacune de ces communes mères : 6965, 6518, 7692, 5384, 6556, 15243 et 7873 habitants en plus.

Figure 76 : Localisation des nouvelles communes de la région d'Analamanga



Sources : Tracés des limites administratives fournis par OCHA et localisation des nouvelles communes par Mindat et Wikipedia

b) Modifications apportées aux limites administratives

La commune nommée Androhibe est renommée Androhibe Antsahadinta dans le fichier géoréférencé OCHA afin de correspondre aux résultats du RGPH-3.

La commune d'Ankazondady est quant à elle corrigée en Ankazondandy dans la base de données du recensement.

Enfin, les communes d'Ivato Aéroport et d'Ivato Firaisana sont fusionnées en une seule et même commune nommée Ivato.

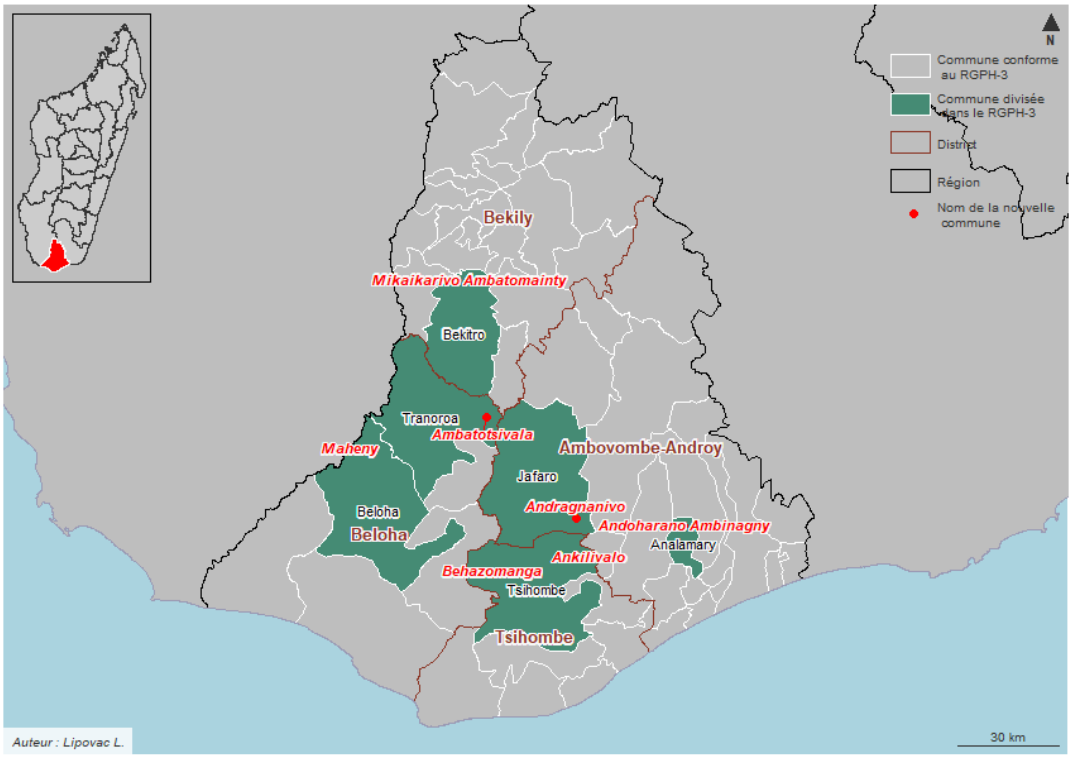
4) Androy

a) Modifications démographiques

Les communes d'Andoharano Ambinagny et d'Andragnanivo* (district d'Ambovombe-Androy), de Mikaikarivo Ambatomainty (district de Bekily), d'Ambatotsivala et de Maheny (district de Beloha), d'Ankilivalo* et de Behazomanga (district de Tsihombe) ont été créées (figure 77).

Ces communes se trouvent respectivement dans les communes d'Analamary (Mindat, 2024q), Jafaro (Google Maps, 2024), Bekitro (Mindat, 2024r), Beloha (Mindat, 2024s), Tranoroa (Mindat, 2024t) et de Tsihombe qui accueille les deux dernières communes (Mindat, 2024u, 2024v). Ainsi, pour chacune des communes mères nous ajoutons respectivement : 7059, 3386, 5056, 8449, 7319, et de 5714 et 8164 habitants en plus.

Figure 77 : Localisation des nouvelles communes de la région d'Androy



Sources : Tracés des limites administratives fournis par OCHA et localisation des nouvelles communes par Mindat

b) Modifications apportées aux limites administratives

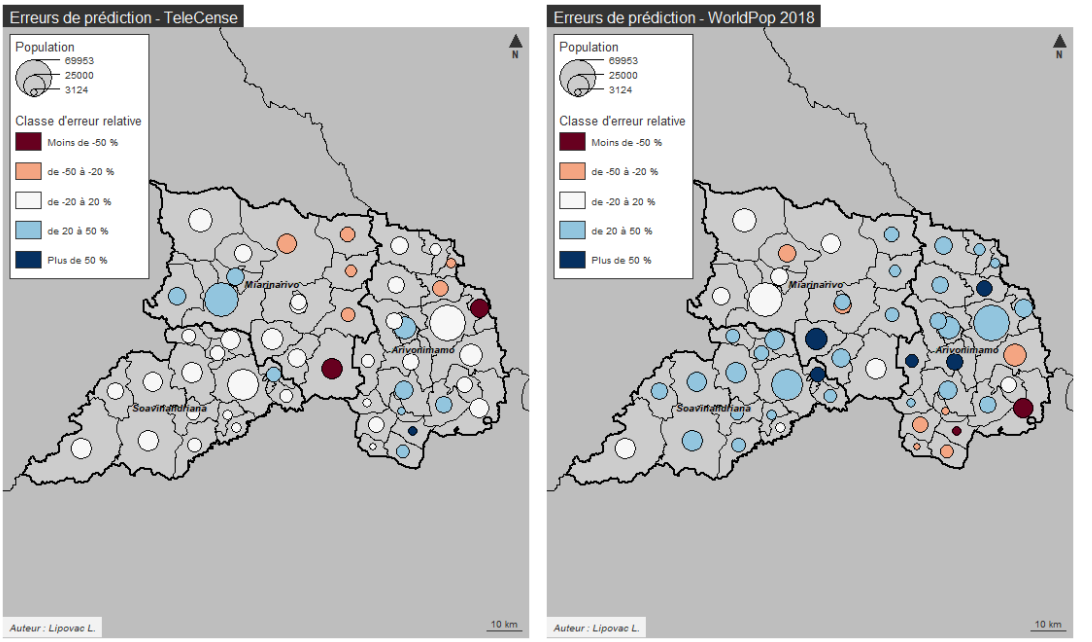
La commune nommée Morafeno Bekily est renommée Bekily-Centrale dans le fichier géoréférencé OCHA afin de correspondre aux résultats du RGPH-3.

Les communes d'Anja-Nord, d'Ikopoky et Belindo/Mahasoa sont quant à elles corrigées en Anja Nord, Kopoky et Belindo Mahasoa dans la base de données du recensement.

Annexe 3 : Comparaison des estimations de la désagrégation TeleCense avec les estimations WorldPop dans les régions de Madagascar

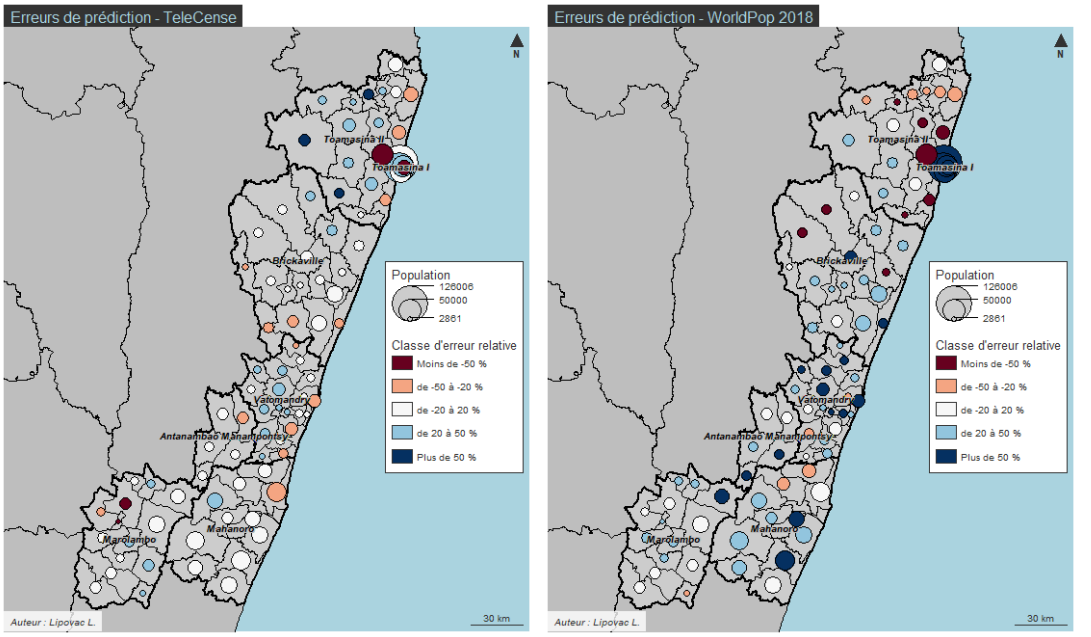
Les erreurs de prédiction sont pour toutes les régions moins importantes avec la désagrégation surfacique qu'avec les estimations non contraintes de WorldPop en 2018.

Figure 78 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Itasy - 2018



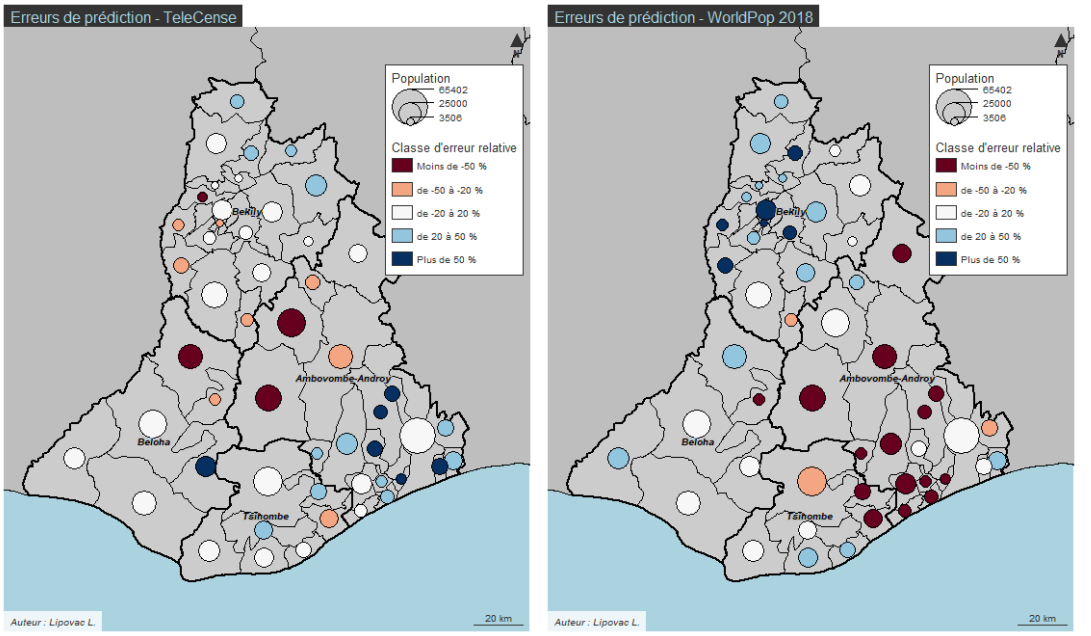
Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection de bâti TeleCense dont la surface est ajustée par la détection Google et des estimations non contraintes de WorldPop en 2018

Figure 79 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Atsinanana - 2018



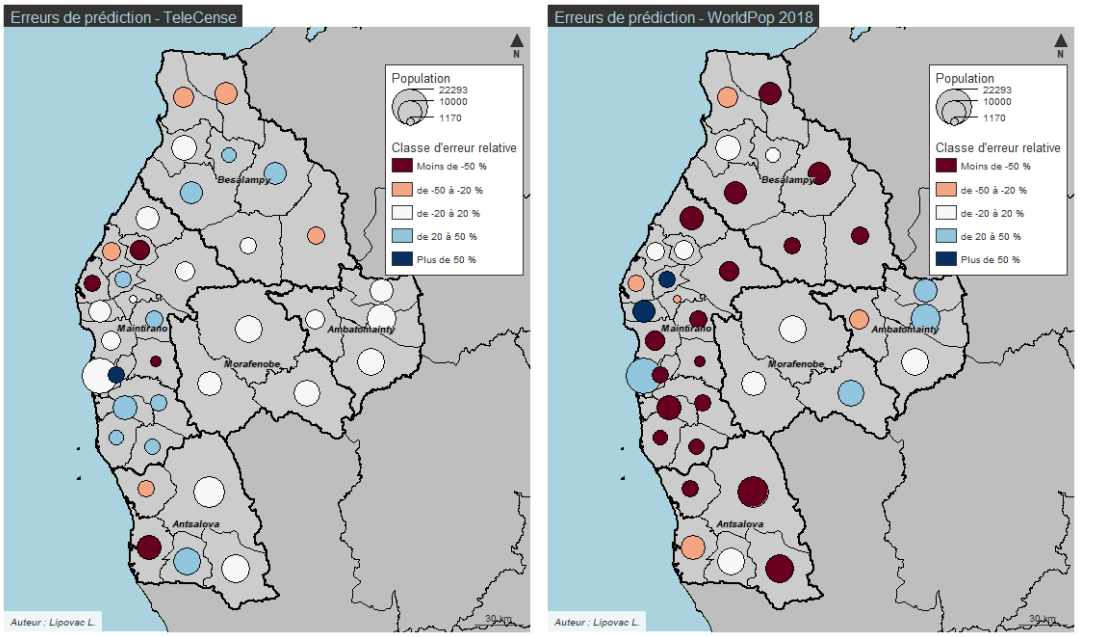
Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection de bâti TeleCense dont la surface est ajustée par la détection Google et des estimations non contraintes de WorldPop en 2018

Figure 80 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Androy - 2018



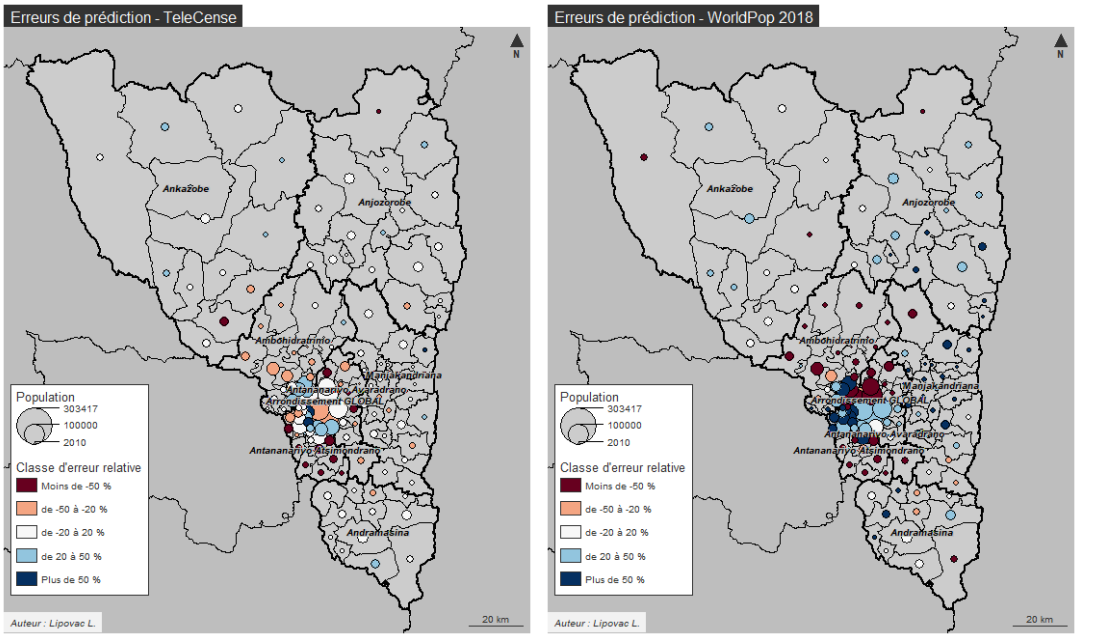
Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection de bâti TeleCense dont la surface est ajustée par la détection Google et des estimations non contraintes de WorldPop en 2018

Figure 81 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région de Melaky - 2018



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection de bâti TeleCense dont la surface est ajustée par la détection Google et des estimations non contraintes de WorldPop en 2018

Figure 82 : Comparaison des erreurs de prédiction entre la désagrégation TeleCense et l'estimation du programme WorldPop – région d'Analamanga - 2018



Sources : Calculs de l'auteur à partir des décomptes de population du RGPH3 (INSTAT), de la détection de bâti TeleCense dont la surface est ajustée par la détection Google et des estimations non contraintes de WorldPop en 2018

Annexe 4 : Détection de Meta (data for good) en 2020 à Madagascar

Voici l'un des nombreux exemples qui montre qu'il n'est pas possible de se servir de Meta pour comparer nos estimations, en tous cas à Madagascar en 2020 (Figures 83 et 84). En rouge ce sont les shapes détectées par TeleCense et en noir c'est ce qui est détecté par Meta (ou alors juste des zones qu'ils ont considérées comme non habitées) ; on observe une grande différence dans la détection des zones bâties.

Figure 83 : Superposition de la détection TeleCense (en rouge et en 2018) et de la détection Meta (en noir et en 2020) dans la région de Diana

Sources : Identification du bâti par le programme TeleCense et identification de Meta – Image issue du logiciel QGIS

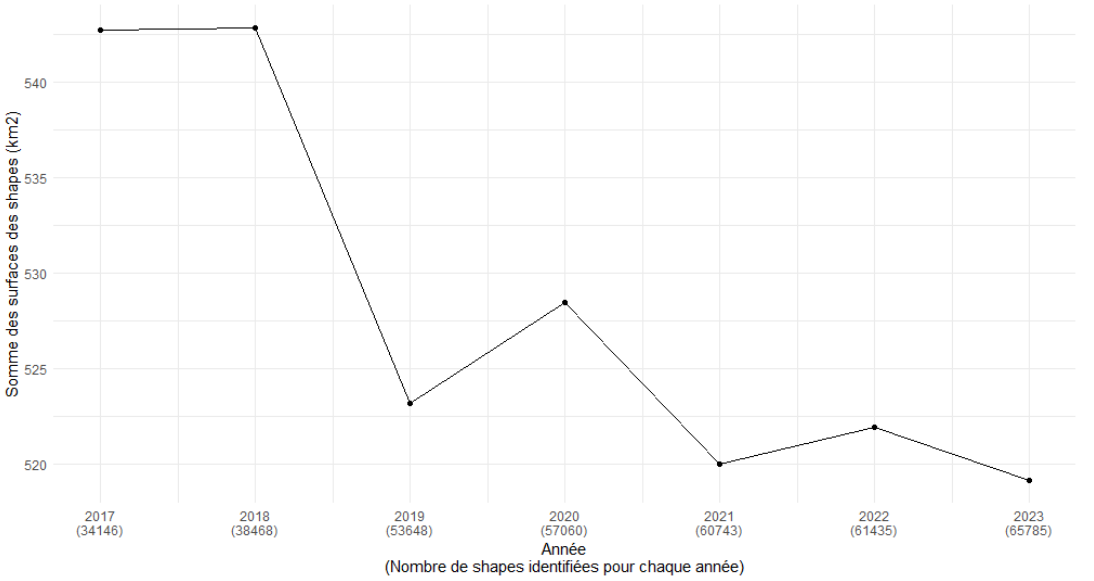
Figure 84 : Image satellite montrant les zones bâties et la détection de Meta (en noir et en 2020)

Sources : Identification de Meta – Image issue du logiciel QGIS

Annexe 5 : Cartographie et surface TeleCense de 2017 à 2023 au Bénin

La surface totale identifiée par le programme TeleCense est bien plus importante que lorsque l'on l'ajuste avec la détection Google.

Figure 85 : Évolution du nombre de shapes et de leur surface totale au fil des années au Bénin avec la détection TeleCense



Sources : Calculs de l'auteur à partir de la détection de bâti TeleCense

Annexe 6 : Projection des décomptes de population des arrondissements du Bénin

Tableau 34 : Projections des décomptes de population des arrondissements du Bénin pour les années 2014 à 2023

Pays	Département	Commune	Arrondissement	Effectif total RGPH4 (2013)	Effectif total RGPH3 (2002)	r	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Benin	Atlantique	Abomey-Calavi	Abomey-Calavi	117824	61450	0,060	124843	132280	140160	148509	157356	166730	176662	187186	198337	210152
Benin	Atlantique	Abomey-Calavi	Akassato	61262	17197	0,120	68586	76785	85965	96242	107748	120629	135050	151195	169271	189507
Benin	Atlantique	Abomey-Calavi	Glo-Djigbe	28103	12827	0,072	30132	32308	34641	37142	39824	42699	45782	49088	52633	56433
Benin	Atlantique	Abomey-Calavi	Godomey	253262	153447	0,046	264797	276858	289467	302652	316436	330849	345918	361673	378146	395369
Benin	Atlantique	Abomey-Calavi	Hevie	67218	13450	0,154	77553	89477	103234	119106	137419	158547	182924	211049	243498	280937
Benin	Atlantique	Abomey-Calavi	Ouedo	27522	10067	0,094	30096	32910	35988	39353	43033	47058	51458	56270	61532	67287
Benin	Atlantique	Abomey-Calavi	Togba	73331	18674	0,129	82812	93518	105609	119263	134682	152095	171759	193965	219042	247362
Benin	Atlantique	Abomey-Calavi	Zinvie	18157	13212	0,029	18677	19213	19764	20330	20913	21512	22129	22763	23416	24087
Benin	Atlantique	Kpomasse	Aganmalome	4523	4566	-0,001	4519	4515	4512	4508	4504	4500	4496	4493	4489	4485
Benin	Atlantique	Kpomasse	Agbanto	5694	4377	0,024	5829	5967	6108	6252	6400	6552	6707	6865	7028	7194
Benin	Atlantique	Kpomasse	Agonkanme	8072	6533	0,019	8225	8381	8540	8703	8868	9036	9208	9382	9560	9742
Benin	Atlantique	Kpomasse	Dedome	7301	5294	0,029	7513	7730	7954	8185	8422	8666	8918	9176	9442	9716
Benin	Atlantique	Kpomasse	Dekanme	9977	8581	0,013	10112	10248	10386	10526	10668	10812	10958	11106	11256	11407
Benin	Atlantique	Kpomasse	Kpomasse	10527	9240	0,012	10650	10774	10900	11027	11155	11285	11417	11550	11685	11821

Benin	Atlantique	Kpomasse	Segbeya	4016	3611	0,009	4054	4093	4131	4171	4210	4250	4291	4331	4372	4414
Benin	Atlantique	Kpomasse	Segbohoue	7347	6058	0,017	7474	7603	7735	7869	8005	8143	8284	8427	8573	8721
Benin	Atlantique	Kpomasse	Tokpa-Dome	10191	8930	0,012	10311	10433	10556	10681	10807	10935	11064	11195	11327	11461
Benin	Atlantique	Ouidah	Avlekete	11453	5636	0,065	12198	12992	13837	14737	15696	16717	17805	18963	20197	21511
Benin	Atlantique	Ouidah	Djegbadji	4997	4170	0,016	5078	5160	5244	5329	5415	5503	5592	5683	5775	5869
Benin	Atlantique	Ouidah	Gakpe	6236	4776	0,024	6386	6539	6696	6856	7021	7189	7362	7538	7719	7905
Benin	Atlantique	Ouidah	Ouakpe-Daho	3473	2941	0,015	3525	3577	3630	3685	3739	3795	3852	3909	3967	4026
Benin	Atlantique	Ouidah	Ouidah i	9224	8188	0,011	9322	9421	9522	9623	9726	9829	9934	10040	10146	10254
Benin	Atlantique	Ouidah	Ouidah Ii	13710	12856	0,006	13789	13868	13947	14027	14108	14188	14270	14352	14434	14517
Benin	Atlantique	Ouidah	Ouidah Iii	15207	9880	0,039	15801	16419	17060	17727	18420	19139	19887	20665	21472	22311
Benin	Atlantique	Ouidah	Ouidah Iv	9475	6723	0,031	9768	10071	10383	10704	11036	11378	11730	12093	12468	12854
Benin	Atlantique	Ouidah	Pahou	78474	14436	0,162	91219	106033	123254	143271	166539	193586	225026	261572	304053	353433
Benin	Atlantique	Ouidah	Savi	9785	6949	0,031	10087	10399	10720	11051	11393	11745	12107	12481	12867	13264
Benin	Atlantique	So-Ava	Ahomey-Lokpo	11026	8760	0,021	11254	11486	11724	11966	12213	12465	12723	12986	13254	13528
Benin	Atlantique	So-Ava	Dekanmey	6617	4241	0,040	6884	7162	7450	7751	8064	8389	8727	9079	9445	9826
Benin	Atlantique	So-Ava	Ganvie i	19155	10280	0,057	20245	21396	22613	23899	25259	26695	28214	29819	31515	33307
Benin	Atlantique	So-Ava	Ganvie Ii	18017	10288	0,051	18937	19904	20921	21989	23112	24292	25533	26837	28207	29648
Benin	Atlantique	So-Ava	Houedo-Aguekon	20909	10610	0,062	22209	23589	25055	26613	28267	30024	31890	33872	35977	38213
Benin	Atlantique	So-Ava	So-Ava	13347	9961	0,026	13699	14060	14430	14810	15201	15601	16012	16434	16867	17312
Benin	Atlantique	So-Ava	Vekky	29476	22175	0,026	30231	31006	31800	32615	33451	34308	35187	36088	37013	37961
Benin	Atlantique	Tori-Bossito	Avame	5351	4444	0,017	5440	5531	5623	5716	5811	5908	6007	6107	6208	6312
Benin	Atlantique	Tori-Bossito	Azohoue-Aliho	3915	2258	0,050	4111	4317	4534	4761	5000	5251	5514	5790	6080	6385

Benin	Atlantique	Tori-Bossito	Azohoue-Cada	8543	6457	0,025	8758	8979	9205	9437	9675	9919	10169	10425	10687	10957
Benin	Atlantique	Tori-Bossito	Tori-Bossito	14844	12481	0,016	15075	15309	15546	15788	16033	16282	16535	16792	17053	17318
Benin	Atlantique	Tori-Bossito	Tori-Cada	15729	11952	0,025	16118	16516	16924	17342	17771	18210	18660	19121	19593	20078
Benin	Atlantique	Tori-Bossito	Tori-Gare	9250	6977	0,025	9485	9726	9972	10226	10485	10751	11024	11304	11591	11885
Benin	Atlantique	Ze	Tangbo-Djevie	14628	9604	0,038	15185	15764	16365	16989	17636	18308	19006	19730	20482	21262
Benin	Atlantique	Ze	Yokpo	9173	4989	0,056	9683	10222	10791	11391	12024	12693	13399	14145	14932	15762
Benin	Littoral	Cotonou	1er Arrondissement	57962	55413	0,004	58194	58427	58661	58896	59132	59369	59607	59846	60085	60326
Benin	Littoral	Cotonou	2ème Arrondissement	61668	53708	0,012	62430	63202	63983	64774	65575	66385	67206	68036	68877	69729
Benin	Littoral	Cotonou	3ème Arrondissement	69991	59830	0,014	70974	71970	72981	74005	75045	76098	77167	78250	79349	80463
Benin	Littoral	Cotonou	4ème Arrondissement	36357	39012	-0,006	36130	35904	35680	35457	35236	35016	34797	34580	34364	34149
Benin	Littoral	Cotonou	5ème Arrondissement	20039	32864	-0,043	19177	18352	17562	16807	16084	15392	14730	14096	13490	12909
Benin	Littoral	Cotonou	6ème Arrondissement	75336	71085	0,005	75726	76118	76512	76908	77306	77706	78108	78513	78919	79328
Benin	Littoral	Cotonou	7ème Arrondissement	27535	36158	-0,024	26876	26233	25606	24993	24395	23811	23242	22685	22143	21613
Benin	Littoral	Cotonou	8ème Arrondissement	32420	37631	-0,013	31993	31572	31157	30747	30342	29943	29548	29160	28776	28397

Benin	Littoral	Cotonou	9ème Arrondissement	57691	61585	-0,006	57357	57025	56695	56367	56040	55716	55393	55073	54754	54437
Benin	Littoral	Cotonou	10ème Arrondissement	38728	41806	-0,007	38466	38205	37946	37689	37434	37180	36928	36678	36430	36183
Benin	Littoral	Cotonou	11ème Arrondissement	34879	36219	-0,003	34762	34646	34530	34415	34299	34185	34070	33956	33843	33730
Benin	Littoral	Cotonou	12ème Arrondissement	97920	76217	0,023	100125	102380	104686	107044	109455	111920	114441	117018	119654	122349
Benin	Littoral	Cotonou	13ème Arrondissement	68486	63572	0,007	68941	69399	69859	70323	70790	71260	71733	72210	72689	73172
Benin	Mono	Bopa	Bopa	11496	9206	0,020	11725	11959	12198	12441	12689	12942	13200	13463	13732	14006
Benin	Mono	Bopa	Possotome	7782	6889	0,011	7867	7952	8039	8127	8215	8305	8395	8487	8579	8673
Benin	Mono	Come	Agatogbo	13126	9758	0,027	13477	13836	14206	14585	14975	15375	15785	16207	16640	17084
Benin	Mono	Come	Ouedeme-Pedah	6784	6179	0,008	6841	6898	6955	7013	7072	7131	7190	7250	7310	7371
Benin	Mono	Houeyogbe	Dahe	19536	14622	0,026	20046	20569	21105	21656	22221	22801	23395	24006	24632	25275
Benin	Oueme	Adjarra	Adjarra i	12226	8651	0,031	12608	13001	13407	13826	14258	14703	15162	15635	16123	16627
Benin	Oueme	Adjarra	Adjarra Ii	11906	7604	0,041	12390	12894	13418	13964	14531	15122	15737	16377	17043	17736
Benin	Oueme	Adjarra	Aglogbe	11850	6759	0,051	12456	13094	13764	14468	15209	15987	16805	17665	18569	19519
Benin	Oueme	Adjarra	Honvie	17938	11635	0,039	18642	19373	20133	20923	21744	22597	23483	24404	25362	26357
Benin	Oueme	Adjarra	Malanhoui	22072	11504	0,060	23388	24783	26261	27827	29486	31244	33107	35082	37174	39390
Benin	Oueme	Adjarra	Mededjonou	21432	13959	0,039	22265	23130	24028	24961	25931	26939	27985	29072	30202	31375
Benin	Oueme	Aguegues	Avagbodji	12335	8668	0,032	12728	13133	13552	13984	14429	14889	15363	15852	16357	16879
Benin	Oueme	Aguegues	Houedome	14782	8309	0,053	15559	16376	17236	18142	19095	20099	21155	22266	23436	24667
Benin	Oueme	Aguegues	Zoungame	17445	9673	0,054	18384	19373	20416	21515	22672	23893	25178	26533	27961	29466

Benin	Oueme	Akpro-Misserete	Akpro-Misserete	41657	22491	0,056	44003	46481	49099	51864	54784	57870	61129	64571	68207	72049
Benin	Oueme	Akpro-Misserete	Gome-Sota	15345	8483	0,054	16175	17050	17973	18945	19970	21050	22189	23389	24655	25989
Benin	Oueme	Akpro-Misserete	Katagon	17860	12173	0,035	18479	19120	19782	20468	21177	21912	22671	23457	24270	25111
Benin	Oueme	Akpro-Misserete	Vakon	38806	20541	0,058	41064	43453	45980	48655	51486	54481	57651	61005	64554	68309
Benin	Oueme	Akpro-Misserete	Zoungbome	13581	8964	0,038	14092	14622	15172	15743	16335	16950	17587	18249	18935	19648
Benin	Oueme	Avrankou	Atchoukpa	35232	19565	0,054	37123	39116	41215	43428	45759	48215	50803	53530	56403	59431
Benin	Oueme	Avrankou	Avrankou	20326	13734	0,035	21047	21793	22566	23366	24195	25053	25941	26861	27814	28800
Benin	Oueme	Avrankou	Djomon	21920	14245	0,039	22776	23666	24590	25550	26548	27585	28662	29781	30944	32153
Benin	Oueme	Avrankou	Gbozounme	9949	6057	0,045	10398	10867	11357	11869	12404	12964	13548	14159	14798	15465
Benin	Oueme	Avrankou	Kouty	18312	12751	0,033	18911	19529	20168	20827	21508	22211	22937	23687	24462	25262
Benin	Oueme	Avrankou	Ouanho	15034	8712	0,050	15781	16565	17389	18253	19160	20112	21111	22160	23262	24418
Benin	Oueme	Avrankou	Sado	7277	5338	0,028	7480	7689	7904	8125	8351	8585	8824	9071	9324	9585
Benin	Oueme	Dangbo	Dangbo	12838	8408	0,038	13330	13841	14372	14923	15495	16089	16706	17346	18011	18702
Benin	Oueme	Dangbo	Dekin	8880	6895	0,023	9082	9289	9500	9716	9937	10163	10394	10630	10872	11119
Benin	Oueme	Dangbo	Gbeko	14861	10324	0,033	15350	15855	16377	16916	17473	18048	18642	19255	19889	20543
Benin	Oueme	Dangbo	Houedomey	17507	12224	0,032	18075	18661	19267	19892	20537	21204	21892	22602	23335	24092
Benin	Oueme	Dangbo	Hozin	16327	10076	0,044	17043	17790	18570	19384	20233	21120	22046	23013	24022	25075
Benin	Oueme	Dangbo	Kessounou	13609	9802	0,030	14012	14427	14854	15293	15746	16212	16692	17186	17694	18218
Benin	Oueme	Dangbo	Zoungue	12404	8326	0,036	12851	13315	13795	14293	14808	15342	15896	16469	17063	17679
Benin	Oueme	Porto-Novo	1er Arrondissement	33161	33830	-0,002	33102	33043	32985	32926	32868	32810	32751	32693	32635	32577
Benin	Oueme	Porto-Novo	2ème Arrondissement	52571	45035	0,014	53299	54037	54785	55544	56313	57093	57884	58685	59498	60322

Benin	Oueme	Porto-Novo	3ème Arrondissement	33535	31680	0,005	33705	33876	34048	34220	34394	34568	34744	34920	35097	35275
Benin	Oueme	Porto-Novo	4ème Arrondissement	63306	57311	0,009	63868	64436	65008	65585	66168	66756	67349	67947	68550	69159
Benin	Oueme	Porto-Novo	5ème Arrondissement	81747	55696	0,035	84583	87518	90555	93697	96948	100311	103792	107393	111119	114975
Benin	Oueme	Seme-Kpodji	Agblangandan	57762	30716	0,058	61097	64625	68357	72304	76479	80895	85567	90507	95734	101262
Benin	Oueme	Seme-Kpodji	Aholouyeme	13218	8844	0,036	13699	14197	14713	15248	15803	16377	16973	17590	18230	18893
Benin	Oueme	Seme-Kpodji	Djeregbe	20462	10527	0,061	21707	23028	24430	25917	27494	29167	30942	32825	34823	36942
Benin	Oueme	Seme-Kpodji	Ekpe	75313	34917	0,071	80639	86341	92447	98984	105984	113479	121503	130096	139295	149146
Benin	Oueme	Seme-Kpodji	Seme-Kpodji	23636	12582	0,058	24998	26440	27964	29576	31280	33084	34991	37008	39141	41397
Benin	Oueme	Seme-Kpodji	Tohoue	32310	17652	0,055	34094	35976	37962	40058	42269	44603	47065	49663	52405	55298
Benin	Plateau	Ifangni	Ko-Koumolou	13751	9030	0,038	14275	14819	15383	15969	16577	17209	17864	18545	19251	19984
Benin	Plateau	Ifangni	Tchaada	11078	7336	0,037	11491	11920	12365	12826	13305	13802	14317	14851	15405	15980

Sources : Calculs de l'auteur à partir des décomptes de population issus du RGPH3 et RGPH4 (INSTAD)

Bibliographie

Africapolis, 2024, "Africapolis", https://africapolis.org/fr/country-report/Madagascar

ALEXANDRE V., 2025, "Combien de personnes vivent à Mayotte ? Sur l'île, une bataille de chiffres entre l'Insee et les acteurs locaux", *leparisien.fr*. https://www.leparisien.fr/societe/combien-de-personnes-vivent-a-mayotte-sur-lile-une-bataille-de-chiffres-entre-linsee-et-les-acteurs-locaux-12-02-2025-J6FY6HKURZCUXIK7RC6MJQ5XNU.php

ARTADJI A., LIPOVAC L., HASINA ANDRIAMANANTENA H., ROUSSE B., 2024, "Can we estimate sub-Saharan Africa's population from remote sensing images and land cover mapping?", 186-194 in: J. Cardi, M. Favrot, B. Gastineau, D. Genin, V. Golaz, & C. Robles (Éd.), *Digressions*. Les Impromptus du LPED, Marseille, France, Laboratoire Population Environnement Développement (LPED). https://lped.fr/?LesImpromptusDuLped8Digressions

BARDY C., 2023, "Pourquoi Mayotte n'est pas prise en compte dans les chiffres de la population de la France", *Ouest-France.fr*. https://www.ouest-france.fr/mayotte/population-de-la-france-toujours-indiquee-hors-mayotte-une-exception-qui-ne-va-pas-durer-f7d81608-a64d-11ee-bd70-78fa15e16d04

BARDY P. RECUEILLIS PAR C., 2025, "Entretien. Cyclone Chido : la population de Mayotte est-elle vraiment sous-estimée ?", *Ouest-France.fr*. https://www.ouest-france.fr/mayotte/entretien-cyclone-chido-la-population-de-mayotte-est-elle-vraiment-sous-estimee-af32c870-c777-11ef-892d-e6c5d734e1e4

BIDOU J.-É., DROY I., FAUROUX E., 2008, "Communes et régions à Madagascar. De nouveaux acteurs dans la gestion locale de l'environnement", *Mondes en développement*, Vol.141, N°1, 29-46.

CHAMBERLAIN H. R., DARIN E., ADEWOLE W. A., JOCHEM W. C., LAZAR A. N., TATEM A. J., 2024, "Building footprint data for countries in Africa: To what extent are existing data products comparable?", *Computers, Environment and Urban Systems*, Vol.110, 102104.

CHECCHI F., STEWART B. T., PALMER J. J., GRUNDY C., 2013, "Validity and feasibility of a satellite imagery-based method for rapid estimation of displaced populations", *International Journal of Health Geographics*, Vol.12, 4.

CICG, 2023, "Restructuration des quartiers précaires : le gouvernement va changer le cadre de vie de milliers de familles", *GOUV.CI*. http://www.gouv.ci/_actualite-article.php?recordID=15617

CNLEGIS, 2018, "Fokontany des communes de Madagascar", https://www.cnlegis.gov.mg/uploads/ANNEXE-L2018-011.pdf

DARIN E., KUEPIE M., BASSINGA H., BOO G., TATEM A. J., 2022, "La population vue du ciel : quand l'imagerie satellite vient au secours du recensement", *Population*, Vol.77, N°3, 467-494.

DESEILLE L., 2025, "VRAI OU FAUX. « On est un demi-million » : le nombre d'habitants à Mayotte est-il sous-estimé, comme l'affirment des responsables politiques ?", *Franceinfo*. https://www.francetvinfo.fr/vrai-ou-fake/vrai-ou-faux-le-nombre-d-habitants-a-mayotte-est-il-sous-estime-comme-l-affirment-certains-responsables-politiques_6999398.html

EICHER C. L., BREWER C. A., 2001, "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation", *Cartography and Geographic Information Science*, Vol.28, N°2, 125-138.

FICK S. E., HIJMANS R. J., 2017, "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas", *International Journal of Climatology*, Vol.37, N°12, 4302-4315.

GAUGHAN A. E., STEVENS F. R., LINARD C., JIA P., TATEM A. J., 2013, "High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015", *PLOS ONE*, Vol.8, N°2, e55882.

GENDREAU F., 2019, "Le recensement de population en Afrique, une opération encore problématique"

GENDREAU F., DACKAM-NGATCHOU R., 2024, "Histoire des recensements de la population en Afrique", https://extranet.puq.ca/media/produits/documents/4480_9782760559844.pdf#page=330&zoom=100,0,0

GIRAUD T., 2024, *mapsf: Thematic Cartography*. https://riatelab.github.io/mapsf/

GOLAZ V., 2009, *Pression démographique et changement social au Kenya: vivre en pays gusii à la fin du XXe siècle*. Paris Nairobi, Karthala IFRA.

Google Maps, 2024, "CR ANDRAGNANIVO", *Google Maps*. https://www.google.com/maps/place/CR+ANDRAGNANIVO/@25.0492556,45.5940929,21z/data=!4m14!1m7!3m6!1s0x21d5d3ed4c07aa0f:0xe6bad3d27bbb55cf!2sCR+ANDRAGNANIVO!8m2!3d-25.0492829!4d45.5941636!16s%2Fg%2F11kq2zx52c!3m5!1s0x21d5d3ed4c07aa0f:0xe6bad3d27bbb55cf!8m2!3d-25.0492829!4d45.5941636!16s%2Fg%2F11kq2zx52c?entry=ttu

HAKLAY M., 2010, "How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets", *Environment and Planning B: Planning and Design*, Vol.37, N°4, 682-703.

HALLOT E., GRIPPA T., STEPHENNE N., WOLFF É., 2019, "Cartographie détaillée de la densité de population : comparaison de méthodes dasymétriques", *Dynamiques régionales*, Vol.8, N°2, 35-56.

HIJMANS R. J., ETTEN J. VAN, SUMNER M., CHENG J., BASTON D., BEVAN A., ET AL., 2023, "raster: Geographic Data Analysis and Modeling", https://cran.r-project.org/web/packages/raster/index.html

HILLSON R., ALEJANDRE J. D., JACOBSEN K. H., ANSUMANA R., BOCKARIE A. S., BANGURA U., ET AL., 2014, "Methods for Determining the Uncertainty of Population Estimates Derived from Satellite Imagery and Limited Survey Data: A Case Study of Bo City, Sierra Leone", *PLOS ONE*, Vol.9, N°11, e112241.

INS, 2021, "Distribution de la population par département et sous-préfecture - RGPH-2021, Résultats Globaux", https://www.ins.ci/RGPH2021/RGPH2021-RESULTATS%20GLOBAUX_VF.pdf

INS, 2022, "Recensement général de la population et de l'habitat - résultats globaux définitifs", https://www.ins.ci/RGPH2021/RESULTATS%20DEFINITIFSRP21.pdf

INSAE, 2004a, "Cahier des villages et quartiers de ville - Département de l'OUEME", https://instad.bj/statistiques/enquetes-et-recensements

INSAE, 2004b, "Cahier des villages et quartiers de ville - Département du LITTORAL", https://instad.bj/statistiques/enquetes-et-recensements

INSAE, 2004c, "Cahier des villages et quartiers de ville - Département de l'ATLANTIQUE", https://instad.bj/statistiques/enquetes-et-recensements

INSAE, 2013, "plaquette_RGPH4.docx", https://instad.bj/statistiques/enquetes-et-recensements

INSAE, 2015a, "Résultats définitifs RGPH4", https://rgph5.instad.bj/wp-content/uploads/2023/03/Resultats_definitifs_RGPH4.pdf

INSAE, 2015b, "RGPH4 : Que retenir des effectifs de population en 2013",

INSEE, 2015, "Chapitre 1, Les unités urbaines - Les zonages d'étude de l'Insee - Insee Méthodes n° 109",

INSEE, 2022, "Recensement de la population 2022 - Mayotte",

INSEE, 2023, "À Mayotte, un recensement adapté à une population aux évolutions hors normes", https://blog.insee.fr/mayotte-recensement-adapte-a-population-hors-norme/

INSTAT, 2021a, "Rapport thématique sur les résultats du RGPH-3, Thème 09 : migration à Madagascar"

INSTAT, 2021b, "Résultats globaux du recensement général de la population et de l'habitation de 2018 à Madagascar (RGPH-3)", Tome 1

INSTAT, 2021c, "Rapport thématique sur les résultats du RGPH-3, Thème 01 : état et structure de la population à Madagascar"

INSTAT, 2021d, "Résultats globaux du recensement général de la population et de l'habitation de 2018 à Madagascar (RGPH-3)", Tome 2

INSTAT, 2021e, "Erratas des rapports sur les résultats globaux du RGPH-3, Tome 1 et Tome 2"

INSTAT, 2021f, "Rapport thématique sur les résultats du RGPH-3, Thème 05 : Habitation et cadre de vie de la population"

INSTAT, 2021g, "Rapport thématique sur les résultats du RGPH-3, Thème 04 : Caractéristiques des ménages et structure familiale à Madagascar"

INSTAT, 2021h, "Population et Démographie après l'année 2018 : Combien sommes-nous actuellement ? Quelle tendance jusqu'en 2050"

KEILMAN N., 2008, "European Demographic Forecasts Have Not Become More Accurate over the Past 25 Years", *Population and Development Review*, Vol.34, N°1, 137-153.

KOUAKOU E., 2019, "Habitat et les espaces habités abidjanais, transformations et réappropriations de l'habitat : faut-il repenser le modèle d'habitat dans la ville ?. Architecture, aménagement de l'espace.", https://dumas.ccsd.cnrs.fr/dumas-03132436/file/M1820194438_ADJOUMANIKouakouEric.pdf

LEASURE D. R., JOCHEM W. C., WEBER E. M., SEAMAN V., TATEM A. J., 2020, "National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty", *Proceedings of the National Academy of Sciences*, Vol.117, N°39, 24173-24179.

LINARD C., GILBERT M., SNOW R. W., NOOR A. M., TATEM A. J., 2012, "Population Distribution, Settlement Patterns and Accessibility across Africa in 2010", *PLoS ONE*, Vol.7, N°2, e31743.

LIPOVAC L., 2020, "Estimation de la population par images satellites, Rapport de stage de Master 1 de Mathématiques Appliquées et Sciences Sociales, Analyse des Populations (MASS), Aix-Marseille Université - Diginove, 42 pages",

LIPOVAC L., 2021, "Détection du bâti et modèles d'estimation démographiques de la population des zones géographiques africaines, une analyse de la région d'Abuja, capitale du Nigéria, Rapport d'alternance de 2ème année du master MASS, Aix-Marseille Université – Diginove, 55 pages.",

LLOYD C. T., CHAMBERLAIN H., KERR D., YETMAN G., PISTOLESI L., STEVENS F. R., ET AL., 2019, "Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets", *Big Earth Data*, Vol.3, N°2, 108-139.

LO C. P., 1995, "Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach", *International Journal of Remote Sensing*, Vol.16, N°1, 17-34.

LUTZ W., SANDERSON W., SCHERBOV S., 2001, "The end of world population growth", *Nature*, Vol.412, N°6846, 543-545.

MAGRIN G., DUBRESSON A., NINOT O., BOISSIERE A., 2022, *Atlas de l'Afrique - Un continent émergent ?*

Mairie Attecoube, 2024, "Présentation de l'urbanisation d'Attecoube", https://mairieattecoube.ci/urbanisation

MANOU-SAVINA A., 1989, "Eléments pour une histoire de la cour commune en milieu urbain : réflexions sur le cas ivoirien",

MAXIMENNE A., DJEGO J., LOUGBEGNON T., SINSIN B., 2017, "Typologie Et Répartition Des Espaces Verts Publics Dans Le Grand Nokoué (Sud Bénin)", *European Scientific Journal, ESJ*, Vol.13, 79.

MENNIS J., 2003, "Generating Surface Models of Population Using Dasymetric Mapping", *The Professional Geographer*, Vol.55, N°1, 31-42.

METZGER N., VARGAS-MUÑOZ J. E., DAUDT R. C., KELLENBERGER B., WHELAN T. T.-T., OFLI F., ET AL., 2022, "Fine-grained population mapping from coarse census counts and open geodata", *Scientific Reports*, Vol.12, N°1, 20085.

MUNOZ J. E. V., SRIVASTAVA S., TUIA D., FALCAO A. X., 2021, "OpenStreetMap: Challenges and Opportunities in Machine Learning and Remote Sensing", *IEEE Geoscience and Remote Sensing Magazine*, Vol.9, N°1, 184-199.

Nations Unies, 2020, *Principes et recommandations concernant les recensements de la population et des logements - troisième révision*. UN. https://www.un-ilibrary.org/content/books/9789210051460

NEAL I., SETH S., WATMOUGH G., DIALLO M. S., 2022, "Census-independent population estimation using representation learning", *Scientific Reports*, Vol.12, N°1, 5185.

OCDE, CLUB DU SAHEL ET DE L'AFRIQUE DE L'OUEST., 2020, *Dynamiques de l'urbanisation africaine 2020: Africapolis, une nouvelle géographie urbaine*. OECD. https://www.oecd-ilibrary.org/development/dynamiques-de-l-urbanisation-africaine-2020_481c7f49-fr

OCHA, 2018a, "Madagascar - Subnational Administrative Boundaries - Humanitarian Data Exchange", https://data.humdata.org/dataset/cod-ab-mdg

OCHA, 2018b, "Madagascar - Subnational Population Statistics - Humanitarian Data Exchange", https://data.humdata.org/dataset/cod-ps-mdg

ONU-Habitat, 2012a, "Côte d'Ivoire : profil urbain d'Abobo", https://unhabitat.org/sites/default/files/download-manager-files/Cote%20d%20Ivoirie%20-%20Abobo.pdf

ONU-Habitat, 2012b, "Côté d'Ivoire : profil urbain de Treichville", https://unhabitat.org/sites/default/files/download-manager-files/Cote%20d%20Ivoire%20-%20Treichville.pdf

ONU-Habitat, 2012c, "Côte d'Ivoire : profil urbain de Port-Bouët", https://unhabitat.org/sites/default/files/download-manager-files/Cote%20d%20Ivoire%20-%20Port%20Bouet.pdf

ONU-Habitat, 2023, "ONU-Habitat Côte d'Ivoire. Rapport pays | 2023",

PELLETIER F., SPOORENBERG T., 2016, "Séance 2 Aperçu sur les méthodes de projection",

PESARESI M., CORBANE C., REN C., EDWARD N., 2021, "Generalized Vertical Components of built-up areas from global Digital Elevation Models by multi-scale linear regression modelling", *PLOS ONE*, Vol.16, N°2, e0244478.

PISON G., PONIAKINA S., 2024, "Tous les pays du monde (2024)", *Revenu national*.

RAISON J.-P., 1974, "L'Afrique des Hautes-Terres",

République de Madagascar, 1995, "Loi n° 94-001 fixant le nombre, la délimitation, la dénomination et les Chefs- lieux des Collectivités Territoriales Décentralisées avec amendements", https://www.assemblee-nationale.mg/wp-content/uploads/2020/11/Loi-n%C2%B0-94-001fixant-le-nombre-la-d%C3%A9limitation-la-d%C3%A9nomination-et-les-Chefs-lieux-des-Collectiv.pdf

République de Madagascar, 2004, "Loi n°2004-001 relative aux Régions et de la mise en œuvre de la décentralisation", https://faolex.fao.org/docs/pdf/mad142823.pdf

République de Madagascar, 2014, "Loi N°2014 - 020 Relative aux ressources des Collectivités territoriales décentralisées, aux modalités d'élections, ainsi qu'à l'organisation, au fonctionnement et aux attributions de leurs organes", https://faolex.fao.org/docs/pdf/mad146083.pdf

République de Madagascar, 2015, "Loi N°2015-002 complétant l'annexe n°01 de la loi n° 2014-020 du 27 septembre 2014 relative aux ressources des Collectivités territoriales décentralisées, aux modalités d'élections, ainsi qu'à l'organisation, au fonctionnement et aux attributions de leurs organes.", https://www.ceni-madagascar.mg/wp-content/uploads/2016/05/Loi_2015-_002.pdf

RUNFOLA D., ANDERSON A., BAIER H., CRITTENDEN M., DOWKER E., FUHRIG S., ET AL., 2020, "geoBoundaries: A global database of political administrative boundaries", *PLOS ONE*, Vol.15, N°4, e0231866.

SALENSON I., 2020, "IV. L'Afrique de demain sera rurbaine", *Repères*, 57-76.

SEGUIN S., GRANJON M., THIBAULT P., 2023, "À Mayotte, un recensement adapté à une population aux évolutions hors normes", https://blog.insee.fr/mayotte-recensement-adapte-a-population-hors-norme/

SIRKO W., BREMPONG E. A., MARCOS J. T. C., ANNKAH A., KORME A., HASSEN M. A., ET AL., 2024, "High-Resolution Building and Road Detection from Sentinel-2", http://arxiv.org/abs/2310.11622

STECK J.-F., 2008, "Yopougon, Yop city, Poy... périphérie et modèle urbain ivoirien", *Autrepart*, Vol.47, N°3, 227-244.

STEVENS F. R., GAUGHAN A. E., LINARD C., TATEM A. J., 2015, "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data", *PLOS ONE*, Vol.10, N°2, e0107042.

SUTTON P., ROBERTS D., ELVIDGE C., BAUGH K., 2001, "Census from Heaven: An estimate of the global human population using night-time satellite imagery", *International Journal of Remote Sensing*, Vol.22, N°16, 3061-3076.

TABUTIN D., SCHOUMAKER B., 2020, "La démographie de l'Afrique subsaharienne au XXIe siècle : Bilan des changements de 2000 à 2020, perspectives et défis d'ici 2050", *Population*, Vol.75, N°2, 169-295.

TATEM A. J., NOOR A. M., HAGEN C. VON, GREGORIO A. D., HAY S. I., 2007, "High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa", *PLOS ONE*, Vol.2, N°12, e1298.

THIBAULT P., 2019, "Quatre logements sur dix sont en tôle en 2017", https://www.insee.fr/fr/statistiques/4202864

THOMSON D. R., LEASURE D. R., BIRD T., TZAVIDIS N., TATEM A. J., 2022, "How accurate are WorldPop-Global-Unconstrained gridded population data at the cell-level?: A simulation analysis in urban Namibia", *PLOS ONE*, Vol.17, N°7, e0271504.

THOMSON D. R., RHODA D. A., TATEM A. J., CASTRO M. C., 2020, "Gridded population survey sampling: a systematic scoping review of the field and

strategic research agenda", *International Journal of Health Geographics*, Vol.19, N°1, 34.

TRENDS, 2020, "Ne rayer personne de la carte - guide des données démographiques maillées pour le développement durable", https://static1.squarespace.com/static/5b4f63e14eddec374f416232/t/5eb9d7ee15dc887a04523d52/1589237775980/Ne+rayer+personne+de+la+carte+-FRENCH.pdf

VALLIN J., MESLE F., TOULEMON L., VERON J., 2011, "Projections de population", 384-386 in: *Dictionnaire de démographie et des sciences de la population*.

VIGNES C., RIMBOURG S., 2013, "Méthodes statistiques d'allocation spatiale : interpolation de données surfaciques",

WARDROP N. A., JOCHEM W. C., BIRD T. J., CHAMBERLAIN H. R., CLARKE D., KERR D., ET AL., 2018, "Spatially disaggregated population estimates in the absence of national population and housing census data", *Proceedings of the National Academy of Sciences*, Vol.115, N°14, 3529-3537.

WATTELAR C., CASELLI G., VALLIN J., WUNSCH G., 2004, "Perspectives démographiques : historique de la méthode et méthodes actuelles", 253-284 in: *Démographie : analyse et synthèse, Vol.5 (Histoire du peuplement et prévisions)*.

WEBER E. M., SEAMAN V. Y., STEWART R. N., BIRD T. J., TATEM A. J., MCKEE J. J., ET AL., 2018, "Census-independent population mapping in northern Nigeria", *Remote Sensing of Environment*, Vol.204, 786-798.

WICKHAM H., 2016, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

World Population Prospect, 2024, "World Population Prospects - Population Division - United Nations", https://population.un.org/wpp/

WorldPop, 2024, "About us", *WorldPop*. https://www.worldpop.org/about/

Sources provenant de sites internet :

Mindat, 2024a, "Talata-Tsimadilo", https://www.mindat.org/feature-1055988.html

Mindat, 2024b, "Ambohidanerana", https://www.mindat.org/feature-1080565.html

Mindat, 2024c, "Antanambe", https://www.mindat.org/feature-1071012.html

Mindat, 2024d, "Ambolobozobe", https://www.mindat.org/feature-1079527.html

Mindat, 2024e, "Befotaka", https://www.mindat.org/feature-1068312.html

Mindat, 2024f, "Antsahabe_1", https://www.mindat.org/feature-1069922.html

Mindat, 2024g, "Antsahabe_2", https://www.mindat.org/feature-1069924.html

Mindat, 2024h, "Antsirasira", https://www.mindat.org/feature-1069101.html

Mindat, 2024i, "Manaratsandry", https://www.mindat.org/feature-1061269.html

Mindat, 2024j, "Antsapanana", https://www.mindat.org/feature-11937356.html

Mindat, 2024k, "Andranambomaro", https://www.mindat.org/feature-1075537.html

Mindat, 2024l, "Ambodinonoka", https://www.mindat.org/feature-1081132.html

Mindat, 2024m, "Anjoma Faliarivo", https://www.mindat.org/feature-11937381.html

Mindat, 2024n, "Morarano Soafiraisana", https://www.mindat.org/feature-11941599.html

Mindat, 2024o, "Andranomiely Atsimo", https://www.mindat.org/feature-1074911.html

Mindat, 2024p, "Mangasoavina", https://www.mindat.org/feature-1060847.html

Mindat, 2024q, "Andoharano", https://www.mindat.org/feature-11941933.html

Mindat, 2024r, "Mikaikarivo Ambatomainty", https://www.mindat.org/feature-1082869.html

Mindat, 2024s, "Ambatotsivala", https://www.mindat.org/feature-1082539.html

Mindat, 2024t, "Maheny", https://www.mindat.org/feature-1062019.html

Mindat, 2024u, "Ankilivalo Sud", https://www.mindat.org/feature-1072091.html

Mindat, 2024v, "Behazomanga", https://www.mindat.org/feature-1068194.html

Wikipedia, 2021a, "Andranomisa", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Andranomisa&oldid=1053670893

Wikipedia, 2021b, "Makaraingo", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Makaraingo&oldid=1054628271

Wikipedia, 2021c, "Ifasina III", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ifasina_III&oldid=1056396293

Wikipedia, 2022a, "Anosy Avaratra", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Anosy_Avaratra&oldid=1070395807

Wikipedia, 2022b, "Antranokoaky", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Antranokoaky&oldid=1077888257

Wikipedia, 2022c, "Ambohidrabiby", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ambohidrabiby&oldid=1116675539

Wikipedia, 2023a, "Satrandroy", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Satrandroy&oldid=1163051238

Wikipedia, 2023b, "Sirama", *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Sirama&oldid=1184407427

Wikipedia, 2023c, "Antsakoabe", *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=Antsakoabe&oldid=210449318

La Nouvelle Chaîne Ivoirienne, 2020, *Cour commune : un symbole du vivre-ensemble*. https://www.youtube.com/watch?v=oP_Swcc2Rno

OCHA Centre for Humanitarian Data, 2020, *HDX Dataset Deep Dive: WorldPop Gridded Population Datasets*. https://www.youtube.com/watch?v=A1AvguSj41Q

Institut français des relations internationales, 2024, *Gouverner à la marge des villes : l'impact de la croissance des villes sur les périphéries urbaines*. https://www.youtube.com/watch?v=or1FAw9oJVw&t=952s