Faculté de géographie et d'aménagement
Université de Strasbourg

MASTER 1

Observation de la Terre et Géomatique

Travail d'Étude et de Recherche

2024-2025

LEDERMANN Quentin

Sujet du TER : Construire un modèle de cartographie dynamique des populations par machine learning

Encadrants : Kenji FUJIKI
Laboratoire Image, Ville, Environnement
UMR7362 CNRS-Unistra
3 rue de l'Argonne F-67000 Strasbourg

Romain WENGER
Laboratoire Image, Ville, Environnement
UMR7362 CNRS-Unistra
3 rue de l'Argonne F-67000 Strasbourg

Laboratoire image, ville, environnement | LIVE de l'Université de Strasbourg et du CNRS

Table des matières

Table des illustrations

Table des annexes

Remerciements

Je tiens tout d'abord à remercier mes encadrants, Kenji Fujiki et Romain Wenger, pour leur accompagnement tout au long de ce mémoire. Leur disponibilité, leurs conseils méthodologiques avisés et leur exigence scientifique ont grandement contribué à structurer ma réflexion et à approfondir mes choix analytiques.

Je remercie également l'ensemble de l'équipe pédagogique du Master OTG de l'Université de Strasbourg, pour la qualité de leur enseignement et la richesse des échanges au fil de cette année.

Un grand merci à mes proches, amis et collègues de promotion, pour leur soutien constant, leurs relectures bienveillantes et les discussions qui ont nourri ce travail, parfois bien au-delà des pages qui suivent.

Enfin, ce mémoire est aussi le fruit de nombreuses heures d'exploration, d'erreurs, d'apprentissage autonome, et de doutes surmontés. Il marque une étape importante dans mon parcours, et je suis reconnaissant d'avoir pu le mener jusqu'à son terme.

1. Introduction

1.1. Contexte scientifique : comprendre les limites de la représentation statique de la population

Dans un monde où les données numériques sont omniprésentes, comprendre la répartition des populations ne devrait plus relever du défi. Pourtant, derrière l'apparente abondance des chiffres, se cache une réalité géographique difficile à saisir : les individus se déplacent, se concentrent, s'éloignent et la majorité des outils de mesure continuent de les figer dans un lieu de résidence (Panczak et al., 2020). Ce décalage entre les outils statistiques et les rythmes de la vie urbaine est désormais reconnu, « *la population résidente est une fiction utile, mais inadéquate pour comprendre les dynamiques urbaines et contemporaines* » (Batista e Silva et al., 2020). À l'heure où les crises urbaines exigent des réponses agiles et localisées, une cartographie réellement dynamique des populations devient un impératif scientifique et opérationnel.

1.1.1. La prédominance des données de résidence

Cartographier la population est une tâche ancienne de la géographie, mais comprendre où se trouvent réellement les individus à chaque instant de la journée reste aujourd'hui encore un défi scientifique majeur. Actuellement, la donnée de population la plus facilement accessible, c'est le recensement. La quasi-totalité des politiques publiques (aménagement, mobilité, sanitaire…) et des analyses spatiales en France, s'appuient sur les données INSEE, et en particulier le recensement. Comme son nom l'indique, le recensement mesure la population résidente, soit la population présente sur un territoire généralement la nuit, au lieu d'habitation principal, sans tenir compte des dynamiques temporelles. Par exemple, un quartier résidentiel apparaît densément peuplé sur les données du recensement, mais il peut être quasiment vide en journée, tandis qu'un quartier d'affaires aura une forte densité diurne non-visible sur les données de recensement (Boeing, 2018). Cette citation, *« Mapping population at residence fails to capture the geography of urban life. »* (Batista e Silva et al., 2020), résume bien le fait que les données de recensement échouent à capturer les dynamiques de population. Cette représentation biaisée est problématique, notamment pour les gestions des flux (transports, énergies, services publics), la planification urbaine (équipements, sécurité) et la gestion de crise (évacuation, Covid-19). Cependant, il existe une donnée nous permettant de connaître les déplacements des populations : la téléphonie mobile. Ces données ont montré une bien meilleure captation des mouvements quotidiens, mais ne sont pas accessibles pour la recherche libre (Deville et al., 2014). Ainsi, si la population résidente constitue un socle statistique essentiel pour les politiques publiques, elle offre une vision figée et partielle du territoire, incapable de rendre compte des dynamiques quotidiennes qui façonnent l'espace urbain.

1.1.2. Le besoin d'approches temporelles et spatialisées

L'espace urbain est fondamentalement dynamique : sa fréquentation varie selon l'heure, le jour, la saison ou le type d'activités. Cette variabilité est structurante pour les usages (transports, commerces, services…), mais elle demeure largement invisible dans les données statistiques traditionnelles. Pour qualifier cette instabilité, les géographes parlent de temporalités différenciées de l'espace (Lussault, 2007 ; Vallée & Lenormand, 2024), soulignant que les lieux se transforment en permanence selon les rythmes de vie sociale. Face à cette complexité, des solutions alternatives aux données de recensement sont nécessaires. Une piste prometteuse, bien qu'imparfaite, est offerte par le Mobiliscope. Cette plateforme publique repose sur des enquêtes de mobilité menées par le CEREMA et l'INSEE, disponibles dans plusieurs grandes agglomérations françaises (Vallée et al., 2024). Elle permet de reconstituer les déplacements individuels sur 24 heures et de modéliser les présences dans l'espace à une maille fine, à partir de typologies socio-spatiales. Si ces données sont ouvertes et précieuses, elles présentent certaines limites : elles peuvent être datées, agrégées à des niveaux peu fins, et nécessitent des interpolations pour être spatialisées uniformément. Néanmoins, leur existence souligne un besoin croissant d'indicateurs dynamiques, capables de mieux outiller les politiques publiques face aux enjeux contemporains (mobilité, planification, climat, crises). L'essor de la smart city et des outils géonumériques renforce cette exigence : produire des données de présence en temps quasi-réel, à partir de sources ouvertes, transparentes et reproductibles (Goodchild, 2013 ; Panczak et al., 2020), devient un impératif scientifique et opérationnel.

1.2. Problématique : modéliser la présence humaine avec des données indirectes

Formuler une problématique ne revient pas seulement à poser une question, c'est choisir un angle, définir une trajectoire de recherche, et circonscrire un objet complexe. Dans le cas présent, il s'agit d'interroger notre capacité à estimer la présence humaine dans l'espace urbain en s'appuyant uniquement sur des données indirectes et accessibles.

1.2.1. Formulation du problème de recherche

Face à l'indisponibilité de données directes sur les mobilités individuelles à grande échelle, une alternative émerge. Elle consiste à modéliser la présence humaine à partir de données ouvertes, indirectes et hétérogènes. Ces données ne donnent pas un accès direct à la population, mais à des indices de sa présence potentielle : densité d'établissements, morphologie urbaine, accessibilité, etc. En mobilisant ces variables, il devient possible d'estimer la population présente, notamment en journée, à une échelle fine. Dans ce contexte, la question centrale de ce mémoire est la suivante :

Comment modéliser de manière fine la distribution spatio-temporelle de la population en mobilisant des données hétérogènes et des méthodes d'apprentissage automatique ?

Répondre à cette question, implique d'identifier les variables spatiales les plus explicatives de la densité dynamique, choisir une méthode d'apprentissage (interprétable et efficace) et de proposer un protocole reproductible, généralisable à d'autres territoires

1.2.2. Défis liés à l'absence de données directes

Si l'approche par modélisation indirecte s'avère prometteuse, elle suppose de relever plusieurs défis théoriques, méthodologiques et pratiques. En effet, ces données ne mesurent pas directement la présence humaine en tant que telle, mais des éléments susceptibles de la favoriser (ex : morphologie urbaine, services…). Pour cela on fait l'hypothèse que certains espaces sont « *plus susceptibles d'être fréquentés* » en raison de leur composition fonctionnelle ou physique. C'est une démarche de désagrégation indirecte, qui repose sur une inférence spatiale fragile. Panczak et al. (2020) insistent sur les risques de surestimation ou d'interprétation erronée des proxies. De plus, les données indirectes sont proposées à des échelles différentes et formats variés. Elles nécessitent donc des traitements de nettoyages, d'agrégation ou de pondération. Pour ce qui est du Mobiliscope, il est basé sur une modélisation indirecte de la population présente, via des enquêtes de mobilité. Il attribue à chaque secteur typologique une courbe horaire de population. Cette construction introduit des biais (dépendance à la typologie initiale, homogénéisation des comportements) qui limitent la précision ; le Mobiliscope reste cependant une base fiable et ouverte. Ces défis renforcent la nécessité d'une démarche explicite et documentée, d'un choix rigoureux des variables explicatives et d'une validation méthodique du modèle final. Ils rappellent aussi que le but n'est pas d'atteindre une vérité absolue, mais de proposer une approximation robuste, compréhensible et utile à la décision territoriale.

1.3. Objectifs du mémoire : construire, modéliser, interpréter

Toute recherche ne se réduit pas à une question, elle implique des objectifs concrets, traduisibles en méthodes, en tests et en résultats. À partir de la problématique formulée précédemment, ce mémoire poursuit un ensemble d'objectifs opérationnels, scientifiques et techniques articulés autour de la construction d'un modèle reproductible de la population dynamique.

1.3.1. Objectif principal

L'objectif central de ce mémoire est de concevoir, mettre en œuvre et évaluer un modèle géostatistique ou machine learning, capable de prédire la densité de population présente à une heure donnée en s'appuyant sur des données indirectes, une variable cible dérivée du Mobiliscope et une grille spatiale fine. Il en convient aussi de répondre à trois enjeux. Spatiale premièrement, en produisant un résultat géographiquement fin, lisible, intégré à une représentation cartographique continue du territoire. Ensuite méthodologique, en testant un modèle robuste mais interprétable, afin de ne pas recourir à une « *boîte noire* »

algorithmique. Et enfin opérationnel, en proposant une méthode réutilisable par des chercheurs, des collectivités dans d'autres contextes territoriaux. Ce travail s'inscrit dans la lignée de projets comme les travaux de Batista e Silva et al. (2020), qui combinent données de bâtiments et points d'intérêts pour prédire la densité en Europe. Ou encore Cheng et al. (2022), qui comparent plusieurs algorithmes sur des proxies spatiaux. Ce mémoire reprend ces logiques mais les adapte à un contexte français. En résumé, l'objectif est de développer un modèle spatialement explicatif, basé sur des données ouvertes, pour estimer la répartition de la population, dans une logique de transparence, de reproductibilité et de transférabilité territoriale.

1.3.2. Objectifs secondaires

Un des objectifs secondaires est de créer un pipeline automatisé, où chaque étape de traitement est isolée dans un module ou script dédié. Il doit permettre de réexécuter l'ensemble de la chaîne sur un autre territoire ou à une autre échelle, de tester différentes configurations et de documenter les choix techniques. Il s'appuie sur des bibliothèques open source telles que : `geopandas`, `scikit-learn`, `numpy`, `requests`, etc. La transférabilité du projet repose aussi sur son adaptabilité à plusieurs échelles. Un autre objectif est d'identifier les prédicteurs les plus explicatifs. Pour cela des analyses statistiques bivariées et multivariées nous permettent de connaître la qualité de chaque variable. Au-delà des performances chiffrées, l'objectif est aussi d'interroger la robustesse de la méthode. Quelles variables sont sensibles aux effets de bords ? Que se passe-t-il dans les zones peu couvertes en données ? Cette évaluation vise à préparer une discussion technique dans les parties dédiées, mais aussi d'ouvrir une perspective de transposabilité, en proposant une méthodologie généralisable à d'autres contextes et territoires.

1.4. Hypothèses formulées : relations attendues entre formes urbaines et densités mobiles

Formuler des hypothèses revient à anticiper des relations entre variables, à tester des intuitions théoriques par des moyens empiriques. Dans le cadre de ce mémoire, les hypothèses permettent de guider la construction du modèle en identifiant les facteurs spatiaux susceptibles d'expliquer la répartition de la population présente, et en anticipant la manière dont le modèle devrait réagir à ces entrées.

1.4.1. Hypothèses sur les déterminants spatiaux

A l'issue des objectifs évoqués, il convient de formuler des hypothèses sur les variables expliquant la répartition de la population. La première étant : **la densité d'emplois estimée est un facteur explicatif fort de la présence humaine diurne**. En effet, on peut supposer que plus un secteur accueille d'emplois, plus il est susceptible de voir sa population augmenter en journée. L'objectif est de tester si cette variable est positivement corrélée avec la densité observée dans le Mobiliscope. Boeing (2018) montre que la distribution des lieux d'emplois est une variable clé dans l'estimation de la

population jour. En second, on peut émettre que : **le score des points d'intérêts reflète le potentiel d'attractivité fonctionnelle**. Les points d'intérêts (POI) recensés via OpenStreetMap (OSM) indiquent la présence de commerces, services, équipements, qui génèrent des flux humains. Une forte densité de POI implique une probabilité plus forte d'occupation spatiale. Sun et al. (2024) démontrent la validité de l'usage des POI pour estimer la population présente, en lien avec la fréquentation urbaine. Enfin : **la compacité bâtie est proxy d'intensité fonctionnelle**. La compacité désigne ici le ratio surface bâtie sur la surface de la maille. Un espace très compact tend à correspondre à un centre urbain ou un pôle structurant, cette variable est liée à la capacité à concentrer de la population de jour.

1.4.2. Hypothèses sur la performance du modèle

Modéliser la distribution de la population présente à partir de données indirectes ne repose pas seulement sur le choix des variables, mais aussi sur la capacité du modèle à en extraire une structure explicative pertinente. Le choix d'un modèle simple et interprétable, comme la régression linéaire multiple, engage plusieurs hypothèses sur la qualité des résultats, leur stabilité et leur robustesse. La première étant : **la régression linéaire multiple peut expliquer une part significative de la variance**. Un modèle linéaire bien construit, avec un jeu de variables thématiquement cohérent, peut produire un coefficient de détermination (R^2) supérieur à 0,60. La majorité des travaux existants obtiennent des R^2 compris entre 0,4 et 0,7 en utilisant des modèles linéaires simples (Cheng et al., 2022). Cette méthode permet une interprétation directe du poids des variables, essentielle pour discuter des logiques spatiales. Ces hypothèses sur la performance permettent d'encadrer les attentes du modèle. Elles sont également indispensables pour anticiper les limites, les marges de progression.

Ce mémoire propose donc une démarche originale consistant à entraîner un modèle de prédiction à l'échelle nationale, en mobilisant l'ensemble des secteurs issus du Mobiliscope comme base d'apprentissage supervisé. Une fois validé, ce modèle est appliqué localement sur un territoire plus restreint et à une résolution spatiale choisie. Cette stratégie permet de tester à la fois la portabilité géographique d'un modèle fondé uniquement sur des données ouvertes, et sa capacité à produire une estimation fine et continue de la population à une échelle infra-urbaine.

Dans cette perspective, la première partie du mémoire propose un état de l'art des notions, données et méthodes existantes relatives à la population dynamique. La deuxième partie détaille la construction du pipeline de données et la génération des variables explicatives. La troisième est consacrée à l'entraînement, l'évaluation et l'interprétation du modèle. Enfin, la dernière partie discute les apports, les limites et les perspectives de cette démarche exploratoire.

Afin de fonder théoriquement cette démarche, il convient d'abord d'examiner les concepts, les données et les méthodes existants relatifs à la population dynamique et à sa cartographie, en mobilisant les travaux récents dans le champ de la géographie quantitative et des sciences spatiales.

2. État de l'art – Fondements conceptuels et méthodologiques

2.1. Clarification du sujet : définitions de population dynamique, cartographique et machine learning

Modéliser la présence humaine à partir de données indirectes suppose de s'appuyer sur des concepts solides, des données accessibles et des méthodes éprouvées. Avant d'entrer dans la construction du modèle, il est donc nécessaire de clarifier ce que l'on entend par « population dynamique », d'identifier les sources mobilisables pour en estimer la répartition, et d'examiner les approches méthodologiques existantes. Cette partie propose ainsi un état des connaissances à l'intersection de la géographie, de la géomatique et de l'apprentissage automatique.

2.1.1. Notions de population dynamique, temporaire, fonctionnelle

Parler de « *cartographie dynamique des populations* » implique d'abord de clarifier ce que recouvrent les termes de population présente, temporaire ou fonctionnelle. Ces notions renvoient à des réalités différentes de la population résidente, et leur compréhension est essentielle pour fonder théoriquement le modèle développé dans ce mémoire. Pour la population résidente, elle représente un référent statistique mais figé. Définie par l'INSEE comme la population ayant sa résidence principale dans une commune, la population résidente est celle qui figure dans les bases du recensement et les zonages administratifs (INSEE, 2016). Elle structure de nombreux dispositifs publics (allocations, dotations scolaires, documents d'urbanisme), mais reflète une vision nocturne et statique du territoire (Freire, 2010). Sa limite majeure est qu'elle ne prend pas en compte des déplacements quotidiens, ni l'usage réel des espaces pendant la journée ou selon les temporalités saisonnières. Cette critique est récurrente sur les mobilités urbaines (Batista e Silva et al., 2020 ; Boeing, 2018). Une autre notion est celle de la population temporaire, avec une présence discontinue ou irrégulière. Elle inclut notamment les touristes (Batista e Silva et al., 2018), les travailleurs saisonniers ou navetteurs (Panczak et al., 2020) et les étudiants ou personnes en transition (Xuacho et al., 2019). Cette population échappe souvent aux fichiers administratifs, et nécessite des sources alternatives (données téléphoniques, imagerie satellites, POI). Deville et al. (2014) ont montré que les données de téléphonie mobile pouvaient estimer finement ces présences, mais leur usage reste limité en recherche publique. Une troisième notion mobilisable est celle de la population fonctionnelle, qui reflète une lecture territorialisée des usages. Elle est utilisée en aménagement, elle désigne les personnes ayant une activité dans un espace (travail, étude, achats), même si elles n'y résident pas. Cette notion permet de définir des aires urbaines fonctionnelles (INSEE, Eurostat), fondées sur

les bassins de déplacements ou les flux pendulaires. Elle repose sur l'idée d'ancrage fonctionnel dans un territoire, et offre une base pour penser les services publics à l'échelle des usages (Williams et al., 2012). Ces différentes acceptions de la population révèlent que la présence humaine dans l'espace urbain est plurielle, mobile et contextuelle. Dans le cadre de ce mémoire, c'est la population présente à une heure donnée, telle que modélisée par le Mobiliscope, qui sera utilisée comme variable cible, en assumant son caractère modélisé, agrégé et non résiduel.

2.1.2. Cartographie et modélisation en géographie

Cartographier la population mobile ne consiste plus à localiser des individus à leur domicile, mais à représenter leurs présences variables dans l'espace et le temps. Cette évolution implique de passer d'une cartographie descriptive à une cartographie analytique et prédictive. La cartographie classique repose sur la population résidente, décrite par commune ou IRIS à partir des données de recensement. Elle ne permet pas de saisir les rythmes urbains ou la fréquentation horaire. Des outils comme le Mobiliscope (Vallée et al., 2024) proposent désormais des représentations horaires continues, basées sur des enquêtes de mobilité. En géographie quantitative, cette carte permet de repérer les corrélations spatiales, de détecter des effets de seuil ou de discontinuités, et d'évaluer des modèles (cartes des résidus, des erreurs). Cheng et al. (2022) insistent sur la nécessité de croiser visualisation et métriques dans l'évaluation des modèles de populations. Plusieurs travaux modélisent la population à partir de données spatiales ouvertes : POI (Sun et al., 2024), morphologie bâtie (Bilijecki & Chow, 2022), typologie fonctionnelle (Batista e Silva et al., 2020). Ces modèles reposent sur l'idée que la forme et la fonction de l'espace urbain conditionnent sa fréquentation. La carte n'est pas seulement un outil de communication, mais un outil critique de validation. Elle permet le repérage des zones sur ou sous modélisées et c'est un appui à la lecture des biais. La cartographie appliquée à la population dynamique n'est plus une simple projection de données, mais un moyen d'anticiper, de comparer et d'interroger spatialement les résultats de la modélisation.

2.1.3. Machine learning et apprentissage supervisé en sciences spatiales

Les sciences spatiales intègrent de plus en plus d'outils de machine learning (ML), non seulement pour prédire des phénomènes, mais aussi pour en décrypter les structures explicatives. Dans le cadre de la cartographie dynamique des populations, l'apprentissage supervisé offre un cadre méthodologique puissant mais exigeant. L'objectif de l'apprentissage supervisé est de prédire une variable cible (ici la population présente) à partir d'un ensemble de variables explicatives connues. Il repose sur un jeu de données étiqueté, ici ce sont les variables spatiales (POI, densité bâtie, hauteur…) et la variable cible issue du Mobiliscope. Il permet notamment d'évaluer les performances du modèle (R^2, RMSE, validation croisée) et d'analyser la contribution de chaque variable. L'intérêt d'utiliser le machine learning en géographie est de pouvoir gérer des données hétérogènes et corrélées. Et cela part son adaptabilité à différentes échelles

interprétables ou complexes. Cheng et al. (2022) montrent l'intérêt du machine learning pour prédire la population mobile avec des données spatiales croisées. Cependant, il faut effectuer un choix entre simplicité et puissance. Des modèles simples comme la régression linéaire multiple permettent une interprétation directe. Mais d'autres bien plus puissant, comme XGBoost et Random Forest, présentent de meilleures performances mais avec une perte d'explicabilité. Xuacho et al. (2019) combinent POI et imagerie satellitaire avec des modèles complexes pour cartographier la population en Chine. De son côté, Sun et al. (2024) obtiennent de bons résultats en combinant POI et imagerie nocturne. L'apprentissage en contexte spatiale met en lumière des problèmes spécifiques, notamment l'importance de cartographier les erreurs pour identifier les biais spatiaux (Batista e Silva et al., 2020). Le machine learning offre un cadre méthodologique souple et puissant pour modéliser la population dynamique, à condition de respecter les contraintes spatiales, de choisir des modèles interprétables et de valider les résultats dans une logique géographique.

2.2. Limites des données directes : INSEE, téléphonie, Mobiliscope

Comprendre la dynamique de la population dans l'espace urbain suppose d'accéder à des données capables de refléter les mobilités quotidiennes, les rythmes horaires et les usages territoriaux. Plusieurs sources de données dites « *directes* » ont été mobilisées dans la recherche récente pour approcher cette réalité : les données de recensement, les données issues de la téléphonie mobile, ou encore les modèles comme le Mobiliscope. Chacune présente des avantages en termes de couverture, de fiabilité ou de résolution, mais aussi des limites importantes qui justifient le recours à des approches fondées sur des proxies spatiaux.

2.2.1. Forces et faiblesses des données de recensement

Le recensement de la population constitue la base de données sociodémographiques de référence en France. Il alimente la quasi-totalité des analyses de population utilisées dans les politiques publiques. Pourtant, s'il présente de nombreux atouts en matière de fiabilité et de couverture, il montre également d'importantes limites dès lors qu'il s'agit de représenter la mobilité ou la population présente dans les territoires. Produit par l'INSEE, le recensement repose sur un protocole rigoureux, stabilisé et légalement encadré. Il couvre l'ensemble du territoire national avec une granularité descendante jusqu'aux IRIS. Il est répété régulièrement (tous les 5 ans), ce qui fait un repère fiable pour les comparaisons temporelles, et pour de nombreux zonages fonctionnels : bassins de vie, unités urbaines, EPCI. C'est une donnée centrée sur la résidence principale. Le recensement mesure la population résidente, c'est-à-dire attachée administrativement à un lieu de vie nocturne. Il renseigne les lieux de travail et de scolarisation, mais à une échelle relativement grossière (communale). Mais pas les déplacements quotidiens, ni la fréquentation diurne, ni les temporalités saisonnières. Cela pose certaines limites pour le dimensionnement des services de jour (transports, restaurations, soins), l'analyse des

risques et la modélisation réelle de l'espace. Les données du recensement sont agrégées à l'échelle communale ou IRIS, mais depuis l'année dernière ces données commencent à être disponibles sur un carroyage de 1 kilomètre. Ce niveau est inadapté à la cartographie dynamique, pour exemple une commune peut contenir à la fois une zone d'habitat, une zone d'activité et une friche, sans distinction. Cela complique l'usage de ces données comme variable cible pour des modèles à maille fine. La fréquence de mise à jour du recensement (tous les 5 ans) ne permet pas de saisir les effets de crises (Covid-19, télétravail, tourisme ponctuel), les transformations rapides de certains territoires (ZAC, centralités émergentes). Freire (2010) souligne l'obsolescence rapide des bases administratives pour la modélisation des présences. Le recensement reste un indicateur fondamental de la structure démographique, mais il se révèle mal adapté à l'analyse des dynamiques spatio-temporelles fines. Ces limites justifient le recours à des données plus réactives, ou à des modèles indirects comme celui développé dans ce mémoire.

2.2.2. Contraintes d'accès et d'exploitation des données de téléphonie

Avec l'essor des données massives, les données issues de la téléphonie mobile (utilisation du réseau et utilisation des applications) sont apparues comme une source prometteuse pour cartographier la population présente en temps réel. Leur potentiel est largement reconnu dans les recherches sur la mobilité, les flux urbains et la gestion de crise. Pourtant, leur exploitation dans un cadre académique ou public reste encadrée, incomplète ou inaccessible, ce qui en limite fortement l'usage pour une modélisation ouverte et reproductible. Ces données permettent de localiser les usagers via leur téléphone, de pouvoir suivre les flux horaires ou journaliers à grande échelle et de cartographier la population présente avec une granularité spatiale et temporelle inédite. Cependant, elles souffrent d'une accessibilité restreinte et opaque. Elles sont contrôlées par des acteurs privés (Orange, Google, Meta), et difficilement accessibles à la recherche publique. Ou alors moyennant, des contrats payants et fournis sous forme agrégée ou partielle. Goodchild (2013) alerte sur le manque de transparence et les risques de dépendances. Comme les données ouvertes, les données de téléphonie ont des limites techniques et géographiques. Le biais de couverture (zones mal desservies, téléphones inactifs), le géoréférencement flou ou encore les données non conçues pour l'analyse territoriale. Malgré leur richesse, les données de téléphonie ne répondent pas aux critères de reproductibilité, d'ouverture et de compatibilité géographique. Elles ne constituent donc pas une base pertinente pour un modèle accessible, reproductible comme celui visé par ce mémoire. Bien que ce mémoire exclut leur usage pour des raisons d'accessibilité et de reproductibilité, les données directes (téléphonie mobile, GPS) pourraient, à terme, enrichir les prédictions dans un cadre contrôlé, en complétant les approches par proxies.

2.2.3. Apports et limites du Mobiliscope

Entre les données classiques du recensement et les données privées de téléphonie, le Mobiliscope s'impose comme une solution intermédiaire, en étant une donnée ouverte, accessible et spécifiquement pensée pour représenter la population présente à l'échelle

infra-urbaine. Il constitue la base centrale du présent mémoire. Comme explicité en amont, le Mobiliscope repose sur des Enquêtes Mobilité Certifiées Cerema (EMDC), représentatives à l'échelle des agglomérations françaises. Il modélise les déplacements d'individus sur une journée type et restitue la population présente heure par heure sur des secteurs spatiaux typologiques. Ses forces résident dans sa documentation complète, des données libres et accessibles, le pas horaire fin et l'intégration de profils socio-démographiques (CSP, âge, sexe). Vallée et al. (2024) valorisent l'usage du Mobiliscope pour analyser les rythmes urbains quotidiens. Les estimations sont découpées en un maillage sectoriel typologique, pas toujours compatible avec une modélisation fine. Les données sont fournies par secteur urbain typé, qui résultent d'un clustering géographique (densité, forme urbaine, occupation du sol), mais pas d'un carroyage homogène. Cela nécessite une interpolation spatiale pour les intégrer dans un modèle régulier, ce qui induit un risque de perte de précision locale lors du croisement spatial. La base du Mobiliscope étant des enquêtes, et non des mesures directes. Il est construit autour d'hypothèses telles que la population a un comportement stable, et sur une journée type. Comme toute modélisation, elle génère des incertitudes et ne reflète pas les variations contextuelles. Le Mobiliscope constitue une base précieuse, ouverte et bien documentée pour approcher la population mobile. Sa structure sectorielle et son statut modélisé imposent toutefois une interprétation prudente, et justifient l'usage d'un modèle complémentaire visant à prédire ces densités sur une grille spatiale continue.

2.3. Données ouvertes comme proxy potentiel et critiques

Face aux limites des données directes, qu'elles soient administratives ou privées, de nombreuses recherches se tournent vers l'exploitation de données ouvertes spatiales pour estimer indirectement la présence humaine dans l'espace. Ces données ne mesurent pas la population, mais décrivent l'environnement construit, fonctionnel ou économique susceptible d'attirer ou d'accueillir des individus. OpenStreetMap, la base de données topographique (BD TOPO), la base SIRENE ou encore les POI sont ainsi mobilisés comme proxies, c'est-à-dire comme variables indirectes permettant de prédire la densité de présence. Cette approche, encore récente, soulève des questions cruciales sur la qualité des données, leur potentiel explicatif, mais aussi sur leurs limites structurelles et méthodologiques.

2.3.1. Sources disponibles et caractéristiques techniques

Le recours aux données ouvertes pour estimer la population dynamique repose sur un principe simple, celui d'observer les structures et fonctions de l'espace (bâti, services, attracteurs) afin d'en déduire les logiques de présence humaines. Plusieurs bases sont mobilisables dans ce cadre, chacune présentant des avantages spécifiques et des contraintes techniques. Premièrement, les données OpenStreetMap reposent sur une base de données collaborative mondiale, OSM fournit des points d'intérêts (POI), des voies, des bâtiments, des équipements, etc. Ces données ont comme avantages, une

très large couverture, une mise à jour fréquente dans les zones urbaines denses, et la possibilité de filtrer les objets par thématique (ex : commerces, éducation, santé). Mais présentent aussi avec des limites comme la qualité variable selon le territoire et l'absence de métadonnées (ex : nombres d'usagers POI). Sun et al. (2024) montrent que les POI OSM sont corrélés avec la fréquentation diurne dans les zones commerciales. Ensuite, fournie par l'INSEE, la base SIRENE recense les établissements actifs en France avec leur activité (code NAF), localisation et effectif. Elle peut être utile pour construire des variables liées à l'emploi, à la fonction économique et à la typologie d'activités. Mais avec certaines contraintes comme les effectifs souvent imprécis ou absents, et la géolocalisation parfois au siège social plutôt que sur le site actif. On trouve aussi, la base de données topographique produite par l'IGN, qui fournit une cartographie vectorielle des bâtiments, avec des attributs comme la hauteur, la surface au sol ou le type d'usage. Elle permet d'estimer la densité continue, la compacité morphologique et la capacité d'accueil potentielle d'un secteur. Bilijecki & Chow (2022) exploitent la hauteur moyenne pour prédire la densité humaine à Singapour. Enfin, la Base Permanente des Équipements (BPE) décrit la présence d'équipements dans les domaines scolaire, sanitaire, sportif, commercial. Avec une couverture nationale et une structure normalisée, elle est utile pour construire des scores de centralité fonctionnelle. Ces données sont disponibles à l'échelle ponctuelle et communale). Ces bases ouvertes constituent les briques essentielles de la modélisation indirecte de la population. Leur combinaison permet d'approximer le potentiel d'attractivité ou d'occupation d'un territoire, à condition d'en maîtriser les spécificités techniques et les biais structurels

2.3.2. Approches indirectes : proxy et indicateurs de présence

Dans un contexte où les données directes de présence humaine sont rares ou inaccessibles, de nombreuses recherches se tournent vers les approches indirectes, basées sur l'utilisation de variables spatiales servant de proxies de la population. Ces indicateurs ne mesurent pas les individus, mais les conditions favorables à leur présence ou leur fréquentation. Un proxy est une variable indirecte supposée corrélée à un phénomène non observable directement (ici : la densité humaine). En modélisation spatiale, on utilise des informations structurelles ou fonctionnelles comme substituts, comme la concentration de POI qui donne l'attractivité commerciale. La densité d'emplois qui donne la fréquentation diurne et la hauteur de bâti qui donne la capacité d'accueil verticale. Batista e Silva et al. (2020) montrent que la combinaison de POI, réseau de transport et morphologie urbaine permet d'estimer la population active à l'échelle européenne. Dans son étude, Boeing (2018) démontre comment l'intégration de données de recensement et de statistique d'emploi peut fournir une estimation précise de la densité de population diurne à l'échelle locale. La Figure 1, illustre la concentration de la population active dans les zones urbaines centrales de la baie de San Francisco, soulignant l'importance de considérer l'utilisation de données de recensement et d'indicateurs d'emploi.

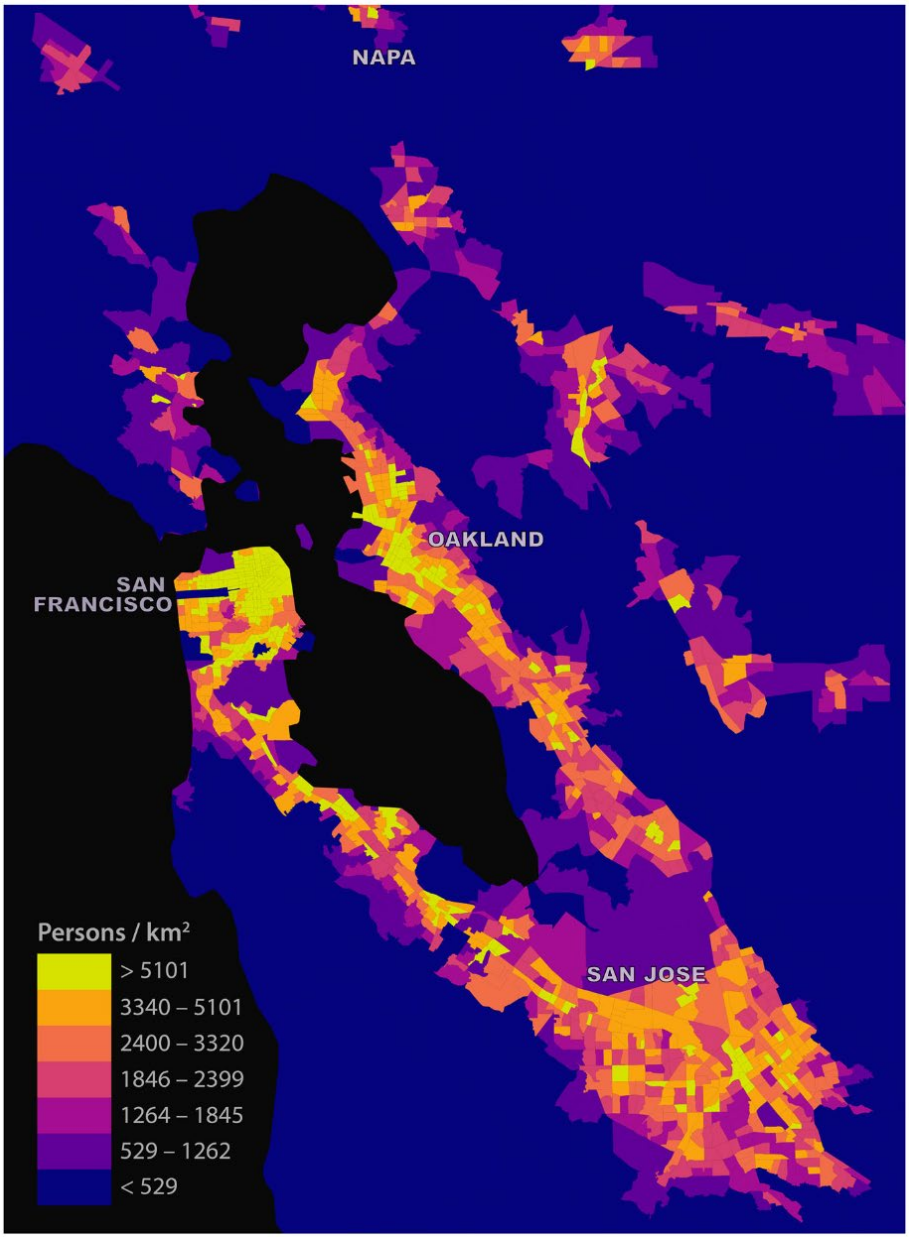


Figure 1: Estimation de la densité de population diurne, dans la baie de San Francisco. (Boeing, 2018)

On distingue généralement trois grands types de proxies. Les proxies fonctionnels, comme le nombre de commerces, établissements scolaires ou services publics (BPE, SIRENE). Les proxies morphologiques : la densité au centre-ville, la proximité des pôles d'attraction (OSM). Les proxies synthétiques : les indicateurs fonctionnels et/ou morphologiques peuvent être pondérés ou combinés (ex : scores POI pondérées par typologie). Pour ce qui est des limites méthodologiques, les proxies sont des approximations, ils ne garantissent ni la fréquentation réelle, ni l'intensité d'usage. Ils présentent des risques de biais de surinterprétation, des effets de redondance entre variables corrélées. Avec une certaine fragilité contextuelle, un proxy peut fonctionner dans un centre-ville mais pas en périphérie. Panczak et al. (2020) soulignent l'instabilité

des proxies selon les territoires étudiés. L'usage de proxies constitue un levier méthodologique puissant pour estimer la population dynamique à partir de données accessibles. Mais il exige une sélection rigoureuse, une pondération contextualisée, et une validation empirique attentive.

2.3.3. Limites épistémologiques et méthodologiques

Si l'utilisation de données ouvertes et de proxies spatiaux ouvre de nouvelles perspectives pour modéliser la population dynamique, elle soulève également des limites fondamentales, tant sur le plan des hypothèses que de méthodes de traitement et d'interprétation. Ces limites doivent être explicitement reconnues pour situer la portée réelle du modèle. Premièrement, les variables utilisées ne mesurent pas directement la présence humaine, elles supposent une relation de causalité ou de corrélation, qui peut varier selon les territoires et les temporalités. Exemple : un quartier dense en POI peut être sous-fréquenté à certaines heures ou durant certaines saisons. La validité empirique d'un proxy dépend donc du contexte local (urbain, rural, périurbain), des rythmes sociaux (horaires d'ouverture, attractivité) et des effets d'échelles. Panczak et al. (2020) alertent sur le caractère instable et contextuel des indicateurs indirects. Ensuite, les données sont inégalement distribuées et parfois incomplètes. Les données OSM ont une richesse variable selon la densité de contributeurs. La base de données SIRENE a des effectifs manquants ou déclaratifs. Et la BD TOPO, est parfois imprécise en zone rurale ou périphérique. Cela engendre des zones vides, de surestimation ou d'erreur, difficiles à corriger sans validation de terrain. Xuacho et al. (2019) montrent que certaines couches spatiales faussent la prédiction dans plusieurs régions chinoises. À force d'utiliser des proxies construits à partir d'objets urbains (commerces, équipements…), on risque de créer un modèle qui décrit la structure urbaine plus que la présence humaine réelle. Ce biais structurel peut conduire à une surévaluation des centralités ou à une sous-estimation des zones hybrides. Goodchild (2013) évoque le danger d'une modélisation « *trop alignée sur ce qu'on peut mesurer* », plutôt que sur ce qu'on souhaite comprendre. Ces limites rappellent que le recours aux proxies n'est pas neutre, il repose sur des choix interprétatifs, des données perfectibles, et une relation incertaine entre espace observé et population réelle. Cela justifie une validation rigoureuse du modèle et une lecture prudente des résultats.

2.4. Méthodes de modélisation spatiale : désagrégation, régression, algorithmes supervisés

La modélisation spatiale permet de reconstruire des phénomènes non observés directement, à partir de variables explicatives disponibles. Dans le cas de la population dynamique, elle offre un cadre pour estimer la répartition fine des présences humaines, en s'appuyant sur des méthodes de désagrégation, de régression ou d'apprentissage automatique. Cette section présente les principales approches mobilisées dans la littérature, de la plus simple à la plus complexe.

2.4.1. Désagrégation spatiale par répartition ou interpolation

Les premières méthodes utilisées pour estimer la population à une échelle plus fine que celle des données disponibles relèvent de la désagrégation spatiale, c'est-à-dire de la répartition d'une valeur agrégée sur un maillage plus détaillée à l'aide de règles géographiques ou statistiques. La désagrégation par répartition proportionnelle est la méthode la plus simple. Elle permet de répartir une donnée agrégée (ex : population communale) en fonction d'une variable spatiale de référence (ex : surface bâtie, emprise au sol des bâtiments). Freire (2010) propose une redistribution selon les zones urbanisées détectées par télédétection. Cela a comme avantages d'être une méthode rapide et reproductible, mais avec une très forte sensibilité au choix de la variable pivot et elle ne prend pas en compte la diversité fonctionnelle ou sociale. Ensuite, certaines approches utilisent l'interpolation par grilles ou modèles pondérés, avec des algorithmes d'interpolation spatiale pondérée (poids inversement proportionnel à la distance, krigeage, voisinage (Thiessen)). Ces méthodes supposent une continuité spatiale du phénomène, ce qui n'est pas toujours le cas pour la population. Cheng et al. (2022) comparent plusieurs méthodes d'interpolation dans la prédiction de la population mobile. Il existe aussi des approches mixtes qui combinent les données de recensement, les variables morphologiques (bâti, voirie) et les données fonctionnelles (POI, SIRENE). Cette désagrégation pondérée est le socle des approches machine learning supervisées, car elle permet de créer des jeux de données d'apprentissage. La désagrégation spatiale constitue une étape fondatrice pour projeter des données à une échelle fine, mais elle a ses limites (absence d'incertitude, simplification des logiques de fréquentation), qui justifient le recours à des modèles plus complexes dans les étapes suivantes du mémoire.

2.4.2. Modèles de régression : linéaire, régularisée

Les modèles de régression sont au cœur de nombreuses approches en géographie quantitative. Leur simplicité, leur lisibilité et leur efficacité en font des outils privilégiés pour expliquer la population présente à partir de variables spatiales. La régression linéaire multiple a pour objectif de modéliser une relation linéaire entre une variable cible (ex : population présente) et plusieurs variables explicatives (ex : POI, emploi, densité, bâtie). Avec comme avantages, une forte lisibilité des résultats (coefficient, signe, significativité), une simplicité d'implémentation, adéquation avec des hypothèses explicatives. Mais elle rencontre certaines limites, notamment sa sensibilité aux corrélations croisées (multi colinéarité), la non-linéarité de certaines relations ignorée (Kutner et al. 2004 ; James et al. 2021). Dans la continuité, on trouve la régression régularisée, qui permet de réduire l'impact des variables redondantes ou peu informatives. Elle existe sous deux formes principales. La Ridge, qui pénalise l'amplitude des coefficients, et la Lasso qui sélectionne automatiquement des variables. Zou & Hastie (2005) combinent les deux logiques, ce qui est recommandé lorsque le jeu de variables est large ou corrélé (Kutner et al., 2004). En ce qui concerne l'application à la dynamique de la population, les régressions sont fréquemment utilisées pour estimer la population diurne à partir de proxies (Batista e Silva et al., 2020), tester l'impact d'une

variable spécifique (Sun et al., 2024), et pour construire un modèle explicatif spatialement interprétable. Les modèles de régression offrent une base robuste et transparente pour modéliser la population dynamique. Leur efficacité dépend toutefois du choix des variables, de la gestion des corrélations et de la capacité à en interpréter les résultats dans une logique spatiale.

2.4.3. Arbres décisionnels et méthodes d'ensemble (Random Forest, XGBoost)

Lorsque les relations entre variables deviennent complexes, non linéaires ou interactives, les modèles de régressions classiques atteignent leurs limites. Les arbres décisionnels et les méthodes d'ensemble constituent une alternative performante, de plus en plus utilisée dans les approches géospatiales. Les arbres décisionnels, comme leur nom l'indique, sont constitués d'un arbre de décision qui segmente l'espace des variables selon des seuils successifs pour prédire une valeur cible. Ils ont comme avantages de pouvoir capturer des relations non linéaires, une robustesse aux effets d'échelle et une interprétabilité partielle (visualisation de l'arbre). Ces avantages sont en contrebalancés par leur dépendance aux jeux de données d'entraînement et par le risque afférent de surapprentissage (James et al., 2021 ; Breiman, 2001). Ici nous explorons Random Forest et XGBoost, tous deux reposant sur une méthode d'ensemble, en agrégeant plusieurs arbres construits sur des échantillons variés. Pour Random Forest, il utilise la moyenne de nombreux arbres pour réduire la variance, il est insensible au bruit et il présente un bon compromis entre biais et variance (Breiman, 2001 ; Biau & Scornet, 2016). Quant à XGBoost, il repose sur un boosting séquentiel (chaque arbre corrige les erreurs du précédent) et il obtient d'excellents scores de performance dans la plupart des compétitions de data science (Fernandez-Delgado et al., 2014 ; Zou & Hastie, 2005 ; Cheng et al., 2022). Dans l'usage pour la modélisation de la population, ces méthodes sont utiles pour traiter des jeux de données complexes ou bruités, gérer des variables catégorielles ou très corrélées et aussi estimer des relations spatiales non linéaires (Xuacho et al., 2019 ; Sun et al., 2024). Les arbres décisionnels et les modèles d'ensemble représentent une évolution naturelle des approches explicatives, combinant puissance et flexibilité.

2.5. Synthèse des approches et justification des choix

Les notions, données et méthodes présentées dans les sections précédentes montrent que la modélisation de la population dynamique est un champ en pleine structuration, à la croisée de la géographie quantitative, de la géomatique et de la data science. Cette dernière section de l'état de l'art vise à situer le présent travail dans cet ensemble, en identifiant à la fois ses affiliations méthodologiques et ses spécificités : types de données utilisées, échelle d'analyse, approche reproductible et objectif d'interprétabilité.

2.5.1. Synthèse des approches comparables

Plusieurs travaux récents ont tenté d'estimer la population mobile ou diurne à partir de données indirectes. Ces approches reposent sur des jeux de données hétérogènes, des objectifs variables (cartographie, gestion, simulation) et des modèles allant du plus simple au plus complexe. Premièrement, les modèles basés sur l'imagerie et les POI. Sun et al. (2024) utilise une combinaison de données de POI issues d'OSM et d'images satellites nocturnes pour prédire la densité de population de jour à l'échelle locale. Xuacho et al. (2019) fait une estimation de la population en Chine via fusion multisource (imagerie multispectrale, POI, typologie fonctionnelle), couplée à des modèles de boosting. Ces approches sont efficaces mais souvent opérées dans un contexte « *data-rich* », difficile à reproduire sans accès à des bases propriétaires. De l'autre côté, se trouve les approches européennes de désagrégation spatiale. Batista e Silva et al. (2020), utilise une désagrégation de la population européenne en combinant bâti, POI, réseaux et occupation du sol via des méthodes statistiques. Les résultats de cette étude visible sur la Figure 2, extraite de Batista e Silva et al. (2020), illustre parfaitement les variations de densité de population en Europe par fusion de données multivariées, et donne aussi une idée de la forme que les résultats peuvent prendre.

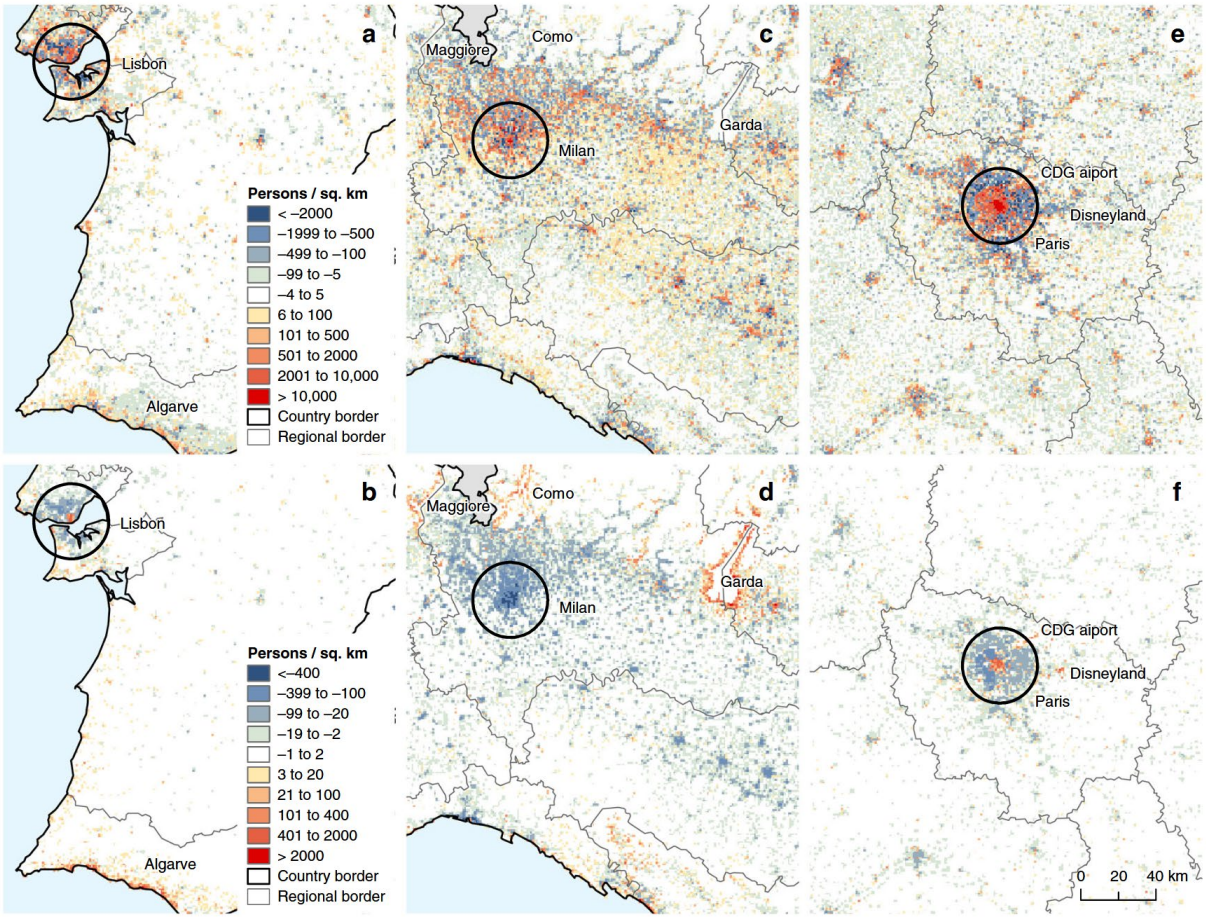


Figure 2: Représentation de la différence jour-nuit de densité de population à Lisbonne, Milan et Paris. (Batista e Silva et al., 2020)

Leurs travaux visent des modèles transférables à grande échelle, mais à une résolution relativement grossière (1 kilomètre). Greger (2015) propose un modèle de répartition intra-urbaine via l'intensité du bâti. Enfin des initiatives reproductibles en open source. Comme le projet DayPop (Fujiki, 2025), un pipeline Python de désagrégation spatiale jour/nuit à partir de données ouvertes (SIRENE, OSM, POI). Le Mobiliscope (Vallée et al., 2024), une plateforme d'exploration des rythmes urbains basée sur des enquêtes de mobilités. Ces approches sont les plus proches du présent travail, en termes de sources, de méthodologie et de finalité exploratoire.

2.5.2. Justification des choix conceptuels et techniques

Le modèle proposé dans ce mémoire repose sur des choix théoriques et techniques assumés, qui viennent compléter certaines approches existantes par leur orientation vers la reproductibilité, la lisibilité géographique et la mobilisation exclusive de données ouvertes. Ce travail adopte une démarche articulée en deux temps : d'abord, l'entraînement et la validation d'un modèle explicatif via régression linéaire multiple et algorithmes de machine learning (Random Forest, XGBoost), puis l'application spatiale des prédictions à différentes mailles. L'apprentissage est réalisé à partir des secteurs Mobiliscope, répartis sur l'ensemble du territoire métropolitain, en utilisant une pondération spatiale des variables explicatives (POI, emplois, morphologie, etc.). Ce choix permet d'accroître la diversité et la robustesse du modèle en intégrant une grande hétérogénéité spatiale. Une fois entraîné, le modèle est appliqué à un carroyage régulier (au choix : 100 m, 200 m ou 500 m) sur un territoire plus restreint. Cette transposition permet de tester la capacité du modèle à produire une estimation fine, localisée et continue de la population présente à un instant donné, en s'affranchissant du découpage sectoriel initial.

L'étude de la population dynamique soulève des enjeux théoriques, méthodologiques et opérationnels majeurs. Les limites des données classiques (recensement, téléphonie), les potentialités mais aussi les incertitudes des approches par proxy, ainsi que la diversité des méthodes de modélisation, dessinent un champ en pleine structuration, à la croisée de la géographie quantitative, de la data science et de la géomatique. Dans ce contexte, le présent mémoire adopte une approche hybride, ouverte et territorialisée, combinant des données hétérogènes, un pipeline reproductible, et une volonté d'interprétation spatiale des résultats.

La section suivante présente la mise en œuvre concrète de cette démarche, en décrivant pas à pas la construction de la base de données, la génération des variables explicatives, la mobilisation du Mobiliscope, et le choix du modèle d'apprentissage.

3. Méthodologie – Mise en œuvre de la modélisation

3.1. Architecture générale du pipeline Python : acquisition, traitement, fusion, modélisation

Ce chapitre expose la méthode employée pour élaborer un modèle prédictif de la population à partir de données ouvertes. L'ensemble du traitement repose sur un pipeline Python modulaire, conçu pour automatiser et structurer les étapes d'acquisition, de traitement spatial, de génération des variables, de modélisation et d'application locale. Si la majorité des traitements a été automatisé en Python via des scripts modulaires (acquisition, prétraitement, génération des variables), certaines opérations ponctuelles ont nécessité un recours à QGIS. C'est notamment le cas de traitements géométriques lourds (fusion de couches, découpage complexe, vérification de topologie) qui se sont avérées trop coûteux ou instables sur Python. Cette approche hybride permet d'optimiser la reproductibilité tout en assurant la stabilité du traitement des couches volumineuses. Chaque section traite en détail des choix techniques et des opérations effectuées, dans le but de garantir la reproductibilité et la lisibilité scientifique.

3.1.1. Logique modulaire du projet

Le projet repose sur une architecture modulaire, pensée pour assurer la clarté, la réplicabilité et la maintenance du code. Chaque étape du traitement est isolée dans un script indépendant, ce qui facilité le débogage, l'évolution des méthodes et la transparence du processus. Le traitement est organisé en modules thématiques, chacun correspondant à une grande phase du pipeline. Chacun de ces types de traitements est exécuté via un script central d'orchestration, qui appelle les sous-modules individuellement, comme expliciter sur la Figure 3.

Schéma technique du pipeline global

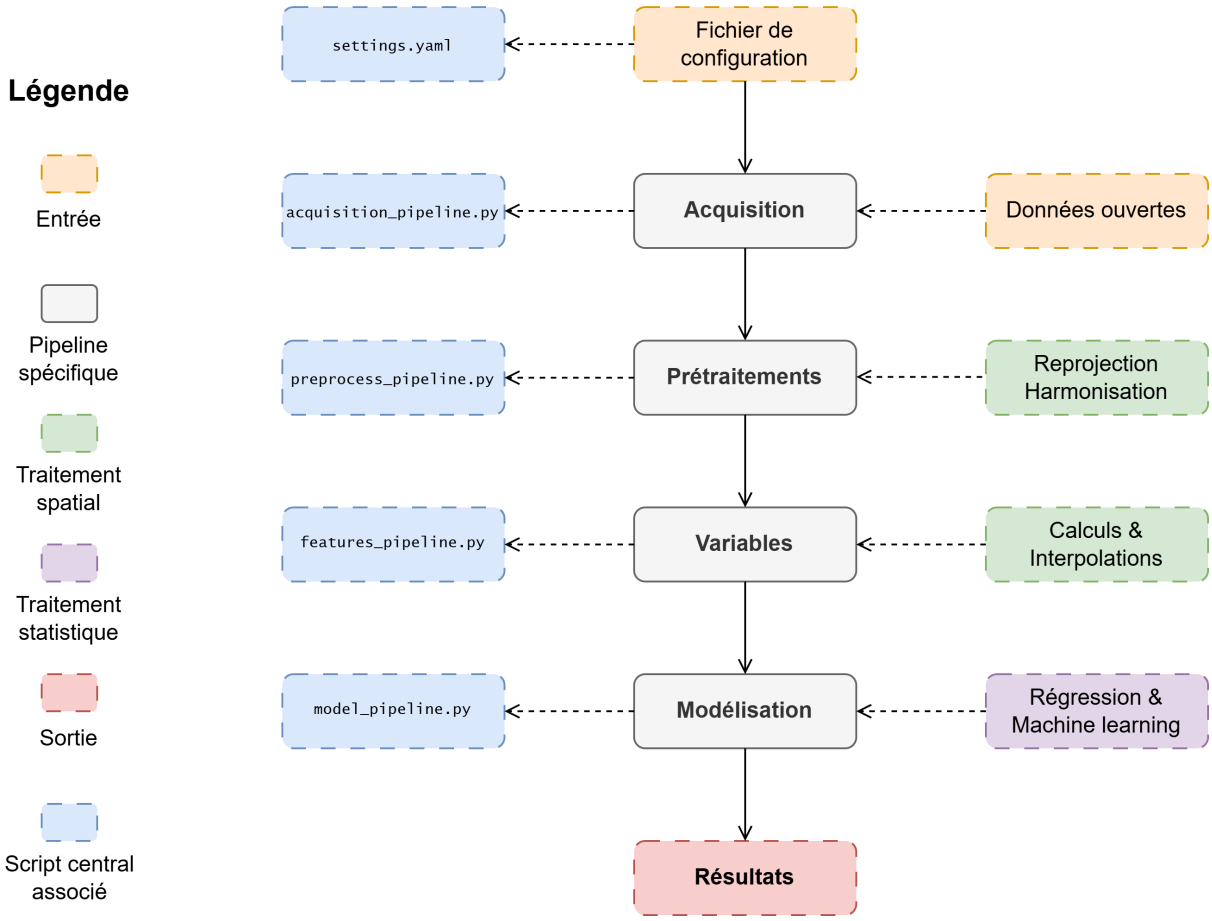


Figure 3: Schéma technique du pipeline global. (Ledermann, 2025)

Les variables globales (ex : maille, territoire ciblé) sont définies dans un fichier unique de configuration, permettant une réutilisation du pipeline sur d'autres territoires. Les résultats intermédiaires (GeoJSON, CSV) sont enregistrés à chaque étape, ce qui facilité la vérification manuelle ou le recalcul partiel. L'arborescence globale du projet est consultable sur le Git-Hub de celui-ci (Annexe 1). Cette organisation modulaire permet de segmenter les responsabilités dans le code, de relancer uniquement certaines étapes et de faciliter la montée en compléxité (ex : ajout d'une nouvelle variable). Cette logique modulaire constitue le socle de la méthodologie : elle garantit à la fois l'efficacité du développement et la traçabilité du traitement, deux exigences clés dans une démarche de modélisation reproductible.

3.1.2. Outils et bibliothèques utilisés

Le traitement des données repose entièrement sur un environnement Python, sélectionné pour sa richesse en bibliothèques géospatiales, sa capacité à automatiser les étapes, et compatibilité avec les formats ouverts. Ce choix garantit une méthode reproductible, évolutive et portable. Le langage principal utilisé est Python 3.12.X, qui offre un écosystème robuste pour le traitement spatial, la manipulation de données et la modélisation statistique. La Figure 4 ci-dessous récapitule les principales bibliothèques mobilisées ainsi que fonctions.

Bibliothèques	
Nom	**Fonctions**
`geopandas`	Manipulation des données spatiales (GeoDataFrame, reprojection, jointures spatiales)
`pandas`	Gestion des tables attributaires, fusion, filtrage
`shapely`	Opérations géométriques (intersections, buffers, union)
`pyproj`	Gestion fine des systèmes de projection (EPSG : 2154 notamment)
`py7zr & zipfile`	Extraction de dossiers compréssés (format .zip et .7z)
`scikit-learn`	Modélisation (régression linéaire, Random Forest, métreiques d'évluation)
`matplotlib & seaborn`	Visualisation (scatterplots, heatmaps, graphiques statistiques)

Figure 4: Tableau des bibliothèques utilisées. (Ledermann, 2025)

L'articulation entre les blocs du pipeline (acquisition, traitement, variables et modèle) suit une logique séquentielle claire, facilitée par l'environnement Python. Le découpage permet de garantir la transparence et la flexibilité de la démarche, tout en favorisant sa réutilisation sur d'autres territoires. Les outils choisis garantissent une chaîne de traitement légère, transparente et portable, sans dépendance à des logiciels propriétaires, conformément aux principes FAIR et à l'esprit open source du projet.

3.2. Données sources utilisées : origines, formats, reprojection, harmonisation

Le modèle développé dans ce mémoire repose exclusivement sur des données ouvertes, sélectionnées pour leur accessibilité, leur richesse spatiale et leur compatibilité technique. Cette section présente les sources mobilisées, leurs formats initiaux, ainsi que les opérations de traitement nécessaires à leur intégration dans le pipeline.

3.2.1. Sources mobilisées et formats initiaux

La modélisation s'appuie sur un ensemble de données hétérogènes, provenant pour la plupart de portails publics nationaux. Chaque source de données apporte une information complémentaire sur l'occupation du sol, la morphologie urbaine ou l'activité économique, servant à alimenter les variables explicatives du modèle. La Figure 5, résume les données utilisées, leurs formats et leur fonction dans la construction des variables explicatives.

Données			
Nom	**Description**	**Format**	**Fonction**
Mobiliscope (UMR Géographie-cités)	Population présente par secteur (jour/heure)	GeoJSON	Variable cible (population présente) pour l'apprentissage
Fichiers fisacles FILOSOFI (INSEE)	Population résidente, âge, densité, logement	CSV / Shapefile	Base démographique, comparaison, pondérations
BD Topographique (IGN)	Polygones de bâtiments avec hauteur et emprise	Shapefile	Calculs morphologiques, densité bâtie, hauteur
Base SIRENE (INSEE)	Établissements, code NAF, tranche d'effectifs	CSV	Estimation d'emplois, mixité fonctionnelle, POI avancées
OpenStreetMap	Points d'intérêts (POI), voirie, typologie fonctionnelle	GeoJSON	Score POI pondérée, densité de commerce
Base BPE (INSEE)	Équipement publics	CSV	Score POI, estimation d'emplois

Figure 5: Tableau des données utilisées. (Ledermann, 2025)

Toutes ces données sont téléchargées, stockées puis traitées localement, les fichiers de données spatiales (GeoJSON et Shapefile), sont tous convertis au format GeoParquet. Le format GeoParquet offre un stockage géospatial compact, rapide et optimisé pour les traitements en série : il permet des lectures/écritures beaucoup plus rapides que les formats traditionnels comme GeoJSON, GPKG ou shapefile tout en garantissant une meilleure compression, une compatibilité cloud, et une intégration native dans les pipelines de data science (Python, Spark, etc.). Cette diversité de source permet de croiser les approches morphologiques, fonctionnelles et démographiques, et de produire un jeu de variables riches fondé sur des ressources ouvertes.

3.2.2. Étapes de traitement et reprojection

Les données brutes utilisées dans ce travail présentent des formats, des systèmes de projection et des périmètres hétérogènes. Avant de pouvoir être exploitées conjointement, elles nécessitent une série d'opérations d'harmonisation spatiale, indispensables à la cohérence du traitement. Pour cette expérimentation sur le territoire français, il est naturel que toutes les données soient converties vers le système de projection Lambert-93 (EPSG : 2154), qui garantit la compatibilité avec les jeux de données IGN et INSEE, et une métrique fiable pour les calculs de surface, de distance ou de densité. Les formats spatialisés (GeoParquet) sont lus via `geopandas`, les CSV géolocalisés sont transformés en GeoDataFrame avec colonnes latitudes/longitudes converties en points projetés. L'harmonisation spatiale garantit une cohérence géométrique et sémantique des jeux de données, condition indispensable à la fiabilité des opérations de jointure et d'analyse spatiale.

3.3. Construction du maillage spatial d'analyse : choix d'échelle, traitement géographique

Le modèle étant entraîné à partir des secteurs Mobiliscope à l'échelle nationale, son application locale nécessite un référentiel spatial régulier, capable de restituer une estimation fine et continue de la population présente. Cette section décrit la création d'un carroyage paramétrable (100 mètres, 200 mètres ou 500 mètres) sur le territoire choisi, utilisé à la fois pour agréger certaines variables et pour projeter les résultats du modèle.

3.3.1. Critères de choix des tailles de maille

Pour projeter localement les prédictions du modèle entraîné, il est nécessaire de disposer d'un maillage spatial homogène, couvrant l'intégralité du département étudié. Le choix d'un carroyage régulier permet de s'affranchir des découpages administratifs, tout en assurant une continuité spatiale. La Figure 6 donne un exemple de carroyage régulier.

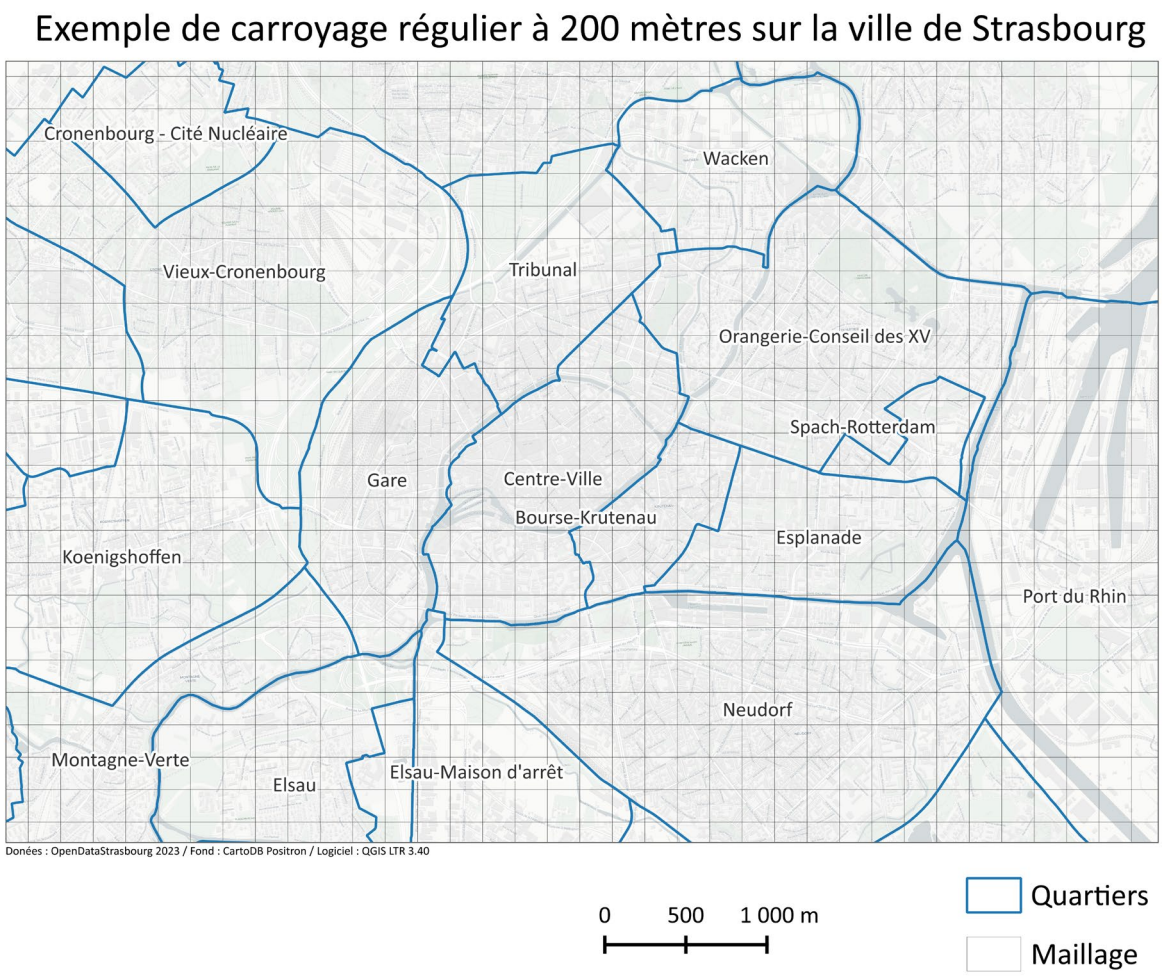


Figure 6: Maillage régulier (200 mètres) sur la ville de Strasbourg. (Ledermann, 2025)

Le modèle est appliqué post-traitements sur un carroyage régulier couvrant toute la zone géographique choisie, qui est lui-même généré à partir d'un script dédié et des informations renseignées dans le fichier de configuration. Même si le pipeline permet de générer différentes tailles de maille, les traitements ont été standardisés sur une maille de 200 mètres pour garantir la stabilité des opérations et la compatibilité avec les données principales. Ce choix d'échelle est également guidé par la spatialisation des données utilisées comme le base fiscale (FILOSOFI) sur un carroyage de 200 mètres, la BD TOPO et les données ponctuelles d'OSM. Enfin, le carroyage est utilisé pour agréger certaines variables spatiales et servir de support pour la génération des fichiers finaux destinés à la cartographie et à l'analyse. L'utilisation d'un carroyage régulier permet de tester la généralisabilité du modèle sur des unités fines, tout en assurant une structure neutre et adaptable à d'autres territoires.

3.3.2. Méthodes d'intersection et de zonage

Une fois le carroyage construit, il doit être croisé avec les autres couches géographiques du projet pour permettre l'agrégation des variables explicatives. Cela nécessite l'usage de méthodes s'intersection spatiale adaptées, qui garantissent une répartition juste et

géométriquement rigoureuse des informations. Le croisement entre le carroyage et les couches sources (bâtiments, secteurs, Mobiliscope, POI, etc.) repose sur plusieurs méthodes. La première est l'intersection géométrique, elle est utilisée pour identifier les objets qui chevauchent une maille et elle est adaptée aux POI, bâtiments ou établissements, dont la présence suffit à affecter la maille. Elle permet de visualiser concrètement les principes de zonage et d'agrégation. Ensuite la méthode d'intersection surfacique, utilisée pour les polygones complexes (ex : secteurs Mobiliscope, zones INSEE), elle permet une pondération au prorata de la surface intersectée, afin d'estimer finement les quantités redistribuées (ex : population). Pour certaines couches (ex : bâtiments) des calculs supplémentaires sont réalisés, comme la surface de bâti par maille ou la hauteur moyenne pondérée par emprise. Enfin le zonage inverse est également utilisé. À partir de chaque maille, on récupère l'ensemble des entités d'une couche intersectée et ces entités sont ensuite agrégées à la maille pour constituer une variable (ex : part de commerces, nombre d'établissements, diversité fonctionnelle). La Figure 7 illustre un exemple d'intersection spatiale entre une maille du carroyage, plusieurs bâtiments, un point d'intérêt et un secteur du Mobiliscope.

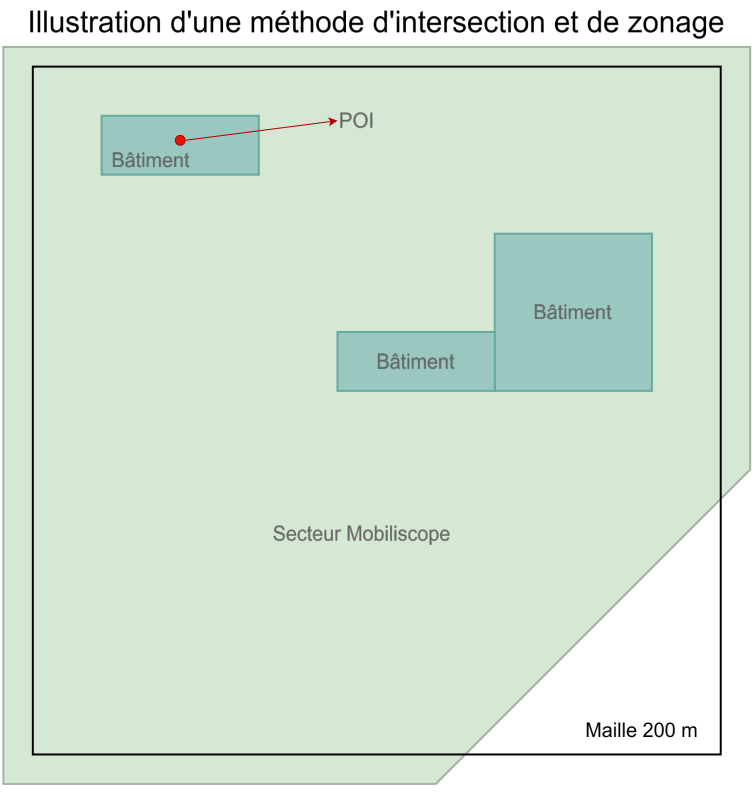


Figure 7: Schéma d'intersection et de zonage. (Ledermann, 2025)

Ces opérations sont automatisées via geopandas, selon la nature des géométries (point ou polygone) et le type de traitement recherché. L'agrégation est toujours facile à l'échelle de la maille, qui devient l'unité centrale d'analyse pour les prédictions. Les opérations d'intersection et de zonage permettent de transférer l'information spatiale vers une structure homogène, le carroyage, tout en respectant la géométrie des objets d'origine. C'est une étape clé pour garantir la pertinence des variables générées.

3.4. Génération des variables explicatives : sélection, calcul, pondération, justification

La construction du modèle repose sur un ensemble de variables dérivées des données spatiales, regroupées par grande thématique : morphologie urbaine, activité économique et composition démographique. Cette section présente les méthodes mobilisées pour produire ces variables de manière systématique et reproductible. Un tableau récapitulatif des variables utilisées est consultable sur le Git-Hub du projet (Annexe 1).

3.4.1. Variables socio-économiques

Les variables socio-économiques mobilisées dans ce travail permettent de caractériser le profil démographique et l'activité des zones étudiées. Issues principalement des données INSEE et SIRENE, elles offrent une lecture quantitative des dynamiques résidentielles et économiques à l'échelle locale. Les données mobilisées proviennent essentiellement de la base fiscale FILOSOFI (carroyage 200 mètres) et de la base SIRENE (établissements actifs, codes NAF, tranches d'effectifs). Les données fiscales nous donnent : la population résidente, la part des jeunes (moins de 20 ans), la part des personnes âgées (plus de 65 ans) et la densité résidentielle. La Figure 8 illustre la distribution de la part des jeunes dans une portion l'Eurométropole de Strasbourg, montrant l'hétérogénéité spatiales des zones démographiques.

Exemple de distribution de la part des jeunes sur une partie l'Eurométropole de Strasbourg

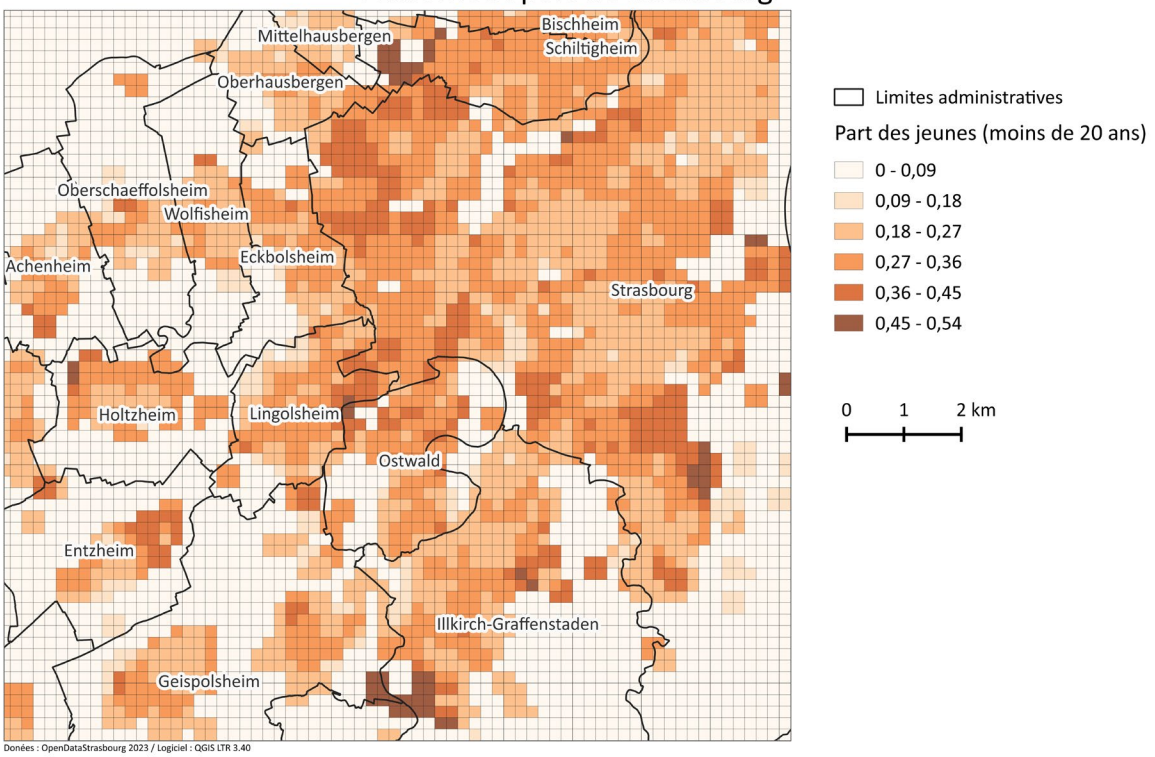


Figure 8: Répartition de la variable (part des moins de 20 ans), sur le territoire de l'Eurométropole de Strasbourg. (Ledermann, 2025)

De son côté la base SIRENE, nous donnent plusieurs informations et données essentielles : le nombre d'établissements actifs dans la maille (toutes catégories confondues), le nombre d'emplois estimés pondérés qui est une estimation indirecte du nombre de travailleurs à partir des tranches d'effectifs du code NAF (pondération selon la branche d'activité), la proportion d'établissement recevant du public qui est approximée à partir des NAF ciblés (commerce, administration, enseignement, etc.) et l'indice de mixité de fonctionnelle calculé avec l'indice de Shanon, en s'appuyant sur les catégories issues des codes NAF. Il est défini par la formule :

$$H = \sum_i p_i \log(p_i)$$

Où p_i représente la proportion d'établissements appartenant à la catégorie i. Plus l'indice est élevé, plus la maille présente une mixité fonctionnelle élevée. Ces variables sont harmonisées pour correspondre à l'unité spatiale cible, avec traitement des valeurs manquantes (Nan) par imputation simple ou exclusion. L'objectif est de traduire des intensités économiques et sociales, qui sont souvent fortement corrélées à la présence réelle de population (travailleurs, clients, usagers, etc.). Les variables socio-économiques constituent un socle essentiel du modèle, elles traduisent à la fois le potentiel attractif d'une zone et la densité de ces fonctions résidentielles ou productives.

3.4.2. Variables morphologiques et de densité

La structure physique de l'environnement bâti influence fortement la présence réelle de population (Biljecki, 2022). Les variables morphologiques permettent de qualifier la densité, la hauteur, la compacité ou la diversité du tissu urbain, à partir des données issues de la BD TOPO. Les géométries des bâtiments servent de base pour plusieurs indicateurs à l'échelle de chaque maille : la hauteur moyenne pondérée à la surface qui reflète une densité verticale, elle est calculée avec la formule suivante :

$$H_{pondérée} = \frac{\sum_i (h_i \times S_i)}{\sum_i S_i}$$

Où h_i est la hauteur du bâtiment i, et S_i sa surface au sol. Cette formule pondère les hauteurs en fonction de la surface, ce qui évite qu'un petit bâtiment très haut fausse la moyenne. Ensuite l'écart type des hauteurs, qui est un indicateur de variation verticale, utile pour différencier les tissus homogènes (zones pavillonnaires) des zones mixtes (centre-ville), il se calcul par cette formule :

$$\sigma_h = \sqrt{\frac{1}{n} \sum_i (h_i - \bar{h})^2}$$

Le shape-index, le rapport entre la surface d'un bâtiment et celle de son enveloppe convexe (compacité), qui se calcul par la formule suivante :

$$SI = \frac{P}{2\sqrt{\pi A}}$$

Où P est le périmètre du bâtiment, et A sa surface. Un shape index proche de 1 indique une forme compacte (ex : cercle ou carré), tandis qu'un indice élevé signale une forme allongée ou fragmentée, ce qui permet encore une fois de données une indication sur la morphologie du bâti. Mais aussi le volume moyen des bâtiments par maille, qui ce calcul avec la formule suivante :

$$V_i = h_i \times S_i$$

Cette variable permet de mieux appréhender la capacité de densité verticale d'une zone, ce qui peut influencer la population présente. La largeur moyenne des rues obtenue avec la BD TOPO, nous renseigne aussi sur la porosité du tissu urbain, l'accessibilité, et la potentielle fluidité des déplacements. Enfin, la distance moyenne entre bâtiments mesure la dispersion du tissu urbain (approchée par barycentres). Ces variables sont normalisées par la surface de la maille pour permettre la comparabilité inter-mailles, et éviter les biais liés à la taille géographique. La Figure 9, montre un exemple de distribution de la hauteur moyenne pondérée sur une portion de l'Eurométropole de Strasbourg, mettant en évidence les contrastes entre tissus denses et zones plus diffusément bâties.

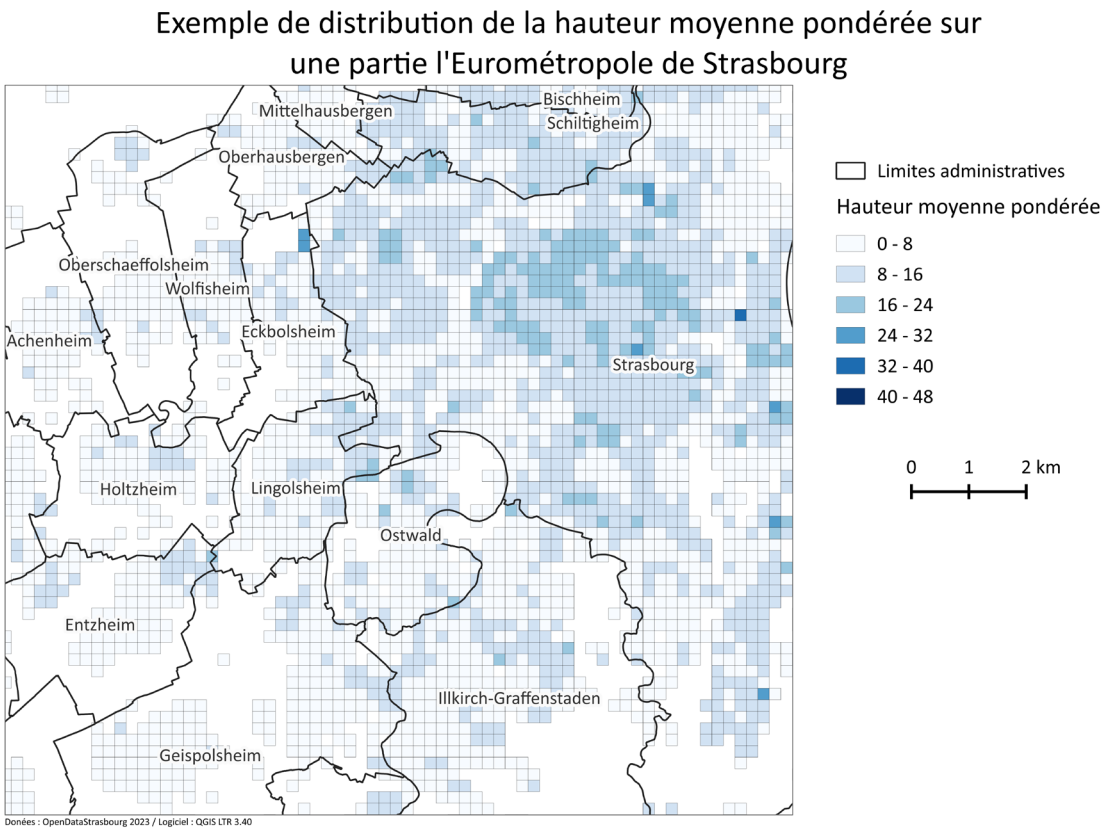


Figure 9: Répartition de la variable (hauteur moyenne pondérée) sur le territoire de l'Eurométropole de Strasbourg. (Ledermann, 2025)

Ces variables servent à objectiver l'intensité morphologique d'une zone, ce qui peut expliquer la présence de population (densité résidentielle, attractivité commerciale, accessibilité verticale, etc.). Les variables morphologiques traduisent la forme et l'intensité du bâti, apportant un éclairage complémentaire aux données socio-économiques, en lien direct avec la structure de l'espace urbain.

3.4.3. Pondérations et agrégations spatiales

Une part importante des variables explicatives repose sur des méthodes d'agrégation spatiale, combinant des objets ponctuels ou polygonaux à des unités de maille régulière. Ces agrégations permettent de produire des indicateurs quantitatifs synthétiques à partir de données brutes hétérogènes. Plusieurs variables sont générées à partir de la densité ou la diversité d'objets spatiaux, notamment les établissements SIRENE et les points d'intérêts OSM. Plusieurs méthodes d'agrégation sont utilisées : l'intersection géographique pour associer des objets à une maille, la pondération par surface intersectée pour des couches polygonales et la bufférisation pour élargir le périmètre de prise en compte autour d'un point (ex : POI dans un rayon de 150 mètres). Les types de POI sont hiérarchisés selon leur pouvoir d'attraction théorique. Par exemple, un établissement scolaire, une gare ou un centre commercial génèrent potentiellement plus de présence humaine d'un distributeur ou un arrêt de bus. Cette pondération permet de différencier les fonctions selon leur intensité d'usage. Chaque catégorie de POI reçoit un

poids ω_i basé sur cette attractivité théorique, le score par maille est calculé avec la formule suivante :

$$Score = \sum_i \omega_i \times N_i$$

D'autres variables sont calculées selon une logique similaire : la part de commerces et d'équipements recevant du public. L'ensemble de ces opérations est automatisé dans les scripts, à partir des couches OSM et SIRENE nettoyées. Les agrégations spatiales et pondérations fonctionnelles permettent de transformer une information brute et diffuse en indicateurs synthétiques et comparables, adaptés à la modélisation.

3.5. Élaboration de la variable cible : extraction Mobiliscope, intersection spatiale, pondération

La modélisation repose sur une variable cible représentant la population moyenne présente sur un territoire donné. Cette section décrit la manière dont cette variable a été construite à partir des données du Mobiliscope, via des opérations d'extraction, d'intersection spatiale et de pondération par surface. Deux variables cibles temporelles ont été construites pour différencier les dynamiques selon les horaires : la population moyenne de jour (10h-16h) et la population moyenne de nuit (00h-06h). Cette distinction permet d'interroger la sensibilité des variables explicatives à la temporalité. Il ne s'agit pas de modéliser une heure précise, mais bien une moyenne agrégée sur des plages horaires typiques : la journée (10h–16h) et la nuit (00h–6h), afin de lisser les effets ponctuels et mieux représenter les dynamiques urbaines récurrentes.

3.5.1. Méthodologie Mobiliscope

La variable cible utilisée pour entraîner le modèle est issue du Mobiliscope, un outil développé par l'UMR Géographie-cités permettant d'estimer la population présente à différents moments de la journée. Afin de l'utiliser comme référence statistique, un travail de traitement et de standardisation est nécessaire. Le Mobiliscope fournit des données agrégées par secteurs statistiques, variables selon les métropoles et les millésimes. Chaque secteur contient un nombre moyen de personnes présentes par heure d'une journée type, ces mêmes secteurs sont fournis sous forme de polygones (GeoJSON). Le traitement préparatoire comprend, le téléchargement de toutes les zones Mobiliscope disponibles, la concaténation des secteurs des différentes agglomérations françaises pour construire un corpus d'entraînement étendu et le nettoyage des géométries. La Figure 10 montre les secteurs Mobiliscope à l'échelle de la France métropolitaine.

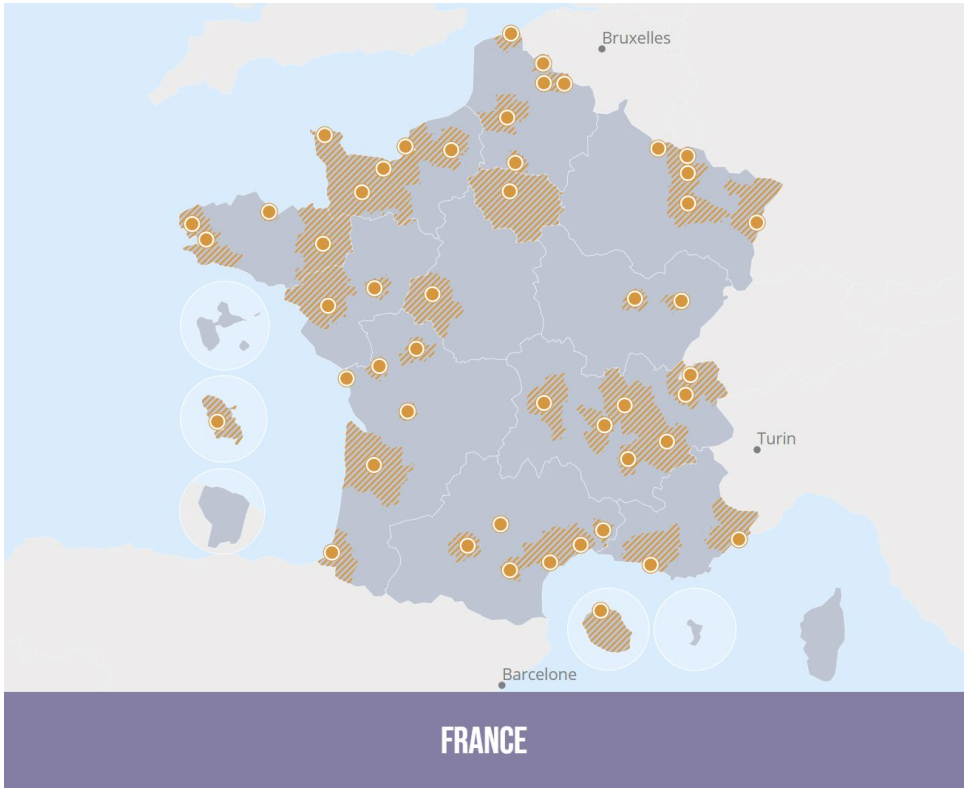


Figure 10: Secteurs Mobiliscope sur la France métropolitaine et DOM-TOM. (Mobiliscope v4.3, 2024)

Les identifiants des secteurs Mobiliscope sont parfois codés de façon hétérogène, il est nécessaire de les standardisés. Ce travail de préparation permet d'obtenir une couche nationale de secteurs avec une valeur cible homogène, associée à une géométrie exploitable dans le pipeline. Le Mobiliscope constitue une source originale, dynamique et spatio-temporelle pour estimer la population présente. Son exploitation nécessite cependant une normalisation rigoureuse pour être mobilisée dans un modèle d'apprentissage.

3.5.2. Intersections géométriques et agrégation

Le modèle est entraîné à partir des secteurs Mobiliscope, tandis que les variables explicatives sont produites à une autre échelle : un carroyage régulier de 200 mètres. Il est donc nécessaire d'effectuer une agrégation ascendante, afin d'adapter les données d'entrée au niveau de la variable cible. Pour entraîner le modèle, chaque secteur Mobiliscope doit recevoir une valeur moyenne de chaque variable explicative, issue des mailles qu'il intersecte. Cette agrégation est réalisée selon une pondération surfacique, permettant de prendre en compte la part de chaque maille incluse dans le secteur, elle faite selon cette formule :

$$X_{secteur} = \sum_{i=1}^{n} (X_j \times \frac{S_{j \cap S}}{S_s})$$

Où X_j est la valeur de la variable dans la maille j, $S_{j \cap S}$ la surface d'intersection entre la maille j et le secteur s, et S_s la surface totale du secteur. Cela garantit que chaque secteur reçoit une valeur moyenne pondérée des mailles qu'il recouvre. Le résultat final est un tableau d'apprentissage où chaque ligne représente un secteur Mobiliscope, chaque colonne une variable explicative moyennée. Cette étape assure la cohérence géométrique entre la source des variables et celle de la cible, en transférant les informations des mailles vers les secteurs de modélisation.

3.6. Choix du modèle : régression linéaire multiple et alternatives (Random Forest, XGBoost)

L'objectif de cette section est de présenter le modèle statistique retenu pour estimer la population présente à partir des variables explicatives spatiales produites précédemment. Elle précise les critères qui ont guidé le choix de la méthode ainsi que sa formulation mathématique et les hypothèses associées. L'ensemble de la modélisation est conçu dans une logique de reproductibilité, avec une attention particulière portée à l'interprétabilité des résultats, tout en testant des approches complémentaires plus performantes en apprentissage automatique.

3.6.1. Critères de choix du modèle

Le choix d'un modèle statistique ne repose pas uniquement sur ses performances prédictives, mais aussi sur sa capacité à être interprété, transposé et documenté. C'est pourquoi la régression linéaire multiple (RLM) a été retenue comme approche de base dans ce mémoire. La RLM offre un bon compromis entre la simplicité de mise en œuvre, la clarté des résultats (coefficients interprétables), et la rapidité d'entraînement, idéale pour un volume important de données. Elle permet de quantifier la contribution individuelle de chaque variable explicative, ce qui est cohérent avec l'objectif de comprendre les liens entre forme urbaine, activité et population présente. Ce modèle est aussi bien adapté à un jeu de données avec un nombre modéré de variables, sans sur-paramétrisation excessive, dans un cadre d'analyse explicative plus que purement prédictive. Le choix de la RLM est également motivé par la structure du jeu de données et par sa compatibilité avec des analyses complémentaires comme l'analyse bivariée (R^2 individuel), l'analyse en composantes principales (ACP), ou encore des tests croisés avec des modèles non linéaires. En complément, deux méthodes d'apprentissage automatique ont été intégrées à des fins comparatives : le Random Forest et XGBoost. Ces approches offrent des performances supérieures sur des jeux de données hétérogènes, au prix d'une perte d'interprétabilité directe.

3.6.2. Formulation mathématique et hypothèses associées

Au-delà du choix méthodologique, il est essentiel de formuler explicitement le modèle retenu et les hypothèses qu'il implique. La régression linéaire multiple repose sur une structure mathématique rigoureuse, qui conditionne l'interprétation statistique des résultats. Le modèle utilisé est une régression linéaire multiple, formulée comme suit :

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$$

Où Y est la variable cible (population moyenne présente), X_i sont les variables explicatives, β_0 est l'ordonnée à l'origine, β_i sont les coefficients associés à chaque variable et ε est le terme d'erreur aléatoire. Le modèle estime les paramètres β par moindres carrées ordinaire (OLS), c'est-à-dire en minimisant la somme des carrés des résidus :

$$min_{\beta} = \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2$$

Ce cadre repose sur plusieurs hypothèses statistiques classiques : la linéarité, la relation entre chaque X_i et Y est supposée linéaire, l'indépendance des erreurs, ε est supposé indépendant et non auto-corrélé, l'homoscédasticité de la variance des erreurs, la normalité des erreurs et l'absence de colinéarité forte entre les X_i. La formalisation du modèle linéaire fournit un cadre rigoureux à l'analyse statistique. Ces hypothèses sont testées à posteriori à l'aide de matrices de corrélation, des graphes de résidus et de tests de normalité. Le modèle permet de dériver des quantités utiles : les coefficients estimés $\hat{\beta}_i$, le coefficient de détermination R^2 et les valeurs résiduelles.

3.6.3. Modèles non linéaires testés : Random Forest et XGBoost

En complément de la régression linéaire, deux modèles non linéaires ont été testés dans ce travail : Random Forest et XGBoost. Le Random Forest est une méthode d'agrégation d'arbres de décision fondée sur un principe de bagging. Elle repose sur l'entraînement de nombreux arbres indépendants sur des sous-échantillons aléatoires du jeu de données, puis sur la moyenne des prédictions. Ce modèle est robuste aux corrélations et aux valeurs extrêmes, et performant en termes de prédiction. Il fournit un indicateur d'importance globale des variables, basé sur la réduction d'impureté moyenne. Le XGBoost, quant à lui, repose sur un principe de boosting : chaque arbre est entraîné pour corriger les erreurs du précédent. Il offre des performances particulièrement élevées, grâce à des mécanismes de régularisation intégrés et une optimisation du gradient. Il est souvent utilisé dans les contextes de compétition en machine learning pour sa capacité à capter des interactions complexes entre variables. Ces deux modèles sont mobilisés ici dans une perspective comparative. Ils permettent d'évaluer la capacité prédictive brute des données spatiales, sans imposer de contrainte de linéarité ou d'indépendance. En revanche, ils ne produisent pas de coefficients interprétables, ce qui limite leur usage

dans un cadre explicatif. Leur lecture repose davantage sur des indicateurs comme les « feature importances ».

La régression linéaire multiple constitue la base interprétative du modèle, en raison de sa lisibilité et de son adéquation au cadre explicatif du mémoire. L'ajout de modèles non linéaires comme Random Forest et XGBoost permet de tester la robustesse des données spatiales dans une approche plus prédictive, et d'élargir les perspectives d'analyse sans compromettre la lisibilité générale de la démarche.

3.7. Métrique d'évaluation : R^2, RMSE, validation croisée, robustesse

L'évaluation du modèle repose sur un ensemble de métriques statistiques visant à mesurer à la fois sa capacité explicative, sa précision prédictive et la structure interne des relations entre variables. Cette section présente les indicateurs mobilisés, ainsi que les méthodes complémentaires d'analyse statistique qui permettent de mieux comprendre les performances du modèle et la pertinence des variables explicatives.

3.7.1. Indicateurs de performance globale

Pour évaluer la qualité du modèle, plusieurs métriques classiques sont mobilisées, permettant de quantifier la part de variance expliquée, l'erreur de prédiction et la distribution spatiale des résidus. En premier, le coefficient de détermination R^2 est utilisé pour mesurer la proportion de variance de la variable cible expliquée par le modèle, il s'exprime de la manière suivante :

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Il varie entre 0 et 1, où 1 signifie que le modèle explique parfaitement la variable cible. En second, on trouve le RMSE (Root Mean Square Root) qui est utilisé pour quantifier l'erreur moyenne de prédiction, il se définit comme suit :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Il est exprimé dans les mêmes unités que la variable cible (ici : nombre de personnes) et donne une idée de l'amplitude des erreurs. Ces indicateurs sont calculés pour le modèle globale et comparés entre les deux variables cibles : population de jour et population de nuit. En complément, une analyse spatiale des résidus est réalisée avec notamment une carte de résidus et une carte d'erreur absolue. Ces cartes permettent d'identifier des zones d'ajustement insuffisant, souvent liées à des contextes particuliers (zones périurbaines, secteurs atypiques, etc.). L'analyse conjointe du R^2, du RMSE et des cartes de résidus permet une évaluation complète du modèle, aussi bien statistique que

spatiale, et alimente les réflexions pour la suite (amélioration des variables ou des méthodes).

3.7.2. Analyse bivariée par variable explicative

Au-delà de la performance globale du modèle, il est utile d'évaluer l'impact individuel de chaque variable explicative sur la population estimée. L'analyse bivariée permet d'isoler la contribution de chaque variable, à travers le calcul de coefficients de corrélation et d'indicateurs de qualité d'ajustement. Pour chaque variable X_i on calcule le coefficient de corrélation linéaire de Pearson r, avec la formule suivante :

$$r = \frac{Cov(X_i, Y)}{\sigma_{X_i} \times \sigma_Y}$$

Le R^2 bivarié, en ajustant un modèle linéaire simple, comme suit :

$$Y = \beta_0 + \beta_1 X_i + \varepsilon$$

Et aussi le RMSE bivarié, pour estimer la précision individuelle d'une variable. Cette approche permet de classer les variables par pertinence statistique, détecter celles dont la contribution est négligeable (ex : $R^2 < 0{,}1$) et de justifier une éventuelle sélection ou élimination en amont de la modélisation multivariée. Des visualisations complémentaires sont utilisées : scatterplots annotés (nuage de points), graphique barre des R^2 par variable. L'analyse bivariée permet d'identifier les variables les plus explicatives en elles-mêmes, et d'enrichir la lecture du modèle global. Elle constitue un outil de diagnostic préalable à la modélisation multivariée.

3.7.3. Analyse factorielle des variables explicatives

L'analyse factorielle permet d'explorer la structure interne du jeu de variables explicatives. En complément des mesures bivariés, elle permet d'identifier d'éventuelles redondances, des axes latents dominants, et des regroupements interprétables d'indicateurs spatiaux. L'analyse mobilisée est une Analyse en Composantes Principales (ACP). Les objectifs de cette analyse sont d'identifier les axes factoriels expliquant une part importante de la variance, regrouper les variables qui évoluent de manière cohérente et de visualiser la position relative des variables dans l'espace factoriel. Cette analyse permet aussi de détecter des corrélations implicites non visibles via les coefficients de Pearson et aussi d'envisager des agrégats thématiques ou des regroupements interprétatifs. L'ACP complète les analyses précédentes en apportant une vision synthétique de la structure des variables, utile pour guider l'interprétation des résultats et affiner la sélection des prédicateurs pertinents.

L'ensemble de la méthodologie présentée dans cette troisième partie repose sur une chaîne de traitement complète, allant de l'acquisition des données brutes à l'entraînement du modèle. Chaque étape a été pensée dans une logique de reproductibilité, de rigueur statistique et de cohérence spatiale.

Les variables explicatives produites couvrent un large spectre d'indicateurs morphologiques, socio-économiques et fonctionnels, et sont croisées à une donnée cible issue du Mobiliscope, redistribuée par agrégation spatiale.

Le choix d'un modèle explicatif, basé sur la régression linéaire multiple, permet d'interpréter les résultats avec précision, tout en ouvrant la voie à des extensions plus complexes si nécessaire.

La partie suivante présente les résultats issus de la mise en œuvre du modèle, en s'attachant à évaluer la qualité de la prédiction, à interpréter les coefficients obtenus, et à identifier les variables les plus pertinentes à travers des analyses croisées.

4. Résultats – Évaluation du modèle

4.1. Performances globales du modèle

Cette section présente les performances des trois modèles mobilisés dans le cadre de ce mémoire : la régression linéaire multiple, Random Forest et XGBoost. Ces performances sont évaluées à l'échelle nationale, sur l'ensemble des secteurs Mobiliscope, pour les deux variables cibles temporelles (population de jour et de nuit). L'objectif est de comparer leur capacité à expliquer la distribution spatiale de la population présente, en s'appuyant sur deux métriques classiques: le coefficient de détermination R^2 et l'erreur quadratique moyenne (RMSE). Ces résultats permettent d'évaluer la qualité d'ajustement des modèles et de justifier leur mobilisation conjointe dans une logique complémentaire.

4.1.1. Qualité de l'ajustement (R^2, RMSE)

L'évaluation des modèles permet ici de comparer leur qualité d'ajustement sur l'ensemble des secteurs Mobiliscope. Les valeurs de R^2 et de RMSE ont été calculées pour les deux variables cibles temporelles. La Figure 11 récapitule les résultats obtenus par les modèles.

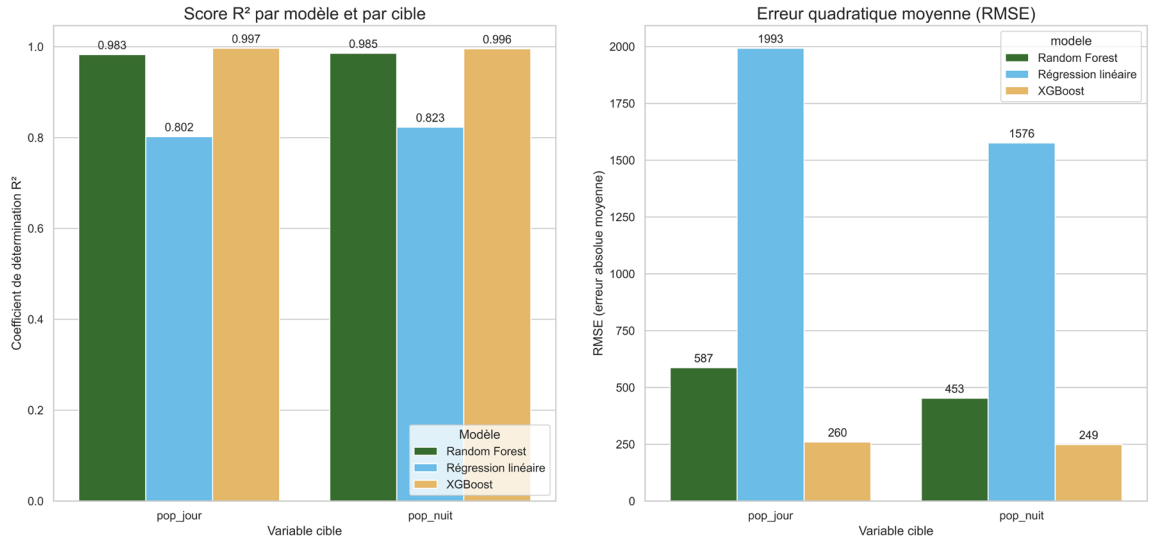


Figure 11: Graphique récapitulatif des métriques d'évaluation (R^2 et RMSE) pour les trois modèles. (Ledermann, 2025)

Les résultats obtenus montrent une nette hiérarchie entre les modèles. La régression linéaire multiple obtient des performances satisfaisantes pour la population de nuit (R^2 égal à 0,82), mais légèrement en dessous pour la population de jours. Néanmoins, l'erreur quadratique moyenne est très élevée pour ce modèle, notamment pour la population de jour (RMSE égal à 1993). Le modèle Random Forest améliore grandement ces résultats, il affiche un R^2 quasiment égal pour la population de jour et nuit (R^2 environ égal à 0,98). De plus, l'erreur quadratique moyenne est en nette baisse, mais elle reste supérieure le jour

(RMSE égal à 587 de jour contre 453 de nuit). Enfin, XGBoost surpasse légèrement Random Forest en montrant un R^2 autour de 0,99 de jour comme de nuit. C'est aussi le cas pour l'erreur quadratique moyenne, qui descend à 260 pour le jour et 246 pour la nuit. Ces premiers résultats confirment que les modèles non linéaires captent mieux les dynamiques spatiales complexes. Ils justifient la poursuite de l'analyse par une lecture croisée des variables explicatives et une application locale.

4.1.2. Comparaison des modèles (RLM, RF, XGBoost)

Après avoir évalué séparément les performances de chaque modèle, cette sous-partie propose une comparaison structurée entre la régression linéaire, Random Forest et XGBoost, à partir des métriques et d'une interprétation spatiale des erreurs. La Figure 12 ci-dessous est le tableau récapitulatif des métriques par modèle.

Modèle	R² moyen	RMSE moyen
Random Forest	0.984	520.0
Régression linéaire	0.813	1784.0
XGBoost	0.996	254.0

Figure 12: Tableau récapitulatif des métriques d'évaluation par modèle. (Ledermann, 2025)

Sur les pages suivantes, les Figures 13, 14, 15, 16, 17 et 18 représentent les résidus et résidus absolus de chaque modèle pour la population de jour et de nuit. Les cartes présentent trois échelles différentes. La première, nationale, avec la moyenne des résidus normaux et absolus par ville. Ensuite pour permettre plus de détails tout en facilitant la lecture deux sous-échelles sont proposées : l'île de France et l'agglomération de Lyon.

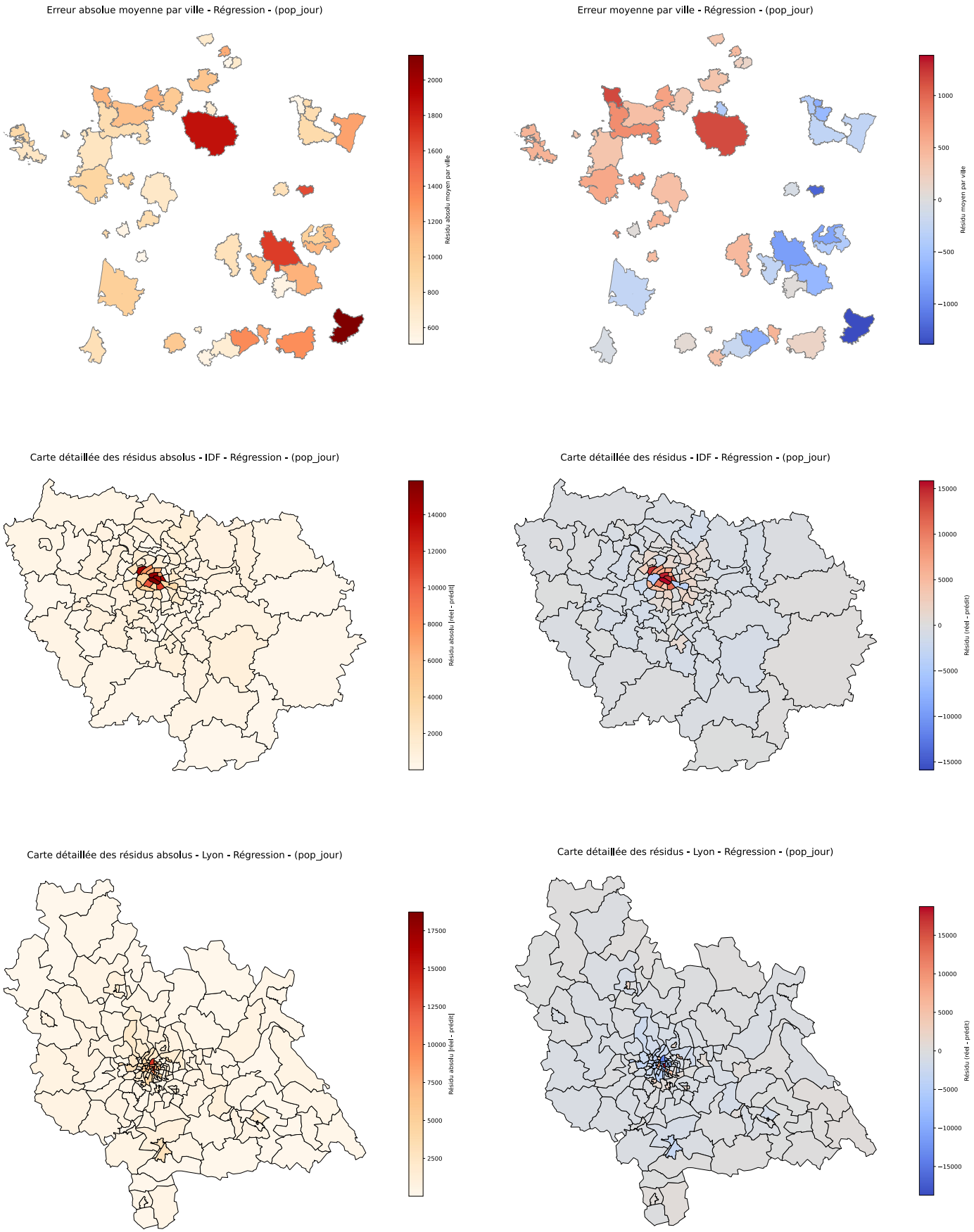


Figure 13: Cartographie des résidus et résidus absolus - Régression Linéaire - Population de jour. (Ledermann, 2025)

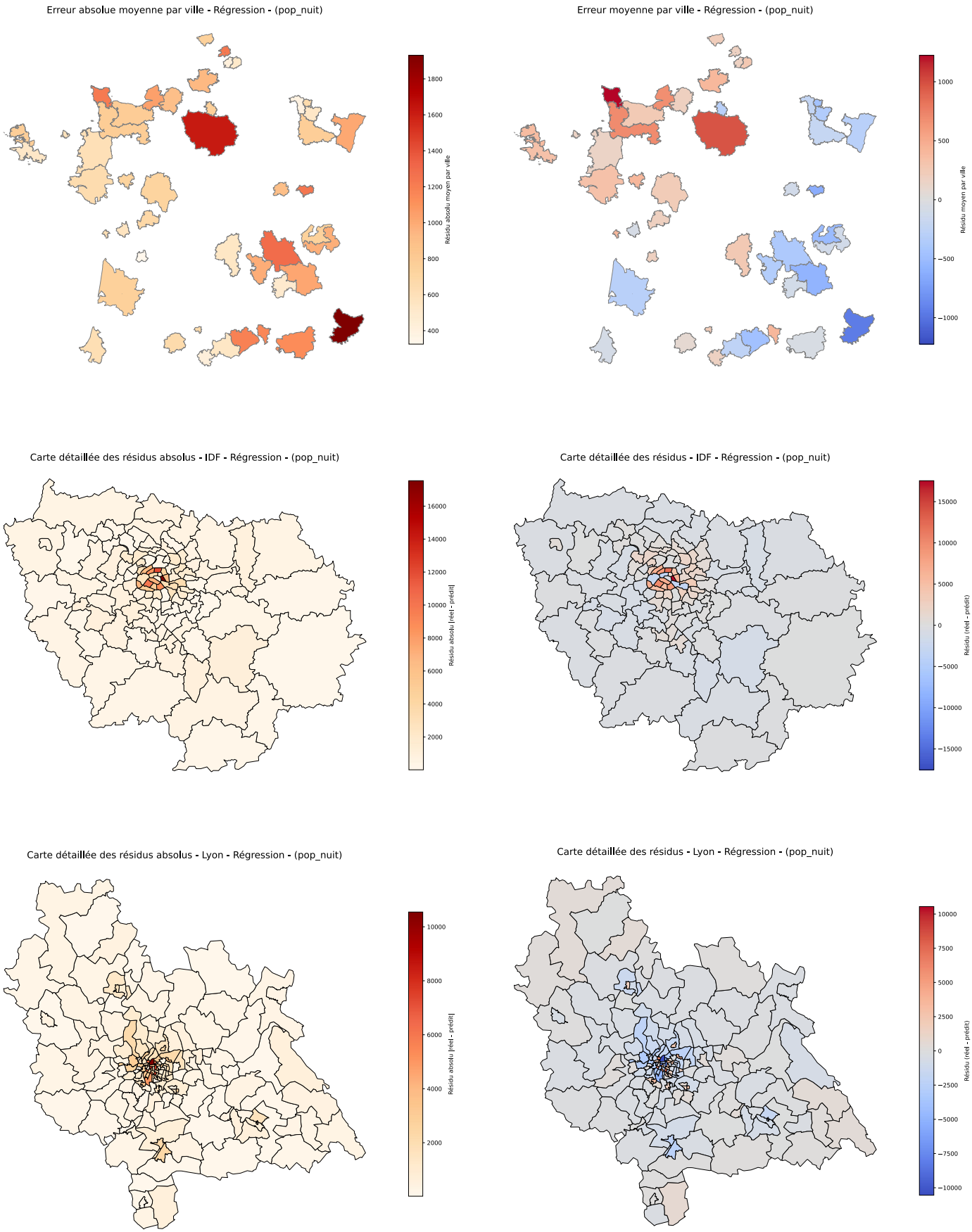


Figure 14: Cartographie des résidus et résidus absolus - Régression Linéaire - Population de nuit. (Ledermann, 2025)

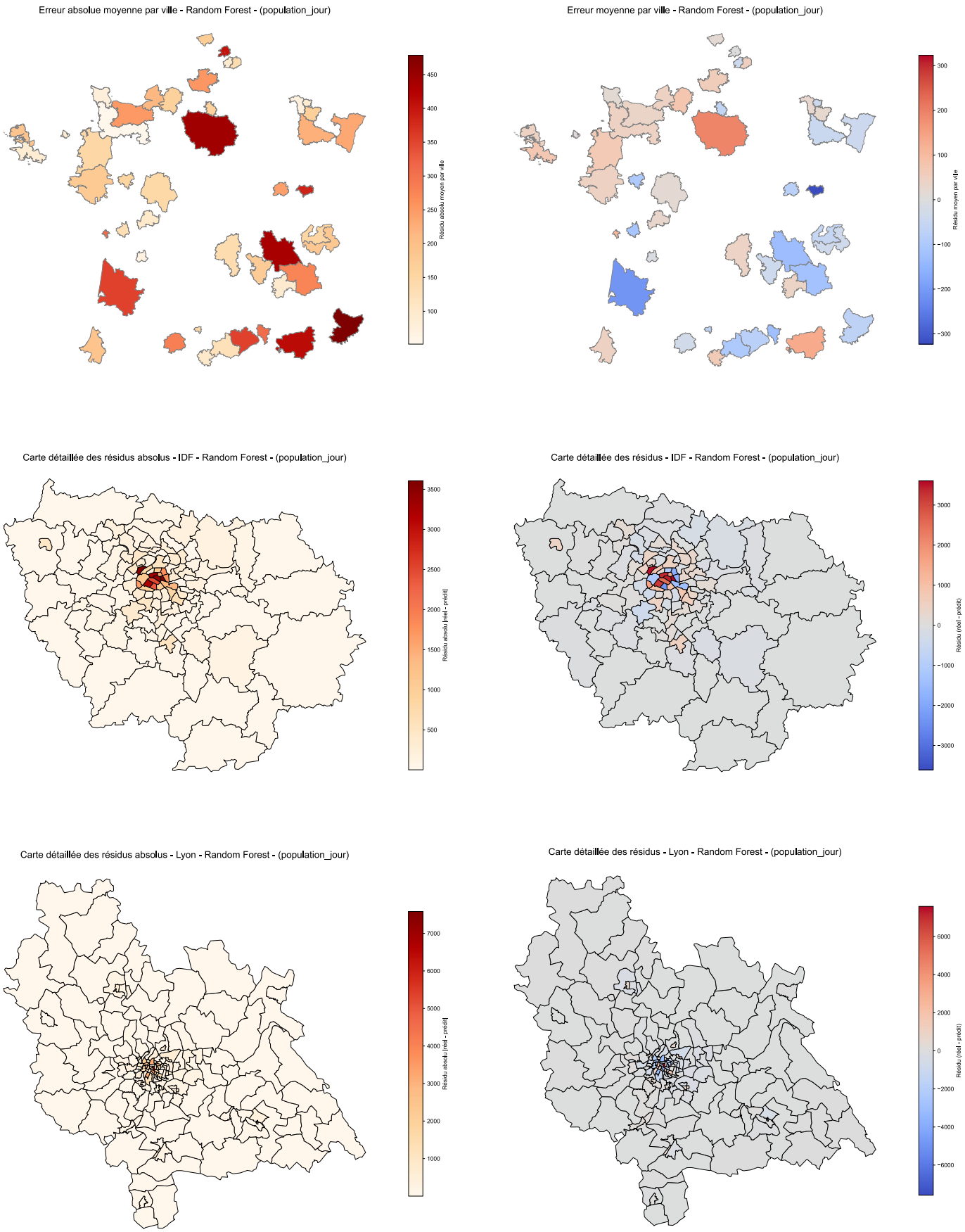


Figure 15: Cartographie des résidus et résidus absolus - Random Forest - Population de jour. (Ledermann, 2025)

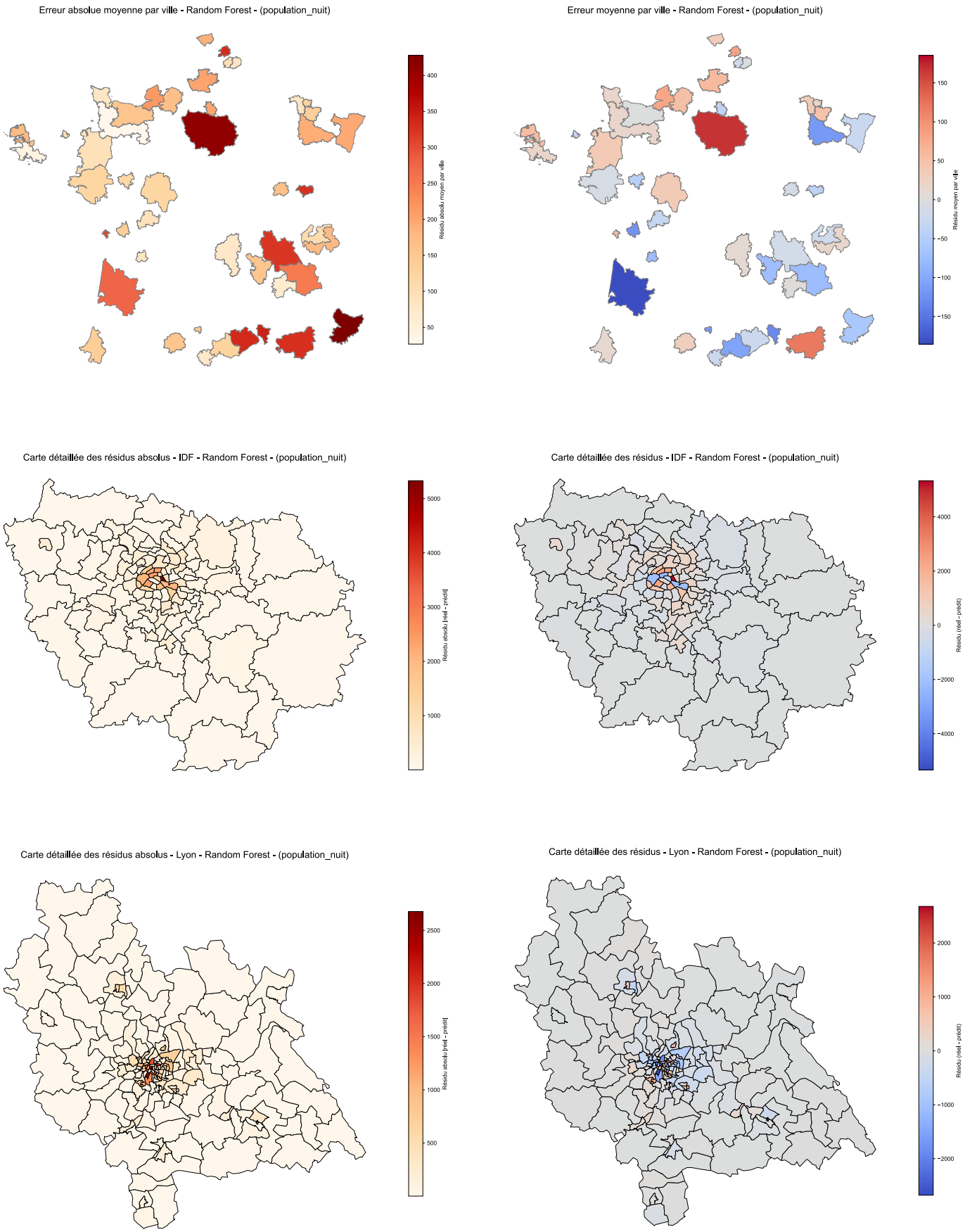


Figure 16: Cartographie des résidus et résidus absolus - Random Forest - Population de nuit. (Ledermann, 2025)

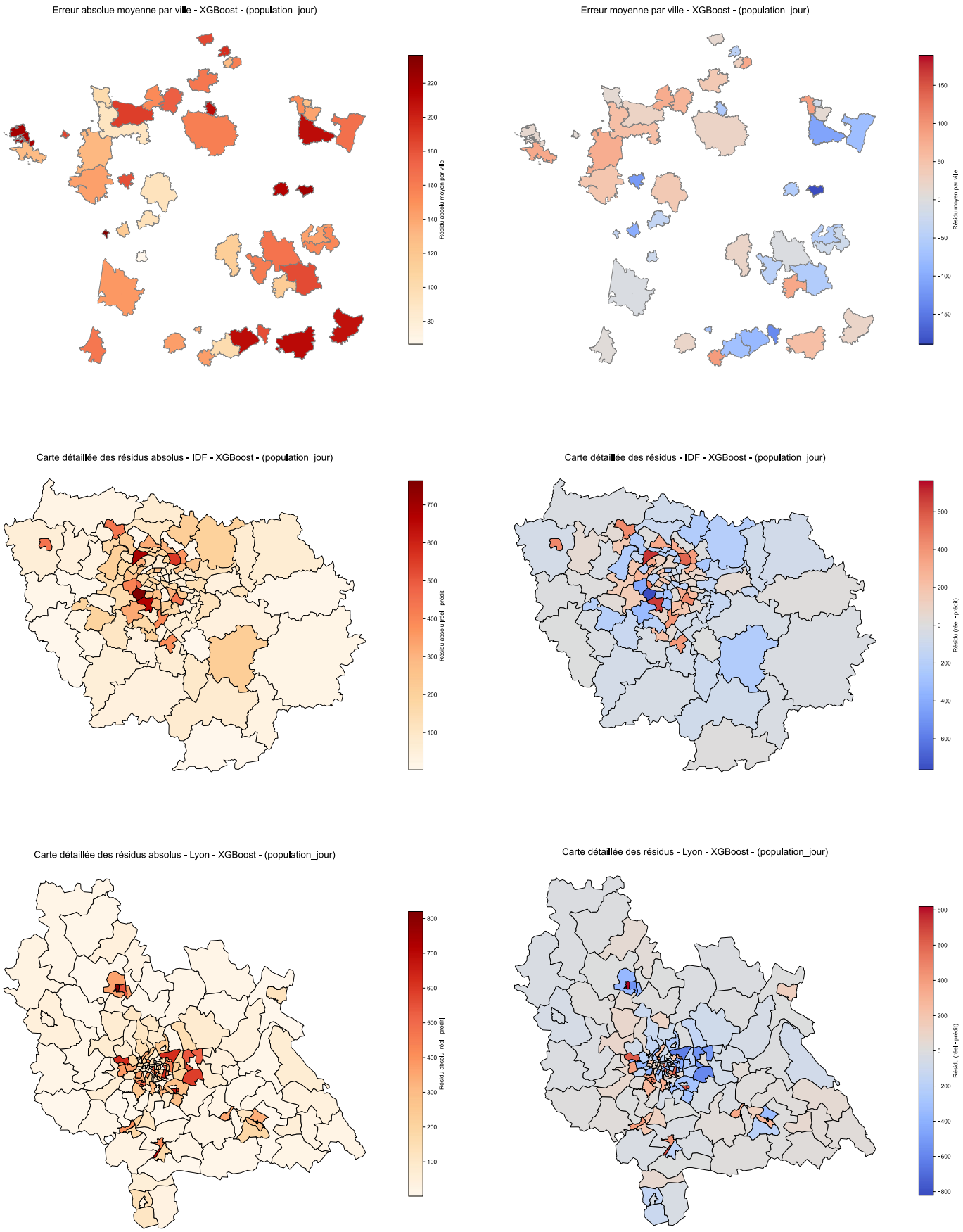


Figure 17: Cartographie des résidus et résidus absolus - XGBoost - Population de jour. (Ledermann, 2025)

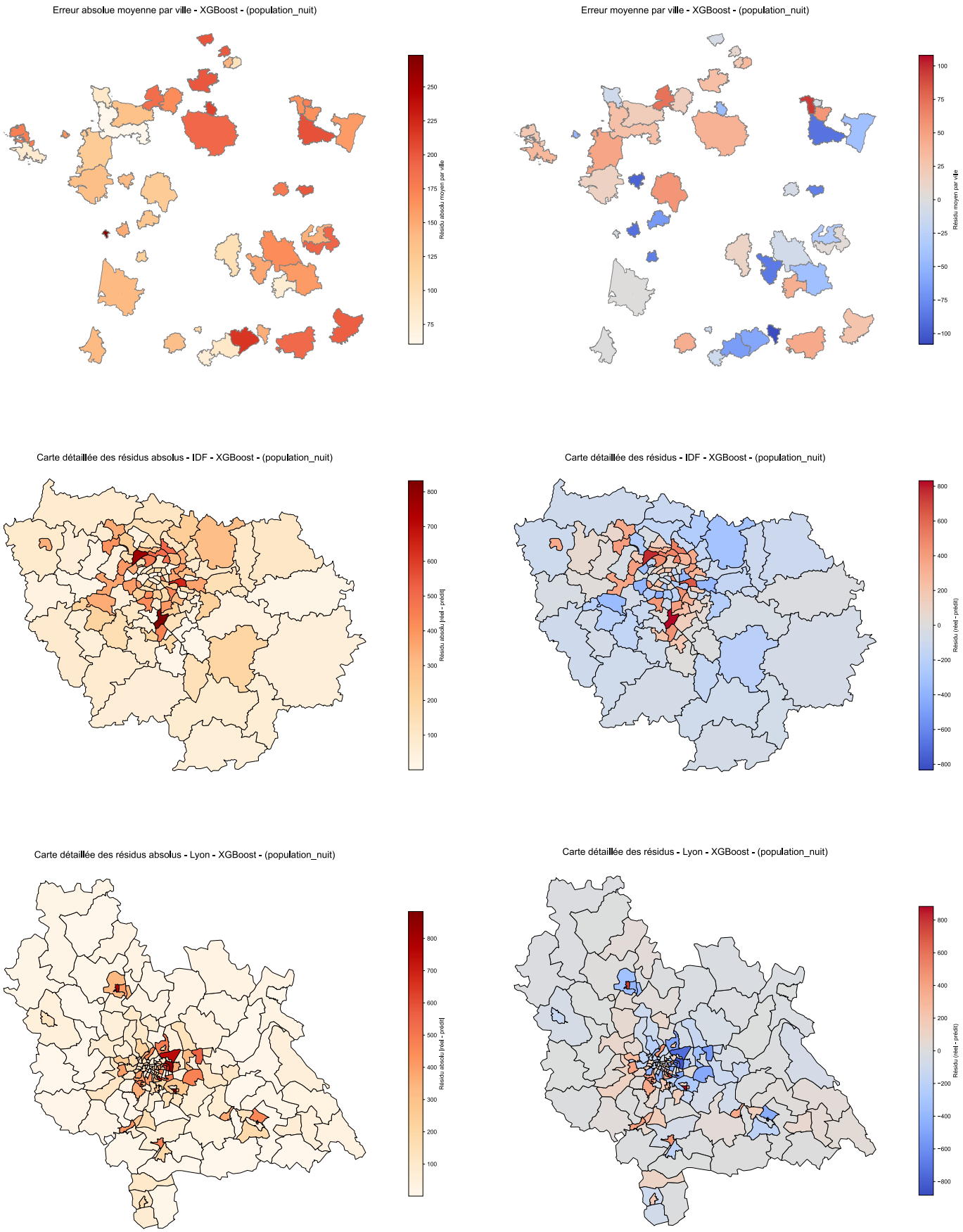


Figure 18: Cartographie des résidus et résidus absolus - XGBoost - Population de nuit. (Ledermann, 2025)

L’analyse des métriques permet de montrer plusieurs aspects des modèles utilisés ici. Tout d’abord, la régression linéaire multiple se montre comme un bon modèle explicatif, mais très sensible à l’hétérogénéité spatiale. Pour Random Forest, on trouve un net gain de précision grâce à la gestion des non-linéarités. Enfin, XGBoost sort du lot avec d’excellente capacité d’ajustement, y compris en période nocturne. L’analyse spatiale des erreurs permet aussi de bien comprendre les modèles et leurs capacités à prédire la population diurne et nocturne. La régression linéaire multiple concentre ces erreurs dans les centres urbains (Figure 13) et dans les espaces très denses comme l’île de France, Lyon et Nice (Figure 13). Les résidus sont très hétérogènes spatialement, aussi bien de jour que de nuit mais restent globalement plus élevés le jour que la nuit (Figure 14). Le modèle Random Forest réduit clairement l’amplitude des erreurs (Figure 11), mais il conserve tout de même ces erreurs dans les centres urbains avec aussi bien de forte sous-estimation et surestimation (Figure 15). A l’échelle nationale, ce sont les régions urbaines denses comme l’Île-de-France, Bordeaux et Lyon qui présentent les résidus les plus importants (Figure 15 et Figure 16). Enfin, XGBoost, réduit encore une fois l’erreur moyenne quadratique (Figure 11). On observe globalement une erreur mieux répartie spatialement, mais toujours plus importante dans les centres urbains (Figure 17 et Figure 18). La comparaison des modèles met en évidence la supériorité du XGBoost pour la cartographie dynamique de la population. Sa précision, sa stabilité spatiale et sa capacité à modéliser des configurations urbaines complexes en font l’outil le plus adapté à une projection à une maille fine. En revanche, son opacité justifie le recours complémentaire à des méthodes explicatives comme l’analyse en composantes principales. Les résultats présentés montrent une hiérarchie nette entre les modèles testés. La régression linéaire, bien qu’interprétable et globalement cohérente, présente des limites dans les contextes urbains denses ou morphologiquement atypiques. Le Random Forest améliore considérablement l’ajustement, mais reste sensible aux variations locales. Le XGBoost se distingue comme le modèle le plus performant et le plus stable, avec une très faible marge d’erreur, quel que soit le contexte spatial ou la période horaire. Ces performances globales doivent être complétées par une analyse plus fine des variables explicatives.

4.2. Analyse des variables explicatives

Pour compléter l'évaluation globale des modèles, cette section s'intéresse à l'influence des variables explicatives sur les prédictions. À travers une analyse bivariée, factorielle et issue des modèles non linéaires, il s'agit d'identifier les facteurs les plus déterminants dans la répartition spatiale de la population présente, ainsi que les redondances ou complémentaires entre indicateurs.

4.2.1. Analyse bivariée (R^2, RMSE individuels)

L'analyse bivariée consiste à mesurer la force du lien entre chaque variable explicative et la population présente, indépendamment de tout modèle multivarié. Pour cela, une régression linéaire simple a été effectuée pour chaque variable, avec calcul du coefficient de détermination R^2 et du RMSE. Les figures 19 et 20 sont les graphiques résultant de l'analyse bivariée pour la population de jour et de nuit.

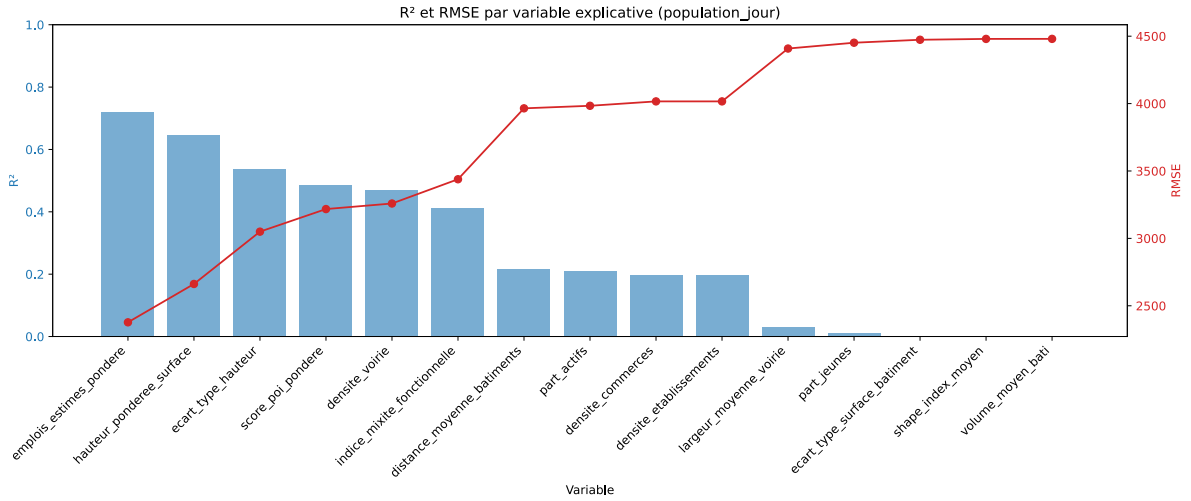


Figure 19: Graphique des métriques par variable - Analyse bivariée - Population de jour. (Ledermann, 2025)

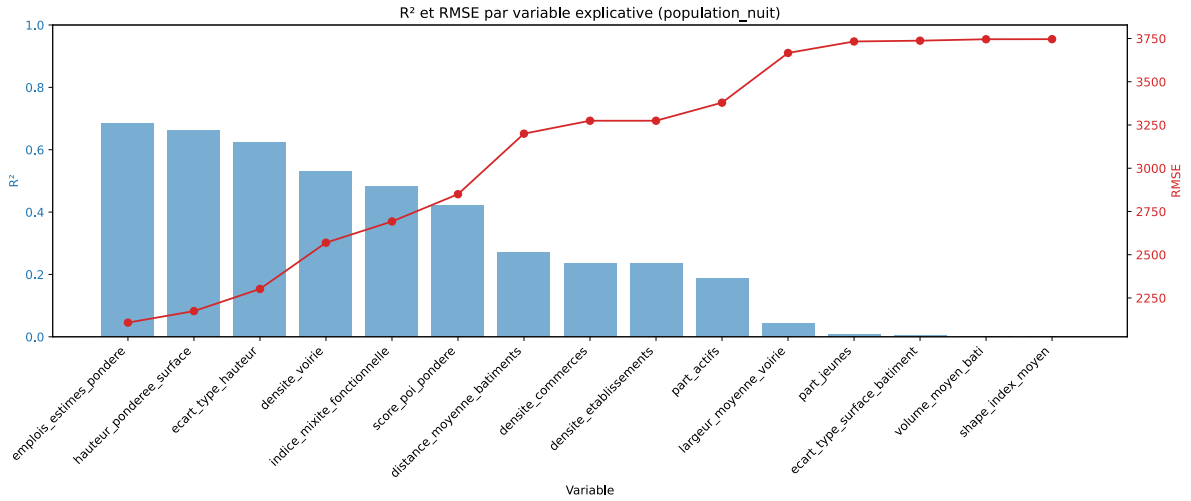


Figure 20: Graphique des métriques par variable - Analyse bivariée - Population de nuit. (Ledermann, 2025)

Pour la population de jour, trois variables se distinguent nettement par leur pouvoir explicatif. L'indicateur *emplois_estimés_pondérés* arrive en tête avec un R^2 de 0,73 et un RMSE de 2391 (Figure 19). Il est suivi par *hauteur_pondérée_surface* (R^2 = 0,66) et *écart_type_hauteur* (R^2 = 0,53), qui confirment le rôle structurant de la verticalité du bâti. La variable *score_POI_pondéré*, construite à partir des données OSM avec une pondération par catégorie, atteint un R^2 de 0,47. D'autres variables présentent des niveaux modérés de corrélation, comme l'indice de mixité fonctionnelle (R^2 = 0,41) ou la densité de voirie (R^2 = 0,47). En revanche, certaines variables initialement intégrées dans le modèle se révèlent peu discriminantes. C'est notamment le cas de *part_jeunes, largeur_moyenne_voirie, shape_index_moyen* et *volume_moyen_bâti*, toutes affichant des proches de zéro. Ces résultats soulignent une faible corrélation directe entre ces indicateurs et la population présente, au moins prise isolément. Pour la population de nuit, la hiérarchie des variables reste globalement similaire. Les emplois estimés et la hauteur pondérée conservent leur première place avec des respectifs de 0,68 et 0,66. Fait notable, l'écart-type de hauteur monte à 0,62, ce qui suggère une certaine influence de l'hétérogénéité morphologique sur la présence nocturne (Figure 20). À l'inverse, le score POI pondéré perd du poids explicatif, tandis que la densité de voirie devient plus significative (0,53). Dans les deux cas, les variables liées à l'activité économique, à la forme urbaine verticale et à la mixité fonctionnelle ressortent comme les plus pertinentes. Cette lecture confirme certaines intuitions initiales, mais permet aussi de clarifier les indicateurs à faible apport. Ces premiers résultats serviront de base à l'analyse factorielle présentée dans la section suivante, qui vise à mieux comprendre les structures latentes entre variables.

4.2.2. Analyse factorielle (ACP) : structure et redondances

Pour compléter l'analyse bivariée et mieux comprendre les relations internes entre les variables explicatives, une Analyse en Composantes Principales (ACP) a été conduite. Elle permet d'identifier les redondances, les groupes de variables homogènes, et les axes latents structurant le jeu de données. La Figure 21, ci-dessous, montre le pourcentage de variance expliquée par composante principale.

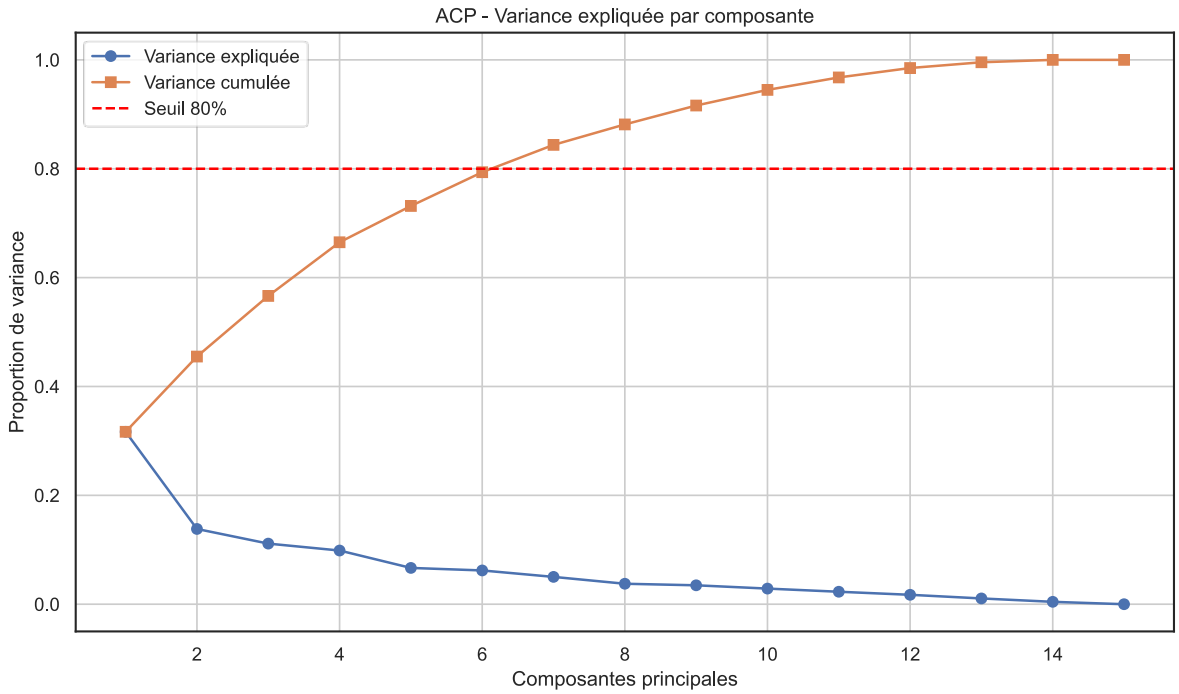


Figure 21: Graphique de la variance expliquée par composante. (Ledermann, 2025)

L'analyse de la variance expliquée par les composantes principales montre que la première composante (PC1) explique 31,7 % de la variance. Puis, la deuxième composante (PC2) explique 13,8 % de la variance. Les six premières composantes cumulent plus de 80 % de la variance, c'est un seuil d'information acceptable pour la réduction dimensionnelle (Figure 21). La Figure 22 ci-dessous permet d'analyser ces composantes.

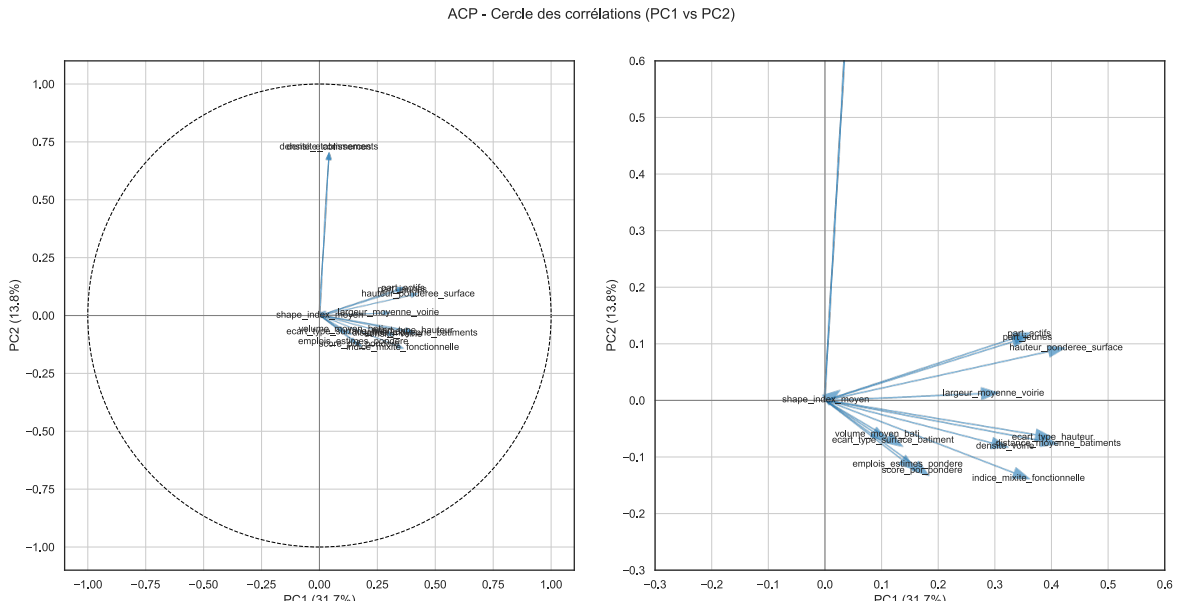


Figure 22: Cercle des corrélations de l'ACP. (Ledermann, 2025)

Le cercle des corrélations permet d'interpréter les composantes. La PC1 regroupe les variables morphologiques et sociales. Elle reflète un gradient de forme urbaine et de jeunesse / mixité. La composante PC2 est dominée par les variables de densité de commerces et d'établissements. Elle décrit un axe d'intensité économique locale, c'est une composante très spécifique. La Figure 23 ci-dessous est la projection des observations issues de l'ACP.

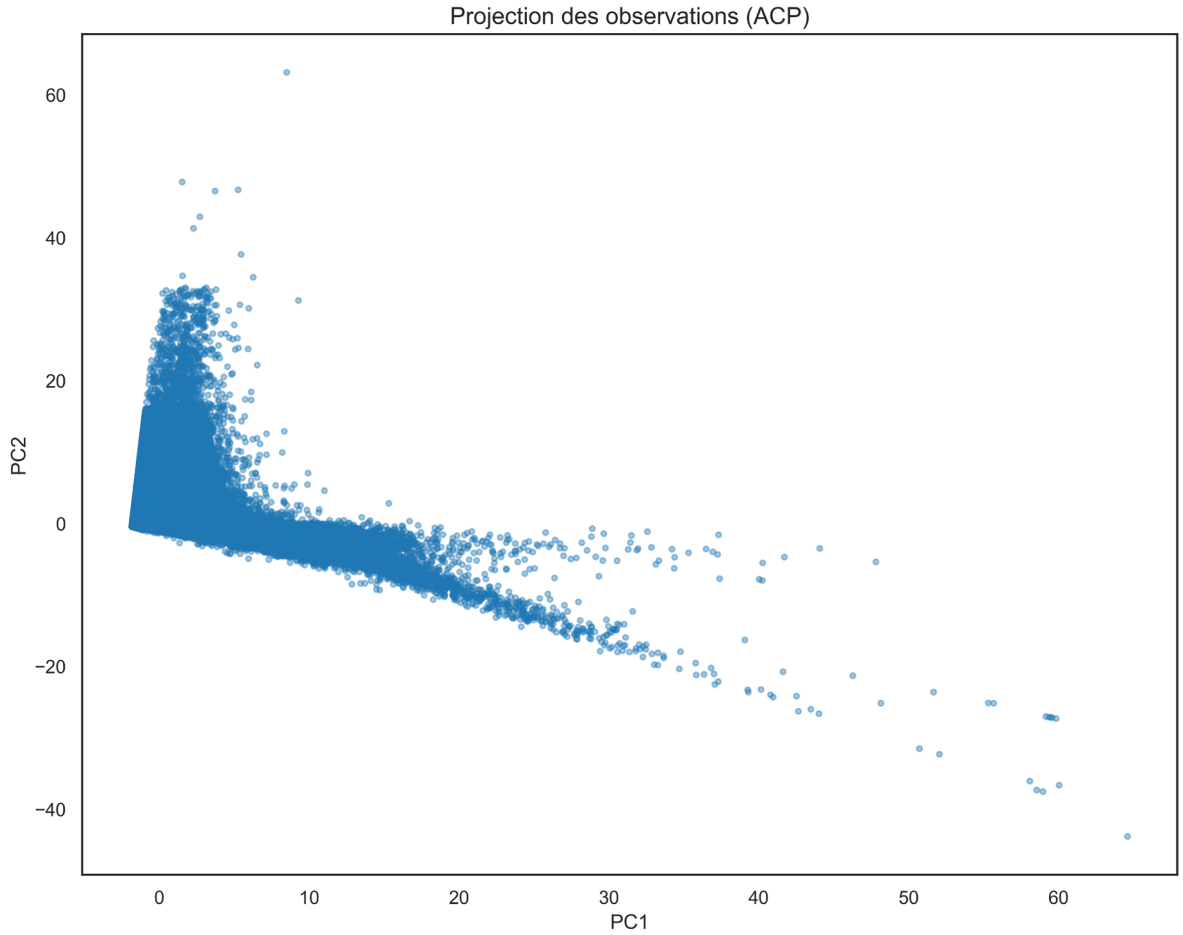


Figure 23: Biplot des observations de l'ACP. (Ledermann, 2025)

La Figure 23, illustre la répartition des variables selon les deux composantes principales. Ici, on remarque une forte densité de points au centre (vers 0,0), ce qui représente les zones urbaines moyennes. On observe aussi la présence de cas extrêmes bien détachés, indiquant des configurations spatiales atypiques comme des zones pavillonnaires éloignées ou des polarités économiques isolées. Les Figure 24 n°24 et 25Figure 25 ci-dessous sont un dernier moyen d'évaluer les résultats de l'ACP, elles représentent la contribution absolue de chaque variable pour chacune des deux composantes principales.

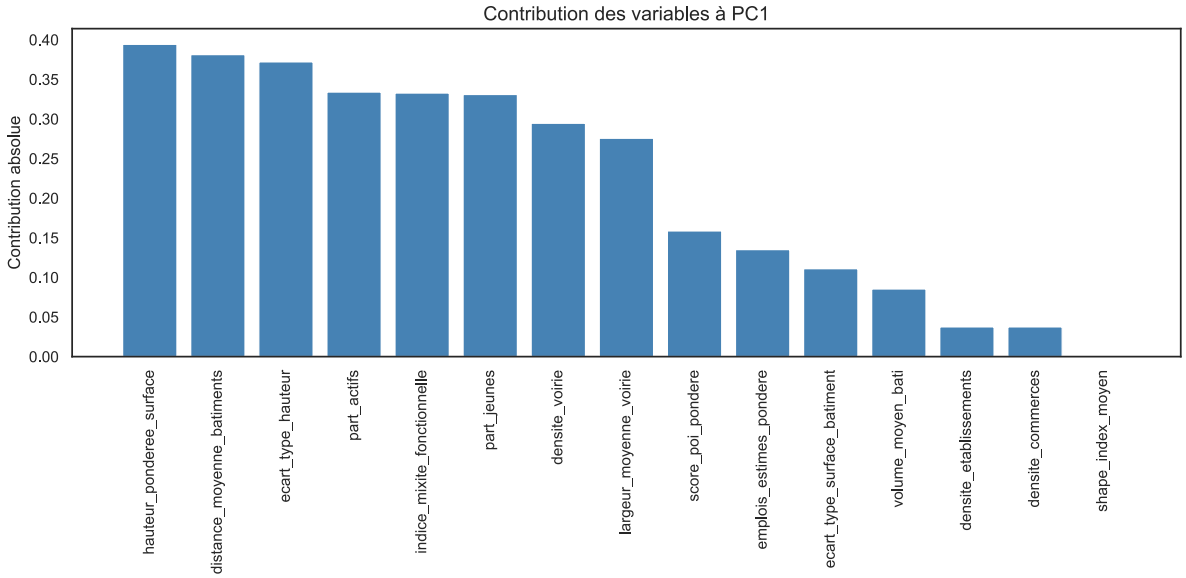


Figure 24: Contribution absolus des variables à PC1. (Ledermann, 2025)

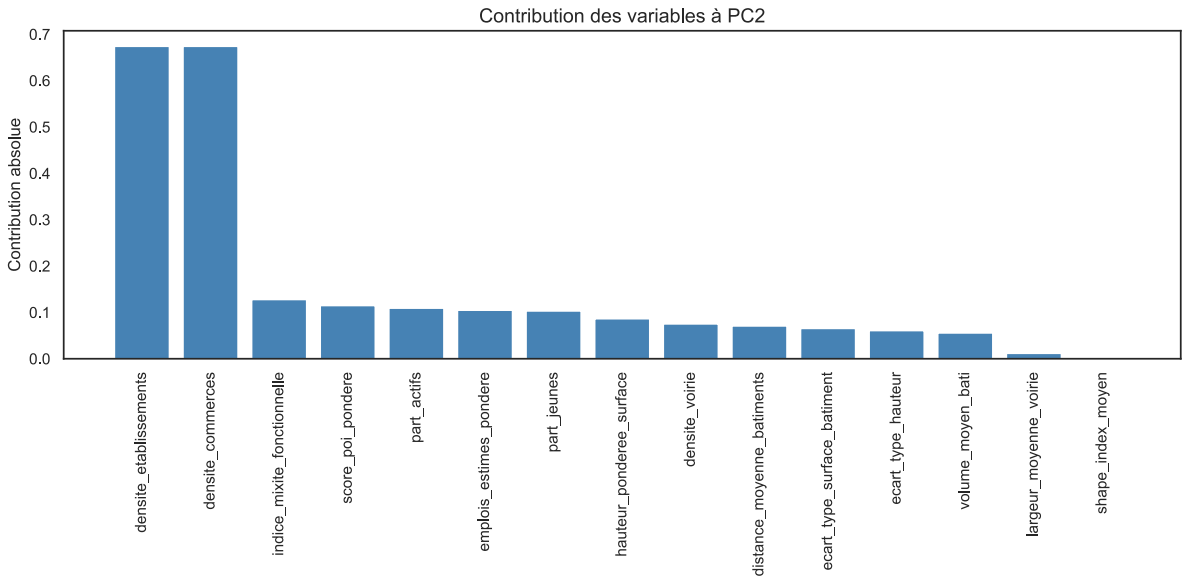


Figure 25: Contribution absolus des variables à PC2. (Ledermann, 2025)

La Figure 24 confirme le poids des variables morphologiques dans la composante PC1. Et la Figure 25, confirme aussi le poids exclusif des variables de densité de commerces et d'établissements. Certaines variables sont faiblement contributives aux deux premiers axes (ex : le score POI, le volume moyen et le shape index), elles peuvent néanmoins porter une information complémentaire dans les composantes suivantes. L'ACP met en lumière deux grands axes de structuration des variables : un axe morpho-social (PC1) et un axe économique (PC2). Cette analyse permet de mieux cerner les logiques de redondance, d'envisager une réduction de dimension et d'interpréter plus finement les comportements observés dans les modèles prédictifs.

4.2.3. Importance des variables selon les modèles non linéaires

Les modèles Random Forest et XGBoost permettent de mesurer l'importance relative de chaque variable explicative dans la prédiction. Cette analyse complète la lecture bivariée et factorielle, en intégrant les interactions complexes et les effets non linéaires. Les importances ont été calculées à partir de la réduction d'impureté pour Random Forest et du gain d'information dans les arbres pour XGBoost. Les Figures 26 et 27 ci-dessous résument l'importance des variables pour les deux modèles avec comme cible la population de jour.

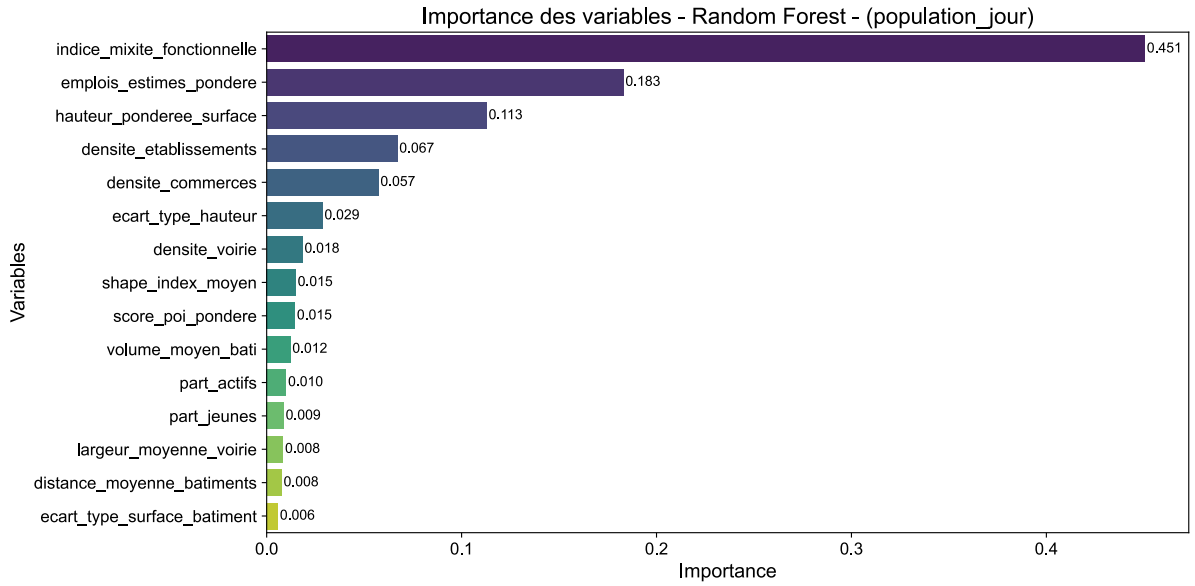


Figure 26: Importance des variables - Random Forest - Population de jour. (Ledermann, 2025)

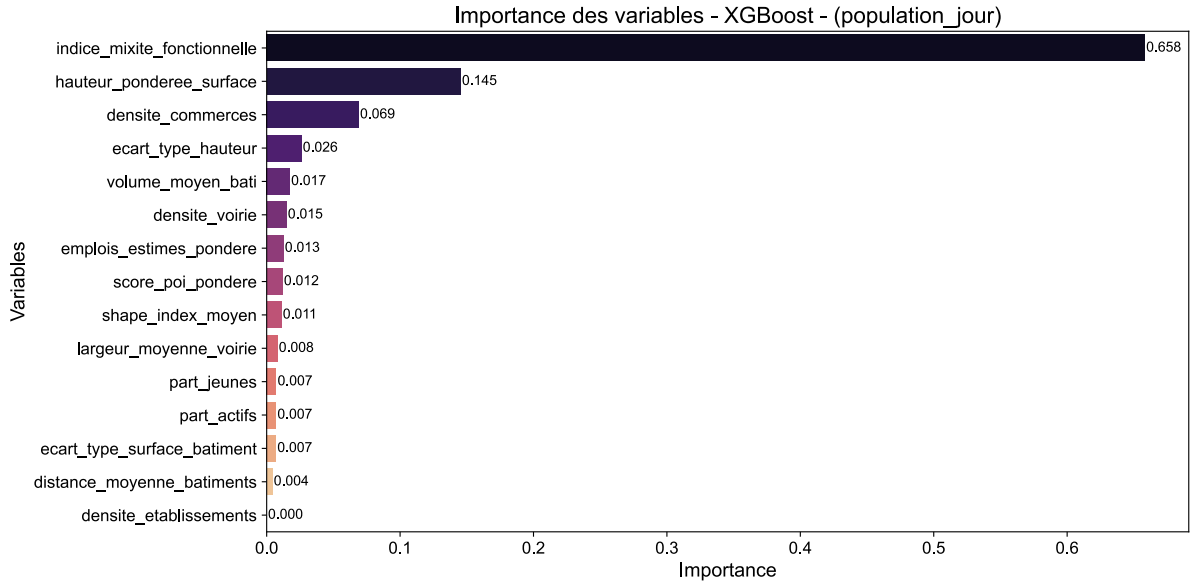


Figure 27: Importance des variables - XGBoost - Population de jour. (Ledermann, 2025)

Avec comme variable cible la population de jour, pour Random Forest l'indice de mixité fonctionnelle arrive largement en tête (45 % de l'importance totale), et il est suivi par les emplois estimés (18 %) et la hauteur pondérée (11 %) (Figure 26). Les variables économiques pèsent moins avec environ 6 % de l'importance. Pour le modèle XGBoost l'ordre change : l'indice de mixité reste la première variable (65 %), suivi par la hauteur pondérée (14 %) et par la densité de commerces (13,5 %), la variable d'emplois estimée est toujours présente mais moins dominante (Figure 27). Les Figures 28 et 29 montrent l'importance des variables pour la population de nuit.

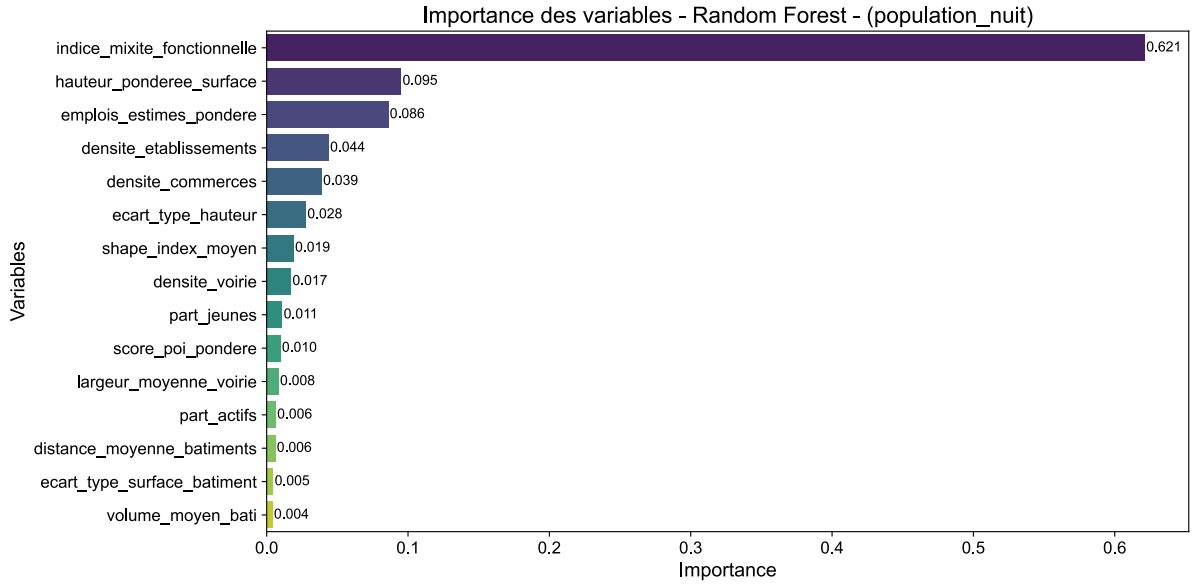


Figure 28: Importance des variables - Random Forest - Population de nuit. (Ledermann, 2025)

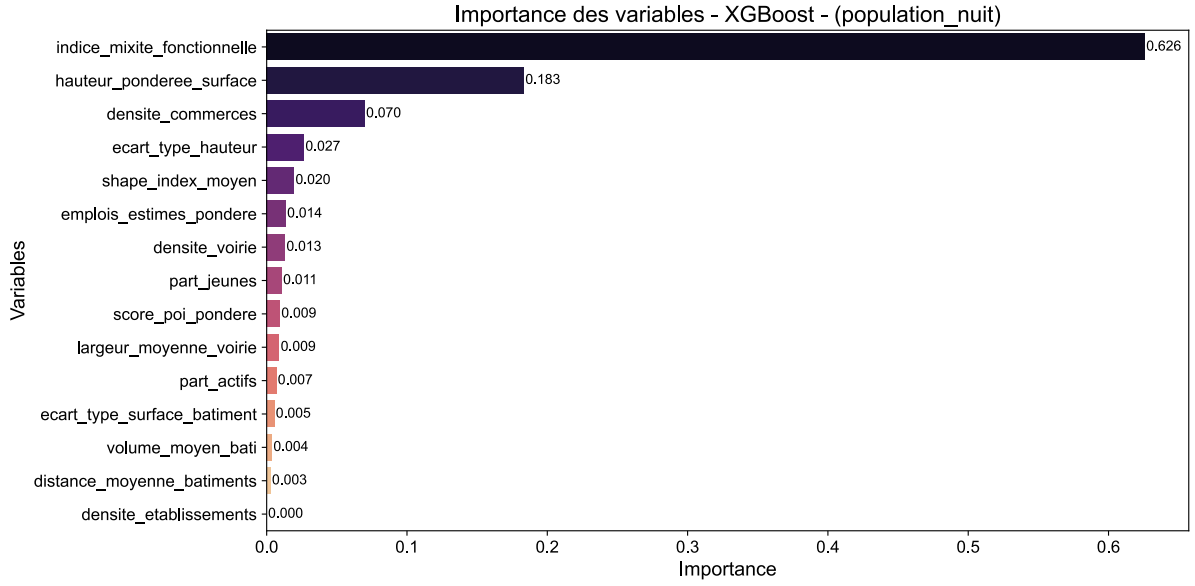


Figure 29: Importance des variables - XGBoost - Population de nuit. (Ledermann, 2025)

Ensuite pour la population de nuit, l'indice de mixité fonctionnelle (62 %) reste en tête pour Random Forest, suivi par la hauteur pondérée (9 %) et les emplois estimés (8 %). L'importance est aussi moins répartie entre les variables (Figure 28). Pour XGBoost, l'indice de mixité (62 %) et la hauteur pondérée (18 %) dominent le haut du graphique, suivi par la densité de commerces avec 7 % de l'importance (Figure 29). Les modèles non-linéaires valorisent différemment les variables selon leur structure. La mixité fonctionnelle, les emplois et la hauteur pondérée ressortent dans tous les cas comme des prédicteurs-clés. L'analyse des importances confirme que les variables les plus utiles ne sont pas toujours les plus corrélées individuellement à la cible. Elle justifie pleinement le recours à des modèles non linéaires pour capturer les interactions complexes entre variables spatiales. Cette hiérarchisation thématique sera mobilisée dans la discussion pour éclairer les dynamiques révélées par la modélisation.

4.3. Application du modèle à une échelle locale (Eurométropole de Strasbourg)

Après l'entraînement et l'évaluation nationale des modèles sur les secteurs Mobiliscope, cette dernière section présente leur application à une échelle locale, sur un maillage de 200 mètres couvrant l'Eurométropole de Strasbourg. L'objectif est de produire une cartographie fine de la population présente, de jour comme de nuit, en projetant les résultats du modèle le plus performant (XGBoost) sur les mailles locales, à partir des variables explicatives préalablement calculées. Cette étape permet de visualiser les dynamiques territoriales modélisées, mais aussi d'analyser les écarts résiduels afin de détecter les zones où le modèle s'avère moins fiable ou moins interprétables.

4.3.1. Cartes de population estimée (jour et nuit)

L'application du modèle XGBoost à l'échelle de l'Eurométropole de Strasbourg permet de produire une cartographie fine de la population présente, différenciée selon les temporalités. Les annexes 2 et 3 renseignent la typologie des quartiers et des communes de l'Eurométropole de Strasbourg afin de faciliter la localisation de ceux-ci. Ces cartes offrent une lecture directe des contrastes entre la structure fonctionnelle de la ville et son occupation résidentielle. La carte de la population de jour (Figure 30) met en évidence une forte polarisation fonctionnelle autour des centralités métropolitaine : Grande Île, gare, institutions européennes, campus Esplanade. Les pics dépassent 20 000 individus par maille, confirmant le rôle d'attraction diurne de ces espaces.

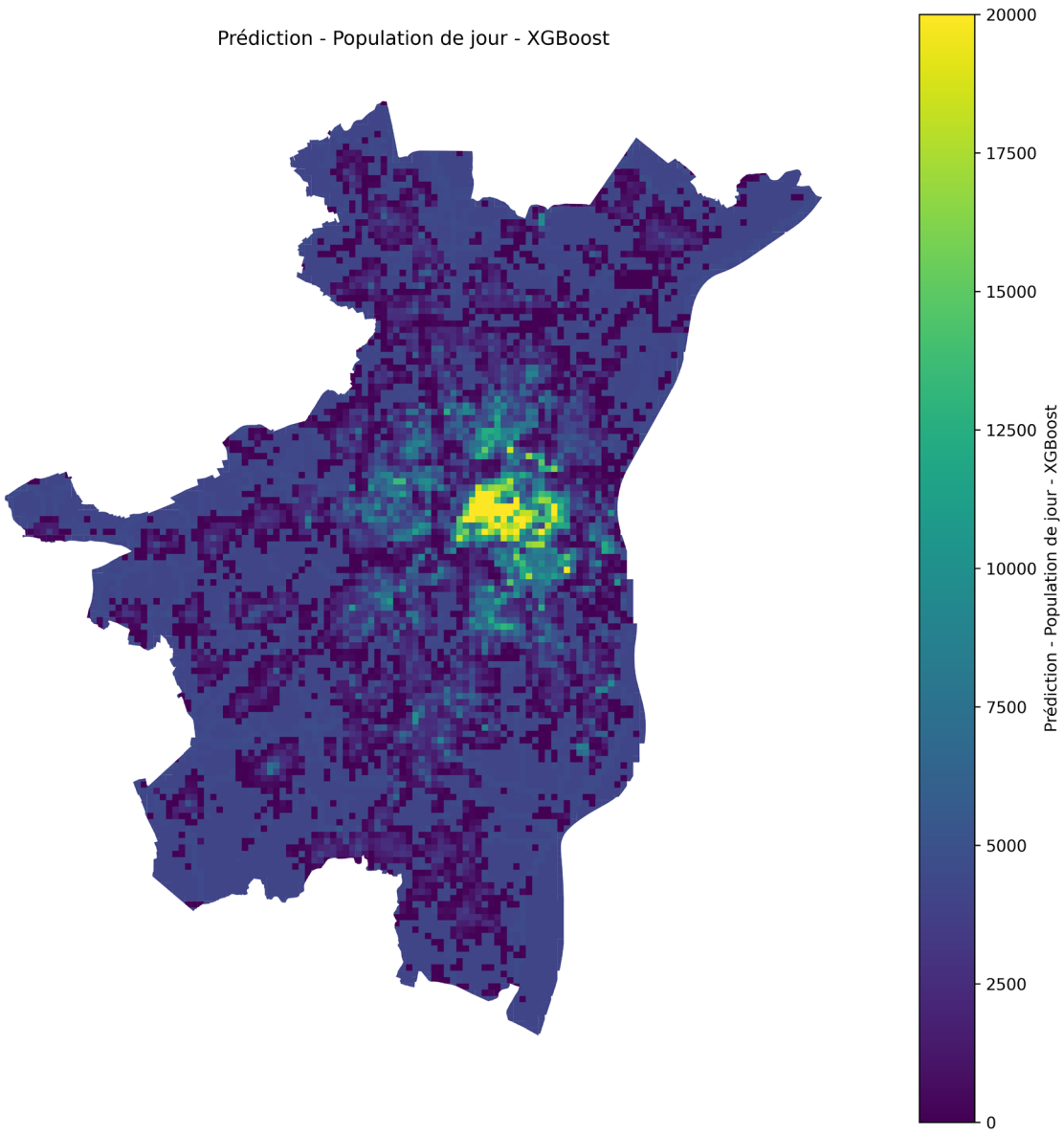


Figure 30: Prédiction de la population diurne - Eurométropole de Strasbourg - XGBoost. (Ledermann, 2025)

La population de nuit montre une redistribution vers les quartiers résidentiels périphériques tels que Neudorf, Hautepierre, Meinau ou la Robertsau (Figure 31). Le Centre-ville, fortement animé le jour, se vide partiellement la nuit, à l'exception de certains secteurs mixtes. Cette bascule entre les deux temporalités permet d'observer une intensité fonctionnelle en journée centrée sur les lieux d'activités économiques, culturelles ou administratifs. Et une reconstruction résidentielle nocturne dans les quartiers d'habitat collectif ou pavillonnaire. Ces deux cartes constituent une base essentielle pour interpréter les dynamiques différentielles du territoire, en rendant visible les cycles de fréquentation qui échappent aux découpages habituels.

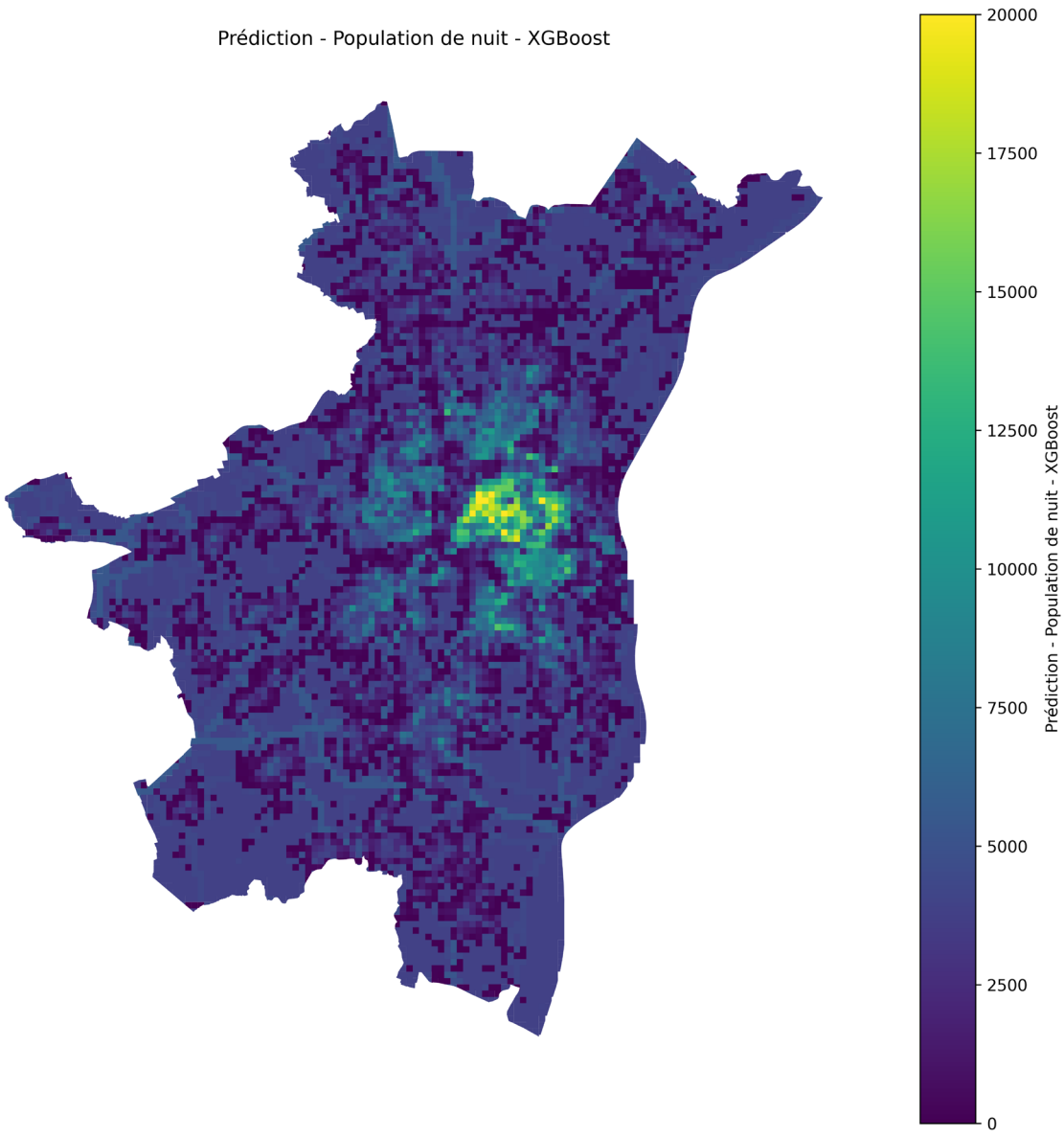


Figure 31: Prédiction de la population nocturne - Eurométropole de Strasbourg - XGBoost. (Ledermann, 2025)

4.3.2. Analyse spatiale des erreurs

Pour approfondir l'analyse, des indicateurs dérivés ont été produits afin de quantifier les écarts entre les deux temporalités. Ces métriques permettent de qualifier les variations locales et de proposer une lecture typologique des mailles du territoire. La différence absolue (Figure 32) met en lumière un gradient net entre hypercentre et périphéries. Cette différence montre des surplus de plus de 10 000 personnes en journée dans les centralités, et des déficits marqués dans les quartiers résidentiels éloignés des fonctions métropolitaines.

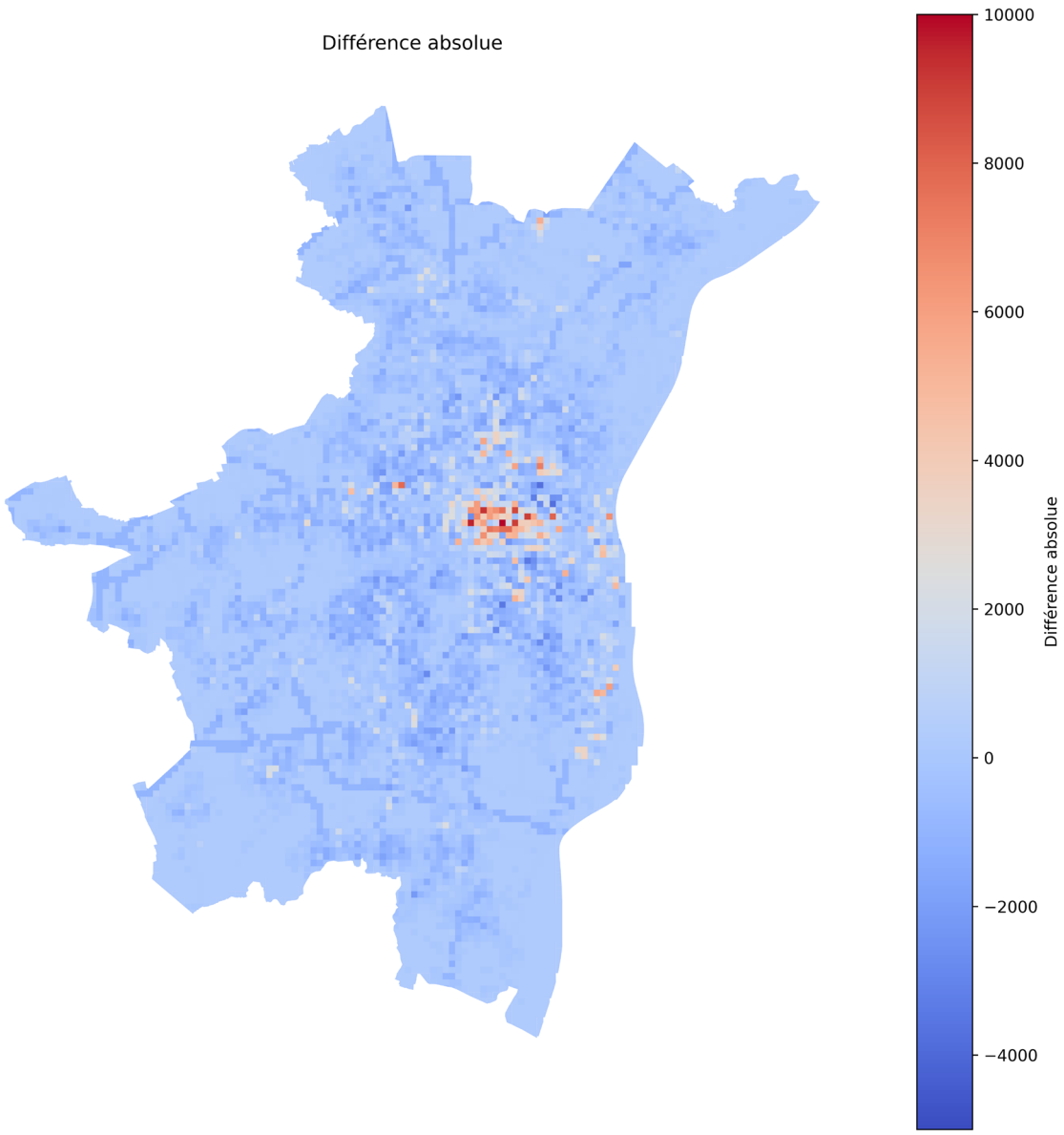


Figure 32: Différence absolue de prédiction - Eurométropole de Strasbourg - XGBoost. (Ledermann, 2025)

La différence relative (Figure 33) permet d'identifier des variations proportionnelles plus fines. Ainsi, des croissances diurnes de +200 à +300 % sont observées dans les zones tertiaires peu peuplées la nuit. A l'inverse des baisses diurnes de -40 à -60 % dans les secteurs d'habitat pur.

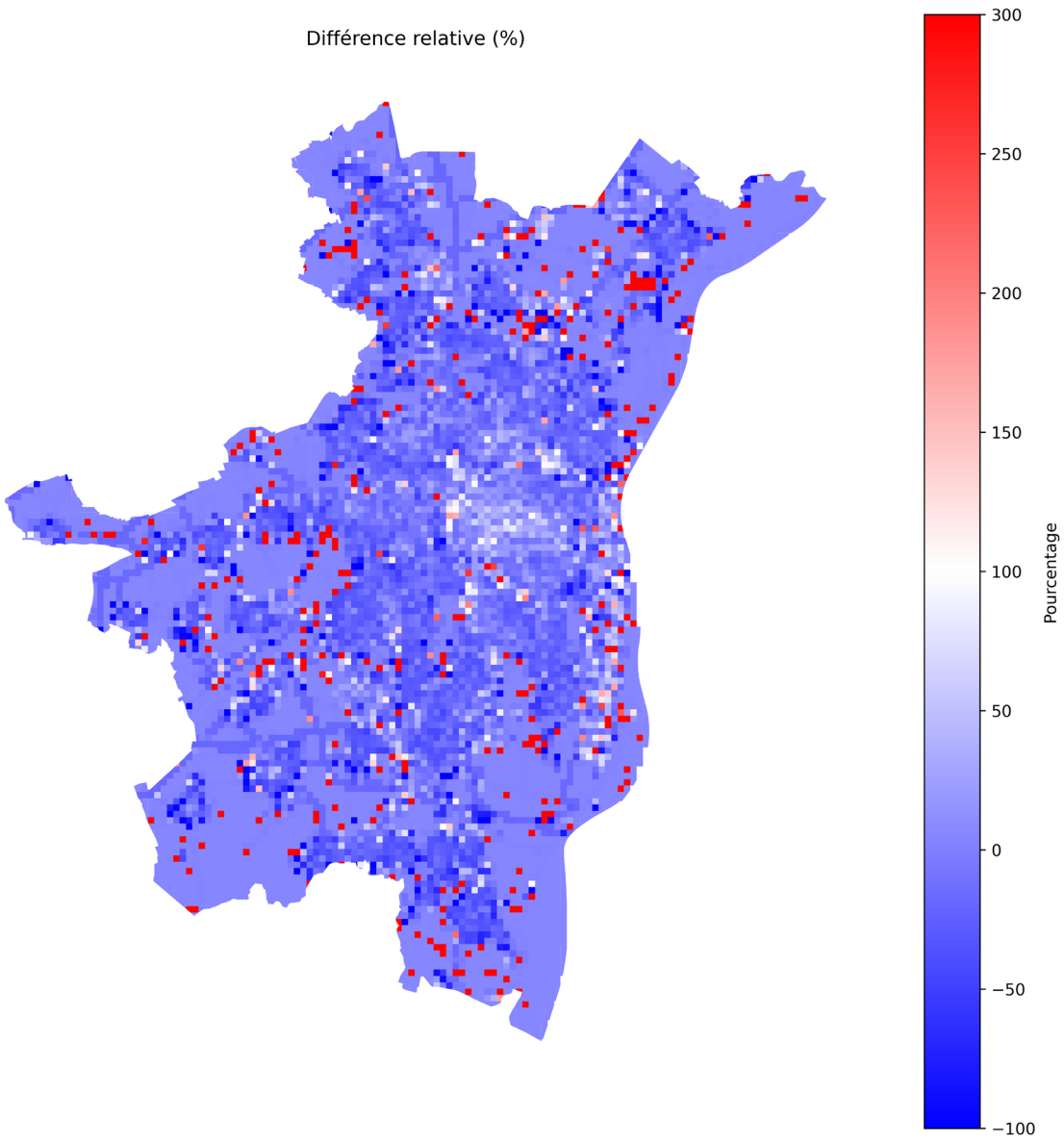


Figure 33: Différence relative de prédiction - Eurométropole de Strasbourg - XGBoost. (Ledermann, 2025)

La Figure 34 établit une typologie selon des seuils d'écarts entre les prédictions diurnes et nocturnes. Cette forme de typologie permet une compréhension d'autant plus aisée des changements journaliers. Ainsi, les centralités ressortent très attractives le jour, les quartiers résidentiels voient leurs populations augmenter la nuit et une part importante atteint le seuil d'équilibre qui pourrait paraître comme des secteurs mixtes à transition douce entre les deux temporalités.

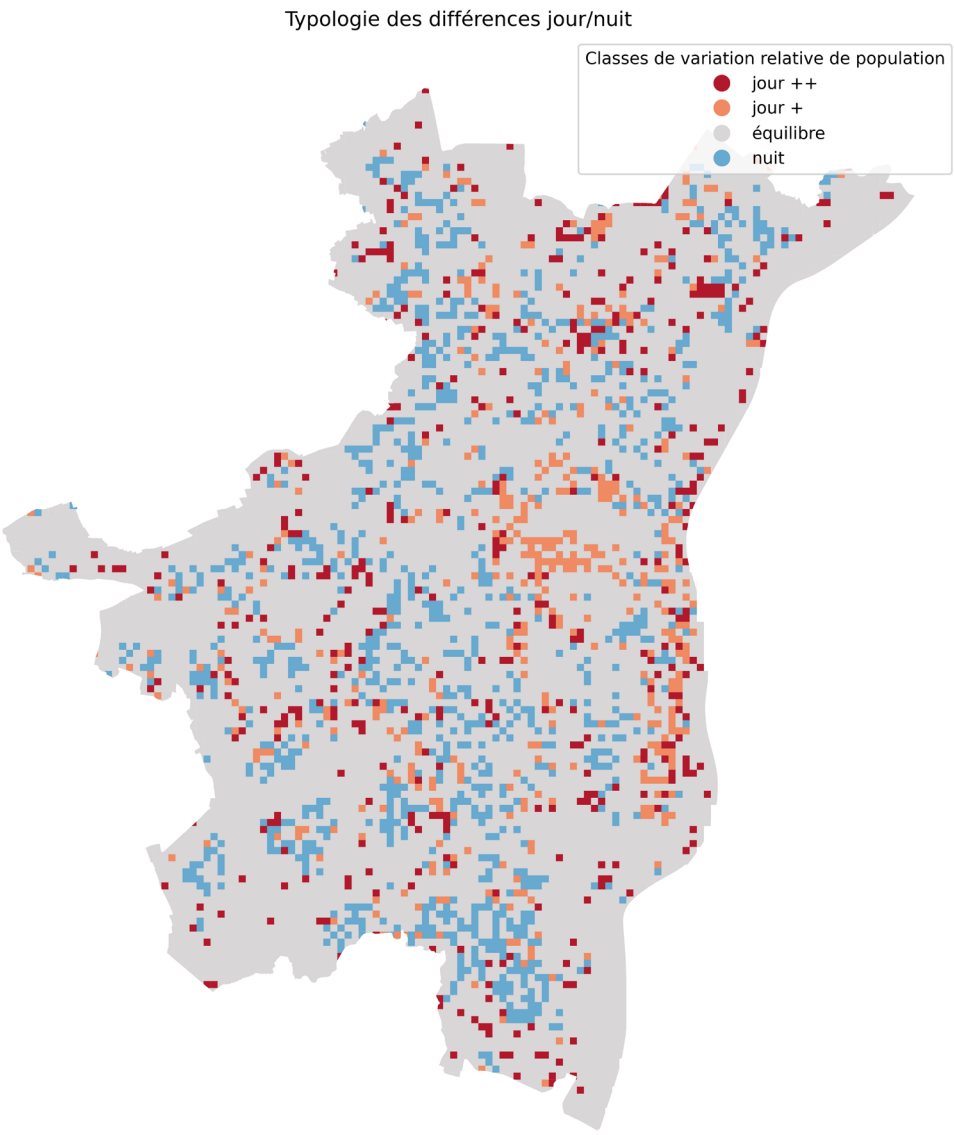


Figure 34: Typologie de différence de prédiction - Eurométropole de Strasbourg - XGBoost. (Ledermann, 2025)

Les résultats confirment la performance du modèle XGBoost, qui dépasse largement les autres approches en précision et stabilité spatiale. L'analyse des variables explicatives met en évidence le rôle central des emplois estimés, de la hauteur du bâti et de la mixité fonctionnelle, renforcé par les résultats de l'ACP et des importances. La projection locale sur un carroyage de 200 mètres à l'échelle de l'Eurométropole de Strasbourg révèle des dynamiques spatiales différenciées entre jour et nuit, avec une polarisation très nette des centralités diurnes et un recentrage nocturne vers les quartiers résidentiels. Les écarts mesurés et typologies proposées permettent une lecture optimale du territoire, au-delà des zonages classiques. Cette base analytique ouvre la voie à une discussion plus critique sur les apports, limites et usages possibles du modèle.

5. Discussion

5.1. Pertinence des variables explicatives

L'objectif de cette discussion est de mettre en perspective les résultats obtenus à travers une lecture critique, thématique et structurée. Elle s'appuie sur les différentes analyses statistiques et spatiales réalisées afin de mieux comprendre les dynamiques explicatives révélées par les modèles. Dans un premier temps, la pertinence des variables explicatives. Ensuite, sur la sensibilité du modèle aux variations de temporalités. Enfin, une réflexion sera menée sur l'apport d'une spatialisation à haute résolution.

5.1.1. Confirmation ou infirmation des hypothèses

L'évaluation des performances des modèles et l'analyse des variables explicatives permettent de revenir de manière critique sur les hypothèses formulées en amont du mémoire, afin d'en confirmer ou d'en relever les limites. La liste ci-dessous rappelle les hypothèses formulées dans la partie 1.4 :

→ **Hypothèse n°1 :** la densité d'emplois estimée est un prédicteur fort de la présence diurne.

→ **Hypothèse n°2 :** le score POI pondéré reflète le potentiel d'attractivité fonctionnelle.

→ **Hypothèse n°3 :** la compacité du bâti est un bon indicateur de l'intensité fonctionnelle.

L'hypothèse n°1 est largement confirmée. L'analyse bivariée montre un R^2 de 0,73 pour la population de jour, ce qui constitue la meilleure performance parmi toutes les variables. L'importance accordée par les modèles non-linéaires (18 % pour Random Forest, 8 % pour XGBoost) confirme son pouvoir explicatif. Cette relation s'inscrit dans la continuité des travaux de Boeing (2018) et de Sun et al. (2024), qui établissent une forte corrélation entre lieux d'emplois et densité diurne. Pour l'hypothèse n°2, celle-ci est partiellement validée. Le score POI présente un R^2 bivarié de 0,47 pour la population de jour, mais son importance chute fortement dans les modèles non-linéaires. Il est probable que ce score soit redondant avec d'autres variables comme l'indice de mixité fonctionnelle. Cela rejoint les précautions méthodologiques avancées par Panczak et al. (2020), qui soulignent les biais potentiels liés à l'usage de proxies fonctionnels. Enfin, l'hypothèse n°3, l'indicateur de compacité, mesuré via le shape index moyen, affiche des résultats faibles dans l'ensemble des analyses (R^2 proche de 0, importance marginale dans les modèles complexes). A l'inverse, d'autres variables morphologiques comme la hauteur pondérée ou l'écart-type des hauteurs jouent un rôle beaucoup plus déterminant. Cette hypothèse est donc infirmée dans sa forme initiale, et mériterait d'être reformulée à partir d'indicateurs plus robustes de densité verticale, comme le volume moyen ou la hauteur moyenne pondérée (Biljecki & Chow, 2022). L'analyse croisée des résultats confirme la validité des hypothèses relatives à l'emploi et à la verticalité bâtie, mais remet en

question la pertinence isolée des POI et de la compacité comme indicateurs explicatifs. Ces écarts soulignent la nécessité d'une sélection rigoureuse et contextuelle des variables proxy, dans la lignée des critiques méthodologiques formulées dans la littérature récente.

5.1.2. Variables les plus robustes

Identifier les variables les plus robustes revient à isoler celles qui conservent une forte capacité explicative, quel que soit le modèle ou la temporalité considérée. Les emplois estimés s'imposent comme la variable la plus performante en analyse bivariée (R^2 = 0,73 de jour, 0,68 de nuit) et restent pertinents dans les modèles complexes, notamment en Random Forest (18 %) et en XGBoost (présent dans le top 5). Sa stabilité entre les modèles et entre les deux temporalités en fait un indicateur transversal de la densité de présence humaine. Ce résultat conforte les analyses de Boeing (2018) et Batista e Silva et al. (2020), qui soulignent le rôle fondamental de l'activité économique dans la structuration des flux urbains. Ensuite la hauteur moyenne pondérée obtient un R^2 bivarié élevé (0 ,66 de jour comme de nuit), et une importance forte dans tous les modèles (11 % RF/14 % XGB de jour, 9 % RF/18 % XGB de nuit) , elle confirme que la verticalité du bâti est un proxy solide de la densité humaine. Sa robustesse sur les deux temporalités et son pouvoir explicatif dans l'ACP renforcent sa légitimité dans le modèle. Cela valide les hypothèses de Biljecki & Chow (2022) sur le lien entre morphologie verticale et intensité d'occupation. Enfin l'indice de mixité fonctionnelle, avec un R^2 plus modeste (~0,41), mais son importance dans Random Forest (45 %) et XGBoost (jusqu'à 65 %) le place au sommet des prédicteurs selon les modèles complexes. Cette forte valeur ajoutée contextuelle et non linéaire conforme son rôle en tant que variable inconditionnelle, synthétisant la diversité des fonctions urbaines (Dovey & Pafka, 2017). Son poids élevé dans la composante principale (PC1) montre également qu'il structure l'espace urbain selon un axe socio-morphologique robuste. Ces trois variables se distinguent clairement par leur robustesse. Leur présence constante dans les résultats bivariés, multivariés et factoriels confirme leur statut de prédicteurs-clés pour la modélisation de la population dynamique, en lien avec les travaux de référence en géographie quantitative et data science spatiale.

5.1.3. Redondances et complémentarités révélées par l'ACP

L'Analyse en Composantes Principales (ACP) permet de dépasser la simple évaluation individuelle des variables pour explorer leur structure latente, identifier les proximités statistiques, et révéler des axes thématiques cohérents. L'ACP se structure autour de deux axes dominants : morpho-social et économique. La PC1 (31,7 % de la variance) regroupe les variables morphologiques (hauteur pondérée, écart-type de hauteur) et sociales (part de jeunes, mixité fonctionnelle), formant un axe socio-morphologique continu. La PC2 (13,8 % de la variance) isole les variables économiques (densité d'établissements, densité de commerces), illustrant une dimension fonctionnelle locale indépendante. Cette structuration confirme une séparation partielle entre forme urbaine globale et polarités économiques ponctuelles, comme observé chez Xuacho et al. (2019).

Plusieurs variables très proches dans le cercle de corrélation (hauteur moyenne pondérée, volume moyen bâti, écart-type de hauteur) traduisent une redondance sur la dimension morphologique. De même, la densité de commerces et la densité d'établissements se recoupent dans leur position sur l'axe économique (PC2), suggérant qu'un indicateur synthétique pourrait suffire. Cette convergence justifie une réduction de dimensionnalité en modélisation, tout en gardant des représentants par grand axe. Il y a des complémentarités révélées entre certains indicateurs, faibles isolés mais utiles combinés. Le shape index ou la largeur moyenne des voiries, peu explicatifs individuellement, prennent plus de sens dans leur interaction avec la densité bâtie ou la mixité. L'ACP permet ainsi de ne pas écarter trop vite des variables à faible pouvoir explicatif isolé, mais intégrables dans des logiques croisées. Cette lecture multidimensionnelle permet une vision plus nuancée que la simple élimination basée sur le R^2 bivarié. L'ACP révèle des structures thématiques robustes et identifie des groupes de variables fortement corrélées, justifiant à la fois des regroupements interprétatifs et des choix d'optimisation du modèle. Elle permet d'enrichir l'interprétation des dynamiques spatiales complexes en tenant compte des effets combinés, rejoignant les recommandations méthodologiques formulées par Cheng et al. (2022) sur la robustesse explicative des structures latentes.

L'analyse croisée des performances statistiques, des modèles prédictifs et de l'ACP permet de confirmer la pertinence de certaines hypothèses initiales, tout en apportant des ajustements critiques à d'autres. Trois variables ressortent comme véritablement robustes : les emplois estimés, la hauteur moyenne pondérée, et l'indice de mixité fonctionnelle. Elles traduisent des logiques socio-économiques et morphologiques bien ancrées dans l'espace urbain. L'ACP révèle par ailleurs des regroupements cohérents et des redondances thématiques qui confortent l'idée d'un modèle lisible, structuré autour d'axes explicatifs clairs. Cette compréhension approfondie des variables alimente la réflexion sur leur comportement différentiel selon les temporalités, enjeu développé dans la section suivante.

5.2. Sensibilité du modèle aux temporalités

L'un des apports majeurs du présent travail réside dans la distinction entre deux temporalités contrastées de la présence humaine : la journée (10h–16h) et la nuit (00h–6h). Cette différenciation n'est pas simplement technique : elle traduit des régimes d'occupation de l'espace fondamentalement différents, entre mobilités fonctionnelles et ancrages résidentiels. Tester le modèle sur ces deux plages horaires permet ainsi d'évaluer sa sensibilité temporelle, c'est-à-dire sa capacité à restituer des logiques spatiales variables en fonction du moment de la journée. Cette partie propose d'abord une comparaison des performances selon la temporalité, avant d'examiner les variables explicatives spécifiques à chaque période, pour enfin interroger les implications socio-spatiales de ces écarts modélisés.

5.2.1. Comparaison des performances

La modélisation séparée de la population de jour et de nuit permet d'interroger la stabilité temporelle du modèle et la capacité des prédicteurs à capter des logiques différenciées d'occupation de l'espace. Les trois modèles testés (RLM, Random Forest, XGBoost) affichent systématiquement des R^2 plus élevés pour la population de nuit. Le modèle XGBoost atteint un R^2 de 0,99 similaire pour les deux temporalités, avec des RMSE respectifs de 249 pour la nuit et 260 le jour. Cette légère supériorité nocturne s'explique probablement par une distribution spatiale plus stable et prévisible de la population résidente, comparée aux flux diurnes plus volatils (cf. Panczak et al., 2020). L'erreur quadratique moyenne (RMSE) est sensiblement plus élevée de jour dans tous les modèles, notamment en régression linéaire (RMSE = 1993 de jour contre 1576 de nuit). Cela témoigne d'une plus grande hétérogénéité spatiale de la présence humaine en journée, liée à des logiques fonctionnelles (emploi, mobilité, services) plus complexes à capter. Ce constat rejoint les critiques adressées aux modèles statistiques dans des contextes de forte variabilité socio-spatiale (Cheng et al., 2022). Dans les deux cas, la régression linéaire reste le modèle le moins performant mais le plus interprétable, tandis que XGBoost conserve la meilleure précision. Cette constance hiérarchique atteste de la robustesse de l'architecture comparative, mais interroge aussi sur l'universalité des prédicteurs face aux variations temporelles. Si la population de nuit se prête mieux à la modélisation du fait de sa stabilité spatiale, les performances observées de jour montrent les limites des proxies face aux dynamiques mobiles. Ces résultats soulignent la nécessité d'intégrer des variables plus sensibles aux rythmes fonctionnels pour améliorer la prédiction diurne.

5.2.2. Variables influentes spécifiques

Comparer les variables influentes selon les temporalités permet de révéler des logiques différenciées d'occupation de l'espace, en lien avec les fonctions urbaines dominantes à chaque moment de la journée. Le modèle de jour est fortement structuré par les variables liées à l'activité : emplois estimés, mixité fonctionnelle, score POI pondéré. Random Forest et XGBoost placent l'indice de mixité en tête, avec 45 % et 65 % d'importance respectivement, suivi par les emplois estimés et les POI. Cela reflète une concentration des présences humaines dans les centralités économiques et les espaces de service, en cohérence avec les résultats de Sun et al. (2024). En période nocturne, la hiérarchie s'inverse partiellement : la hauteur pondérée par surface devient le second prédicteur derrière la mixité fonctionnelle. L'écart-type de hauteur prend davantage d'importance, suggérant une relocalisation des présences vers des tissus résidentiels plus homogènes et structurés. La part d'emplois et les POI, à l'inverse, voient leur influence baisser significativement dans les modèles nocturnes. L'indice de mixité fonctionnelle et la hauteur moyenne pondérée conservent une forte importance de jour comme de nuit, traduisant leur rôle transversal. Ce double ancrage témoigne de leur capacité à capter des dynamiques complexes, combinant attractivité fonctionnelle et densité bâtie, comme l'ont montré Batista e Silva et al. (2020). L'analyse révèle une différenciation claire dans les variables les plus influentes selon la temporalité, les fonctions économiques

structurant l'espace diurne, tandis que les caractéristiques morphologiques du bâti dominent la prédiction nocturne. Ces résultats illustrent la pluralité des régimes de fréquentation et renforcent la nécessité de modéliser la population dans une logique temporelle différenciée.

5.2.3. Interprétation socio-spatiale

Au-delà des métriques de performance, l'analyse des différences jour/nuit soulève des questions fondamentales sur l'organisation socio-spatiale des territoires urbains et les logiques d'occupation différenciée des espaces. La projection cartographique de la population de jour révèle une concentration autour des centralités métropolitaines : cœur de Strasbourg, zone de la gare, campus universitaire, institutions européennes. Ces zones affichent des pics de fréquentation dépassant 20 000 personnes par maille, bien supérieurs aux zones périphériques. Cette polarisation traduit un modèle urbain centré sur les lieux d'activité, conforme aux observations faites dans d'autres métropoles européennes (Batista e Silva et al., 2020). La population de nuit se recentre sur les quartiers d'habitat, tels que la Meinau, Hautepierre, la Robertsau ou encore Neudorf. La dissymétrie entre la carte de jour et celle de nuit révèle des déséquilibres dans l'usage urbain, avec des zones quasi désertées en dehors des heures d'activité. Cela questionne la mixité fonctionnelle réelle de certains quartiers et pose le problème du « vide urbain nocturne » dans les secteurs monofonctionnels. Certaines zones apparaissent comme stables entre jour et nuit, affichant un écart modéré entre les deux temporalités. Ces secteurs, souvent mixtes en termes d'usage du sol (résidentiel + commerces de proximité), constituent des espaces urbains résilients. Ils incarnent le modèle d'une ville du quart d'heure (Dovey & Pafka, 2017), où les fonctions sont accessibles localement et l'occupation spatiale reste continue. Les résultats soulignent que la sensibilité temporelle du modèle permet de mettre en lumière les rythmes sociaux de la ville. Entre centralités polarisantes le jour et quartiers résidentiels refuges la nuit, la lecture socio-spatiale des cartes produites offre une grille d'analyse fine des déséquilibres d'usage et des enjeux de planification urbaine.

L'analyse différenciée des temporalités de jour et de nuit met en évidence une sensibilité réelle du modèle aux régimes d'occupation de l'espace. Les performances légèrement meilleures pour la population nocturne traduisent une stabilité spatiale plus forte, tandis que la période diurne révèle des dynamiques fonctionnelles plus complexes à prédire. Certaines variables, comme les emplois estimés ou la hauteur bâtie, confirment leur rôle structurant quel que soit le moment de la journée. D'autres, en revanche, ne conservent leur pouvoir explicatif que dans des contextes temporels spécifiques, illustrant la variabilité des logiques urbaines. Toutefois, ces résultats doivent être interprétés avec prudence. Le modèle mobilise des données indirectes, agrégées et parfois imparfaites, qui ne capturent qu'une approximation de la réalité. La projection cartographique issue du modèle ne représente pas une vérité objective, mais une estimation fondée sur des relations statistiques entre proxies spatiaux et données Mobiliscope. Comme le rappellent Panczak et al. (2020), ces approches par modélisation indirecte sont puissantes mais contextuelles, et leur usage doit rester critique. Cette prudence n'annule

en rien l'intérêt des résultats. Elle invite au contraire à les considérer comme des outils exploratoires pour comprendre les différenciations spatiales et temporelles de la ville, et non comme des descriptions absolues. Dans cette perspective, la section suivante s'intéresse à l'apport d'une spatialisation fine des présences humaines, en lien avec les usages opérationnels potentiels du modèle.

5.3. Apports de la spatialisation à haute résolution

L'un des objectifs initiaux de ce travail était de dépasser les représentations traditionnelles de la population, souvent limitées à des unités administratives (communes, IRIS), pour proposer une modélisation à haute résolution spatiale. L'usage d'un carroyage de 200 mètres permet une lecture plus fine et continue des dynamiques urbaines, tout en renforçant la lisibilité cartographique des résultats. Cette partie propose d'examiner les bénéfices analytiques et opérationnels de cette spatialisation fine, d'abord en termes de compréhension du tissu urbain, puis en envisageant ses apports pour l'action publique.

5.3.1. Lecture fine du tissu urbain

L'utilisation d'un carroyage de 200 mètres permet d'explorer les variations intra-urbaines de population avec un niveau de détail inaccessible via les zonages statistiques classiques. Le carroyage révèle des transitions progressives ou des ruptures nettes entre zones denses et peu denses, notamment à Strasbourg entre la Grande Île, les quartiers périphériques comme Hautepierre, et les zones en mutation. Cette granularité permet de mieux visualiser la continuité ou la discontinuité du tissu urbain, en lien avec les formes bâties observées (cf. Biljecki & Chow, 2022). Contrairement aux zonages IRIS ou communaux, le maillage fin permet de distinguer des micro-centralités (campus, pôles hospitaliers, zones commerciales) souvent diluées dans des unités plus larges. Il rend aussi visibles des vides fonctionnels ou des franges urbaines déconnectées, qui restent invisibles dans les données agrégées. Les cartes produites à cette échelle permettent une interprétation immédiate des formes urbaines ; allongement linéaire des axes, densités concentrées autour de hubs multimodaux ou dispersion en tissu pavillonnaire. Ce niveau de lecture affine la compréhension des dynamiques spatiales, mais doit être lu en tenant compte du caractère modélisé des données. La haute résolution ne garantit pas une précision absolue, mais une finesse d'interprétation. La spatialisation à 200 mètres offre un compromis pertinent entre finesse géographique et robustesse statistique. Elle permet d'appréhender la complexité du tissu urbain avec une meilleure sensibilité morphologique et fonctionnelle, tout en gardant à l'esprit que cette finesse reste conditionnée par la qualité des données sources et la validité des proxies mobilisés.

5.3.2. Intérêt pour les décideurs territoriaux

L'un des enjeux majeurs d'un modèle de population dynamique n'est pas seulement scientifique, mais également opérationnel. En produisant des représentations spatialisées à haute résolution, il s'agit d'outiller les décideurs locaux face aux défis de planification, de mobilité et d'aménagement. Les cartes produites permettent d'identifier les écarts temporels de présence humaine par maille, et donc de repérer les lieux de surfréquentation ou de sous-utilisation selon l'heure. Ces contrastes peuvent orienter des politiques différenciées, telles que ; le redimensionnement des services publics, l'adaptation des horaires de transport ou la requalification de secteurs peu animés. Cette approche rejoint les logiques de "ville agile" et de planification par la demande, en phase avec les réflexions sur les rythmes urbains (Vallée et Lenormand, 2024). En cartographiant les zones à déséquilibre temporel fort (ex : hypercentres vides la nuit, périphéries désertées en journée), le modèle offre une base empirique pour renforcer la mixité fonctionnelle. Il permet aussi de visualiser la sous-capacité ou la surfréquentation potentielle d'équipements (gares, écoles, hôpitaux) à l'échelle infra-urbaine. Ces éléments peuvent être mobilisés dans le cadre de diagnostics de territoire, de PLUi ou de documents de programmation stratégique. Le modèle ne donne pas un comptage réel mais une estimation statistique fondée sur des proxies et une extrapolation à partir du Mobiliscope. Les décideurs doivent être sensibilisés au caractère incertain des prédictions, et utiliser ces cartes comme un outil d'aide à la réflexion, non comme un outil de pilotage direct. Cette précaution rejoint les critiques de Goodchild (2013) sur le risque d'un usage naïf des big data spatiaux en contexte de décision publique. En dépit de ses limites méthodologiques, le modèle développé offre un potentiel réel pour les acteurs territoriaux : il permet d'introduire une lecture temporelle des dynamiques de peuplement, souvent absente des outils d'aménagement classiques. À condition d'être mobilisé de manière critique et contextualisée, il constitue un levier pour penser une ville plus adaptative, réactive et territorialisée.

La modélisation à 200 mètres permet une lecture fine des contrastes urbains et des rythmes de fréquentation, souvent invisibles aux échelles administratives. Elle offre un outil d'analyse pertinent pour les politiques publiques, à condition de rester conscient de ses limites méthodologiques. Ce niveau de détail n'est pas gage de vérité absolue, mais d'interprétation spatialement nuancée.

Cette discussion confirme l'intérêt d'un modèle explicatif spatialement riche et temporellement différencié, tout en soulignant les précautions d'usage liées aux données indirectes. Si les résultats éclairent certaines dynamiques de peuplement, ils ne sauraient être utilisés sans un regard critique sur leur construction. Il convient désormais de revenir sur les limites de cette démarche, avant d'en esquisser les perspectives futures.

6. Limites , conclusion et perspectives

6.1. Limites méthodologiques

À l'issue de cette démarche exploratoire, il est essentiel de prendre du recul critique sur les résultats produits. Si la modélisation conduite a permis de mettre en lumière des logiques spatiales et temporelles de la population présente, elle repose sur des hypothèses fortes, des approximations et des choix méthodologiques qui en conditionnent la portée. Cette dernière partie vise ainsi à expliciter les principales limites du travail mené, à tirer un bilan synthétique de l'approche développée, et à ouvrir des pistes d'amélioration et de prolongement. Dans cette optique, cette sous-partie s'attache à exposer les limites proprement méthodologiques du modèle, qu'elles soient techniques, statistiques ou épistémologiques.

6.1.1. Contraintes techniques et pratiques

Le développement d'un modèle de cartographie dynamique repose sur une chaîne technique complexe, mobilisant des données massives, des traitements géospatiaux avancés et un environnement de programmation stable. Ces choix techniques ont conditionné la faisabilité et la reproductibilité du projet. Les bases mobilisées (BD TOPO, OSM, SIRENE, Mobiliscope…) présentent des formats variés, des niveaux de détail inégaux, et nécessitent de nombreuses opérations d'harmonisation (reprojection, nettoyage, filtrage). Certains traitements géométriques lourds, comme les intersections surfaciques complexes, ont montré leurs limites en Python, nécessitant un recours ponctuel à QGIS, ce qui rompt partiellement la logique 100 % scriptable. Aussi, le volume des données (plusieurs Go en GeoParquet) a imposé des arbitrages de performance, parfois au détriment de la finesse théorique souhaitée. De plus, la structuration modulaire du pipeline Python a permis une certaine reproductibilité, mais chaque ajout de variable ou changement d'échelle a nécessité des adaptations manuelles. La gestion des erreurs, la maintenance du code et la vérification des résultats intermédiaires ont mobilisé un temps important, peu visible dans le résultat final mais essentiel à la fiabilité globale. La portabilité du pipeline sur un autre territoire est possible, mais conditionnée par la qualité des données locales et l'ajustement des paramètres. Bien que le projet s'appuie exclusivement sur des outils open source (geopandas, shapely, scikit-learn…), leur documentation est parfois lacunaire, et les comportements diffèrent selon les versions. Certaines bibliothèques sont encore instables pour des traitements volumineux ou multi-couches, ce qui limite l'usage du pipeline à des utilisateurs confirmés. Enfin, l'absence d'interface graphique ou de solution "clé en main" limite l'accessibilité du modèle pour les collectivités locales ou les acteurs non techniques. Les contraintes techniques rencontrées n'ont pas invalidé la démarche, mais elles rappellent que le passage du concept au code nécessite des compromis entre rigueur scientifique et faisabilité informatique. Ces limites doivent être prises en compte si l'on envisage une diffusion ou une généralisation du modèle à d'autres contextes territoriaux.

6.1.2. Limites statistiques et géographiques

Au-delà des aspects techniques, la validité du modèle repose sur des hypothèses statistiques fortes et sur une structuration spatiale dont la généralisation doit être interrogée avec prudence. Ainsi, les variables explicatives mobilisées ne mesurent pas directement la présence humaine mais des facteurs supposés la favoriser (emplois, forme urbaine, POI…). Cette logique de modélisation indirecte repose sur des corrélations contextuelles, non sur des causalités établies, comme l'ont souligné Panczak et al. (2020). Un même indicateur (ex : densité de commerces) peut avoir un sens très différent selon les contextes territoriaux (centre historique vs. zone périurbaine). Ensuite, la variable cible est issue du Mobiliscope, lui-même construit à partir d'enquêtes de mobilité modélisées, fondées sur des typologies socio-spatiales. Cela induit une homogénéisation du comportement au sein de secteurs typés, pouvant lisser des différences locales importantes. De plus, l'absence de certains territoires (villes petites ou rurales) dans la base Mobiliscope introduit un biais de couverture spatiale. Le modèle ignore des facteurs exogènes ou événementiels qui influencent la présence humaine : météo, calendrier, contexte sanitaire, télétravail, etc. Il repose sur une journée-type théorique, ce qui réduit sa sensibilité aux temporalités fines et empêche toute analyse diachronique. Cette absence de temporalisation contextuelle limite l'usage du modèle pour anticiper des évolutions ou des ruptures. Enfin, le modèle est entraîné à l'échelle nationale mais appliqué localement. Cette généralisation suppose une stabilité des relations entre variables, ce qui est loin d'être garanti. Les structures urbaines, les fonctions locales et les modes de vie varient fortement entre régions. Ce que le modèle capte à Strasbourg peut ne pas être valable à Marseille ou Dunkerque. Les limites statistiques et géographiques du modèle rappellent qu'il s'agit d'un outil exploratoire, sensible aux contextes et aux hypothèses sous-jacentes. Sa validité dépend autant de la qualité des données que de la pertinence des proxies dans chaque territoire. Il constitue un cadre d'analyse utile, mais ne peut prétendre à une vérité universelle ni à une précision absolue.

6.2. Conclusion générale

Ce mémoire avait pour ambition de proposer une méthode de modélisation de la population présente à une maille fine, en s'appuyant uniquement sur des données ouvertes et reproductibles. En croisant des données morphologiques, fonctionnelles et économiques avec les estimations issues du Mobiliscope, il a permis de construire un modèle statistique capable de restituer les grandes logiques spatiales et temporelles de présence humaine.

Les résultats obtenus sont encourageants. Le modèle XGBoost atteint des niveaux de performance élevés, et les prédicteurs identifiés – emplois estimés, hauteur pondérée, mixité fonctionnelle – confirment la pertinence des hypothèses initiales. La projection à l'échelle de l'Eurométropole de Strasbourg met en évidence des contrastes nets entre zones centrales et périphériques, et révèle des dynamiques de fréquentation différenciées entre le jour et la nuit. Enfin, la spatialisation à haute résolution apporte une

lecture fine du tissu urbain, susceptible d'alimenter des diagnostics territoriaux ou des réflexions d'aménagement.

Mais ces apports doivent être lus avec prudence. Le modèle repose sur des données indirectes, agrégées, souvent approximatives. Il extrapole à partir de comportements moyens et ne peut prétendre à une vérité locale. Il s'agit d'un outil d'exploration, pas d'un outil de mesure. Sa validité dépend du contexte, de la qualité des données sources et des hypothèses méthodologiques sous-jacentes. Il ne saisit ni les temporalités fines, ni les effets conjoncturels, ni la complexité des comportements humains.

Pour autant, il ouvre des perspectives réelles. Il démontre qu'il est possible, avec des ressources ouvertes et des outils accessibles, de produire une cartographie dynamique crédible, cohérente et reproductible. Il montre que la géographie quantitative peut dialoguer avec les sciences de la donnée sans renier sa rigueur critique. Et il affirme la nécessité de dépasser les représentations figées du territoire pour penser la ville dans ses rythmes, ses usages et ses instabilités.

Ce travail constitue ainsi une étape dans la construction d'un outillage méthodologique pour une géographie des présences. Il ne répond pas à toutes les questions, mais il propose une manière d'y entrer. À ce titre, il est autant un modèle qu'une invitation à en imaginer d'autres.

6.3. Améliorations futures : enrichissement des données, modèles alternatifs, temporalités fines

Les limites identifiées dans les sections précédentes ne doivent pas être perçues comme des obstacles définitifs, mais comme autant de leviers d'amélioration. Si le modèle développé repose sur une base solide et reproductible, il n'en demeure pas moins perfectible, tant sur le plan des données que des méthodes et des dimensions temporelles explorées. Cette dernière sous-partie esquisse des pistes concrètes pour renforcer la robustesse, la sensibilité et l'utilité du modèle dans de futures itérations. Ces propositions visent à élargir la portée analytique tout en maintenant l'esprit de sobriété méthodologique et d'ouverture technologique qui a guidé ce travail.

6.3.1. Améliorations du pipeline

Le pipeline mis en place constitue une base fonctionnelle et reproductible, mais plusieurs améliorations techniques permettraient d'en renforcer la flexibilité, la robustesse et l'accessibilité. Actuellement, le pipeline nécessite une vérification manuelle à chaque étape critique (intersection, agrégation, standardisation). L'ajout de tests automatiques (unitaires et spatiaux), de logs structurés et de rapports d'étape faciliterait le diagnostic rapide d'erreurs ou d'anomalies. Cela permettrait également de sécuriser une exécution sur des volumes de données plus importants ou plus bruités. Ensuite, bien que les scripts soient modulaires, leur réutilisation par des tiers suppose une documentation explicite : structure des dossiers, formats attendus, exemples d'usage. Un notebook Jupyter de démonstration ou une interface en ligne de commande

simple pourrait grandement améliorer l'appropriation par des utilisateurs extérieurs. Aussi, le pipeline pourrait intégrer un système de détection automatique des projections, un choix dynamique de la taille de maille selon la densité urbaine, et des fonctions génériques d'import de données locales. Cette souplesse serait nécessaire pour tester le modèle dans des contextes très différents (zones rurales, littorales, transfrontalières). Enfin, les opérations d'intersection spatiale, de pondération surfacique ou de génération de variables à partir d'objets massifs (ex : BD TOPO bâtiments) pourraient être optimisées via des traitements en lot ou parallélisés (Dask, PyGEOS). Cela améliorerait le temps de traitement, actuellement long sur des départements entiers, et permettrait d'envisager des expérimentations à l'échelle nationale. L'amélioration du pipeline ne passe pas tant par une sophistication technique que par un renforcement de sa robustesse, de sa clarté et de sa transférabilité. Ces évolutions permettraient d'ouvrir l'usage du modèle à d'autres utilisateurs, d'autres contextes et d'autres échelles, sans en compromettre la reproductibilité scientifique.

6.3.2. Tests de modèle alternatifs

Le recours à trois modèles dans ce travail (RLM, Random Forest, XGBoost) a permis une première comparaison entre approches explicatives et prédictives. D'autres pistes peuvent être explorées pour affiner les résultats ou mieux saisir la complexité des relations spatiales. Ainsi, l'usage de modèles Lasso ou Elastic Net permettrait de sélectionner automatiquement les variables les plus pertinentes tout en limitant le sur-apprentissage. Cela serait particulièrement utile dans un contexte de forte colinéarité entre proxies morphologiques ou fonctionnels (Zou & Hastie, 2005). Ces modèles maintiennent une bonne interprétabilité, tout en introduisant une rigueur dans la sélection. De plus, les modèles testés ne prennent pas en compte explicitement l'autocorrélation spatiale des résidus, pourtant bien présente dans les cartes produites. Intégrer des modèles de régression spatiale (spatial lag, spatial error) ou des approches bayésiennes spatialisées permettrait de mieux capter les effets de voisinage. Ces modèles restent cependant plus exigeants en termes de calibrage et de ressources computationnelles. Des approches plus avancées, comme les Graph Neural Networks (GNN) ou les CNN spatiaux, pourraient modéliser directement les interactions complexes entre objets géographiques, à condition de disposer de données topologiquement organisées. Cette piste reste prospective, car elle impliquerait une refonte complète du pipeline et un volume de données d'entraînement plus conséquent. Enfin, bien que les modèles aient été évalués sur l'ensemble des secteurs Mobiliscope, une validation croisée géographique (leave-one-city-out) pourrait tester leur robustesse territoriale. Cela permettrait de mieux cerner les biais régionaux et la généralisabilité du modèle. Explorer d'autres modèles n'a pas pour but de maximiser les performances à tout prix, mais de mieux comprendre la structure des données, de tester leur stabilité explicative et de renforcer la légitimité statistique du modèle. À condition de rester lisibles, ces alternatives pourraient approfondir les résultats sans renier l'objectif de transparence méthodologique.

6.3.3. Valorisation possible des résultats

Au-delà de leur valeur académique, les résultats produits par ce travail peuvent alimenter des démarches concrètes d'observation, d'aide à la décision ou de sensibilisation à la dynamique des territoires. Les cartes produites à l'échelle de l'Eurométropole de Strasbourg peuvent servir de base à des diagnostics locaux : surfréquentation diurne de certains pôles, sous-utilisation nocturne, déséquilibres d'usages. Elles peuvent être intégrées dans des observatoires urbains, des études d'impact ou des documents d'urbanisme, à condition d'en expliquer clairement les limites. Une mise en forme plus pédagogique (fiches cartographiques, story map, visualisation interactive) faciliterait leur appropriation. De plus, le croisement jour/nuit permet de rendre visibles les rythmes urbains qui échappent aux représentations statistiques classiques. Ces résultats peuvent nourrir des débats publics sur l'usage des espaces, la mixité fonctionnelle ou les déséquilibres d'aménagement. Une vulgarisation raisonnée, via cartes commentées ou infographies, contribuerait à diffuser une culture des dynamiques spatiales et non simplement des densités. Le modèle proposé n'est pas un outil de pilotage, mais un instrument de lecture critique des formes urbaines et de leurs usages. Sa valorisation future repose sur une double exigence : expliciter ses limites et en faire un support de dialogue entre chercheurs, techniciens et décideurs.

Dans un contexte où la compréhension fine des dynamiques urbaines devient un enjeu central pour la recherche comme pour l'action publique, ce mémoire a proposé une méthode de cartographie dynamique des populations, à partir de données ouvertes, hétérogènes et modélisées. En construisant un pipeline reproductible, combinant traitement géospatial et apprentissage automatique, il a été possible de produire une estimation fine de la population présente à différentes temporalités, et de révéler des contrastes marqués dans l'occupation du territoire.

Le travail mené démontre que les données indirectes, bien que limitées, peuvent offrir un éclairage pertinent sur les logiques socio-spatiales contemporaines. Les résultats valident plusieurs hypothèses structurantes (poids des emplois, de la hauteur bâtie, de la mixité fonctionnelle), tout en appelant à la prudence quant à leur interprétation locale. Le modèle ne produit pas une vérité absolue, mais une grille de lecture probabiliste et contextualisée.

Plus largement, ce mémoire défend une approche critique de la modélisation spatiale : ouverte, transparente, mais pleinement consciente de ses limites statistiques, techniques et épistémologiques. Il invite à ne pas confondre sophistication algorithmique et compréhension des territoires. Il montre que la rigueur du traitement peut coexister avec l'humilité de l'interprétation.

À l'heure où les données abondent mais où leur signification reste souvent floue, cette démarche vise à réconcilier analyse quantitative et lecture géographique. Elle constitue une première pierre, modeste mais structurée, dans la construction d'outils partagés pour penser les territoires en mouvement.

Bibliographie

Batista e Silva, F., Freire, S., Schiavina, M., Rosina, K., Marín-Herrera, M. A., Ziemba, L., Craglia, M., Koomen, E., & Lavalle, C. (2020). Uncovering temporal changes in Europe's population density patterns using a data fusion approach. *Nature Communications*, *11*(1), 4631. https://doi.org/10.1038/s41467-020-18344-5

Batista e Silva, F., Marín Herrera, M. A., Rosina, K., Ribeiro Barranco, R., Freire, S., & Schiavina, M. (2018). Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. *Tourism Management*, *68*, 101-115. https://doi.org/10.1016/j.tourman.2018.02.020

Beaud, M. (2006). *L'art de la thèse. Comment préparer et rédiger un mémoire de master, une thèse de doctorat ou tout autre travail universitaire à l'ère du net*. La Découverte. https://doi.org/10.3917/dec.beaud.2006.01

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Biljecki, F., & Chow, Y. S. (2022). Global Building Morphology Indicators. *Computers, Environment and Urban Systems*, *95*, 101809. https://doi.org/10.1016/j.compenvurbsys.2022.101809

Boeing, G. (s. d.). Modeling and Analyzing Urban Networks and Amenities With OSMnx. *Geographical Analysis*, *n/a*(n/a). https://doi.org/10.1111/gean.70009

Boeing, G. (2018). Estimating local daytime population density from census and payroll data. *Regional Studies, Regional Science*, *5*(1), 179-182. https://doi.org/10.1080/21681376.2018.1455535

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cerón-Palma, I., Sanyé-Mengual ,Esther, Oliver-Solà ,Jordi, Montero ,Juan-Ignacio, & and Rieradevall, J. (2012). Barriers and Opportunities Regarding the Implementation of Rooftop Eco.Greenhouses (RTEG) in Mediterranean Cities of Europe. *Journal of Urban Technology*, *19*(4), 87-103. https://doi.org/10.1080/10630732.2012.717685

Cheng, Z., Wang, J., Zhu, K., Ge, Y., & Zhou, C. (2022). Evaluating spatial statistical and machine learning models in urban dynamic population mapping. *Transactions in Urban Data, Science, and Technology*, *1*(1-2), 37-55. https://doi.org/10.1177/27541231221114169

Cutler, D. R., Edwards Jr., T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, *88*(11), 2783-2792. https://doi.org/10.1890/07-0539.1

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F., Gaughan, A., Blondel, V., & Tatem, A. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, *PNAS Early Edition*. https://doi.org/10.1073/pnas.1408439111

Dovey, K., & and Pafka, E. (2017). What is functional mix? An assemblage approach. *Planning Theory & Practice*, *18*(2), 249-267. https://doi.org/10.1080/14649357.2017.1281996

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133-3181.

Fleischmann, M., Feliciotti, A., Romice, O., & Porta, S. (2021). *Methodological Foundation of a Numerical Taxonomy of Urban Form* (No. arXiv:2104.14956). arXiv. https://doi.org/10.48550/arXiv.2104.14956

Fox, J. (2008). *Applied regression analysis and generalized linear models*.

Freire, S. (2010). Modeling of Spatiotemporal Distribution of Urban Population at High Resolution – Value for Risk Assessment and Emergency Management. In M. Konecny, S. Zlatanova, & T. L. Bandrova (Éds.), *Geographic Information and Cartography for Risk and Crisis Management : Towards Better Solutions* (p. 53-67). Springer. https://doi.org/10.1007/978-3-642-03442-8_4

Fujiki, K. (2025). *DayPop : High resolution day and night population spatial disaggregation*. https://github.com/k-teruo/DayPop

GeoPandas. (2025). *GeoPandas : Python tools for geographic data* [Python]. GeoPandas. https://github.com/geopandas/geopandas (Édition originale 2013)

Geron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow : Concepts, tools, and techniques to build intelligent systems* (Second edition). O'Reilly Media.

Gillies, S., van der Wel, C., Van den Bossche, J., Taves, M. W., Arnott, J., Ward, B. C., & others. (2025). *Shapely : Manipulation and analysis of geometric objects* (Version 2.1.1) [Python]. https://doi.org/10.5281/zenodo.5597138

Goodchild, M. F. (2013). The quality of big (geo)data. *Dialogues in Human Geography*, *3*(3), 280-284. https://doi.org/10.1177/2043820613513392

Gramacki, P., Leśniara, K., Raczycki, K., Woźniak, S., Przymus, M., & Szymański, P. (2023). SRAI : Towards Standardization of Geospatial AI. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (Version 0.9.7) [Python]. Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3615886.3627740 (Édition originale 2022)

Greger, K. (2015). Spatio-Temporal Building Population Estimation for Highly Urbanized Areas Using GIS. *Transactions in GIS*, *19*(1), 129-150. https://doi.org/10.1111/tgis.12086

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning : With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Kutner, M., Nachtsheim, C., & Neter, J. (2004). Applied Linear Regression Model. In *Technometrics* (Vol. 26). https://doi.org/10.2307/1269508

Lévy, J. (2003). *Dictionnaire de la géographie et de l'espace des sociétés [contributions dues à Jacques Lévy]*. Belin. https://infoscience.epfl.ch/handle/20.500.14299/13805

Louppe, G. (2015). *Understanding Random Forests : From Theory to Practice* (No. arXiv:1407.7502). arXiv. https://doi.org/10.48550/arXiv.1407.7502

Mcpherson, T., & Brown, M. (2004). Estimating daytime and nighttime population distributions in U.S. Cities for emergency response activities. In *84th AMS Annual Meeting* (p. 10). https://doi.org/10.1215/9780822384625-001

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Moreau de Bellaing, L. (1997). *Jean-Pierre Fragnière, Comment réussir un mémoire, Paris, Dunod, 1996, nouvelle édition*. https://www.persee.fr/doc/homso_0018-4306_1997_num_126_4_3552

Murphy, K. P. (2012). *Machine Learning : A Probabilistic Perspective*. MIT Press.

Numpy. (2025). *Numpy : The fundamental package for scientific computing with Python* [Python]. NumPy. https://github.com/numpy/numpy (Édition originale 2010)

Panczak, R., Charles-Edwards, E., & Corcoran, J. (2020). Estimating temporary populations : A systematic review of the empirical literature. *Humanities and Social Sciences Communications*, *6*(1), 87. https://doi.org/10.1057/s41599-020-0455-y

Pandas. (2024). *Pandas 2.2.3 Documentation*. https://pandas.pydata.org/docs/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2025). Scikit-learn : Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12, p. 2825-2830) [Python]. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (Édition originale 2010)

Raczicky, K. (s. d.). *kraina-ai/srai-tutorial : A tutorial for the SRAI library*. Consulté 22 mai 2025, à l'adresse https://github.com/kraina-ai/srai-tutorial/tree/main

Requests. (2024). *Requests : A simple, yet elegant, HTTP library* [Python]. Python Software Foundation. https://github.com/psf/requests (Édition originale 2011)

Schläpfer, M., Lee, J., & Bettencourt, L. M. A. (2015). *Urban Skylines : Building heights and shapes as measures of city size* (No. arXiv:1512.00946). arXiv. https://doi.org/10.48550/arXiv.1512.00946

Sun, Y., Xie ,Jing, Wang ,Yu, Chan ,Ting On, & and Sun, Z.-Y. (2024). Mapping local-scale working population and daytime population densities using points-of-interest and nighttime light satellite imageries. *Geo-spatial Information Science*, *27*(6), 1852-1867. https://doi.org/10.1080/10095020.2023.2273826

Vallée, J., Douet, A., Le Roux, G., Commenges, H., Lecomte, C., & Villard, E. (2024). *Mobiliscope, an open platform to explore cities and social mix around the clock*. *Www.mobiliscope.cnrs.fr* (Version v4.3) [Jeu de données]. Zenodo. https://doi.org/10.5281/zenodo.11111161

Vallée, J., & Lenormand, M. (2024). Intersectional approach of everyday geography. *Environment and Planning B*, *51*(2), 347-365. https://doi.org/10.1177/23998083231174025

Williams, A. M., Foord ,Jo, & and Mooney, J. (2012). Human mobility in functional urban regions : Understanding the diversity of mobilities. *International Review of Sociology*, *22*(2), 191-209. https://doi.org/10.1080/03906701.2012.696961

Wu, B., Huang ,Hailan, & and Wang, Y. (2024). Quantifying spatial patterns of urban building morphology in the China's Guangdong-Hong Kong-Marco greater bay area. *International Journal of Digital Earth*, *17*(1), 2392832. https://doi.org/10.1080/17538947.2024.2392832

Xuacho, Y., Tingting, Y., & Naizhuo, Z. (2019). Population Mapping with Multisensor Remote Sensing Images and Point-Of-Interest Data. *ResearchGate*. https://www.researchgate.net/publication/331624261_Population_Mapping_with_Multisensor_Remote_Sensing_Images_and_Point-Of-Interest_Data

YAML. (2024). *Pyyaml : Canonical source repository for PyYAML* [Python]. The YAML Project. https://github.com/yaml/pyyaml (Édition originale 2011)

Zou, H., & Hastie, T. (2005). Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Annexes

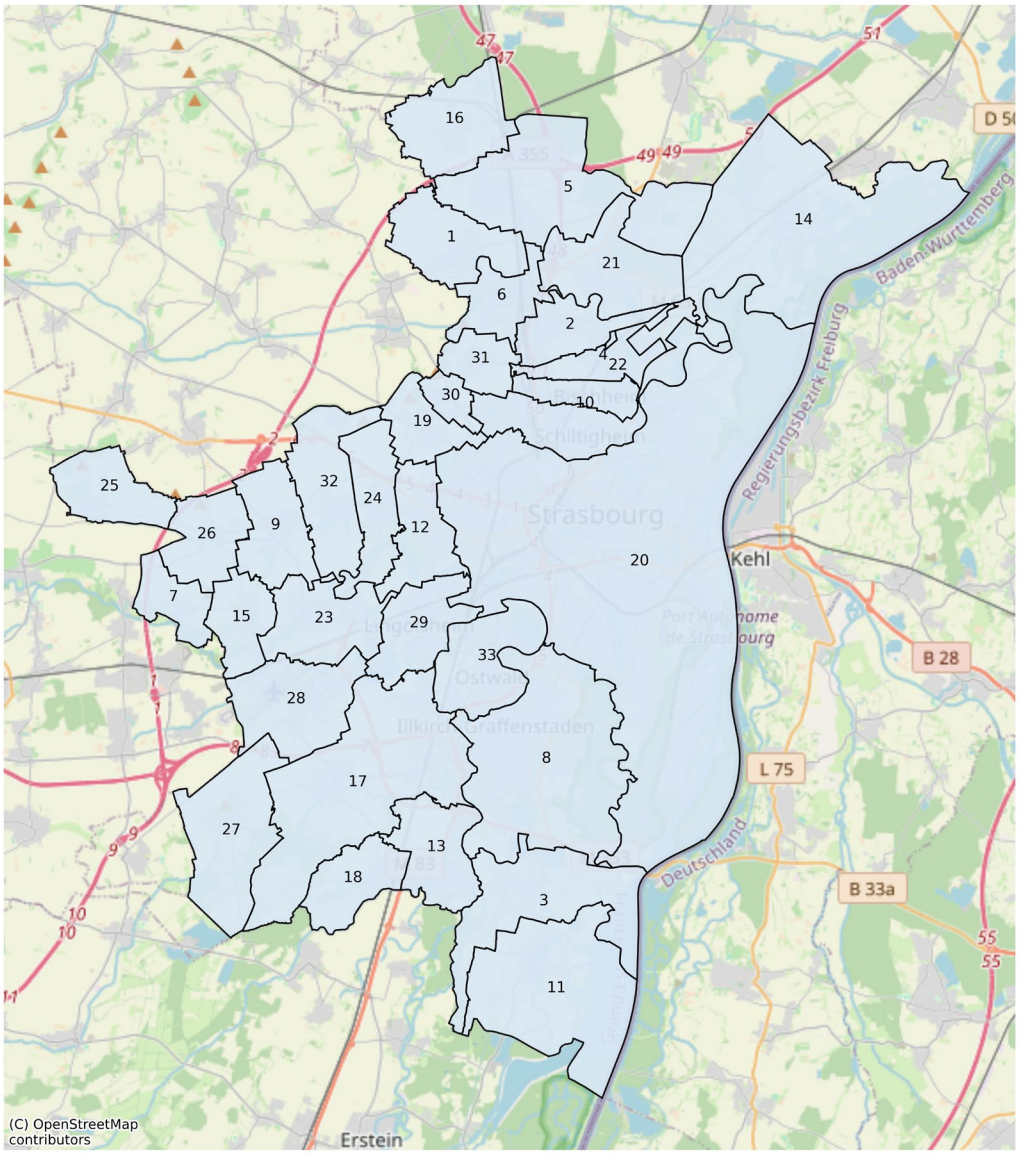
Annexe 1 : QRCode Git-Hub

---

https://github.com/quentinldrm/Model_pop

Annexe 2 : Communes EMS, carte et tableau

Communes de l'Eurométropole de Strasbourg (numérotées)



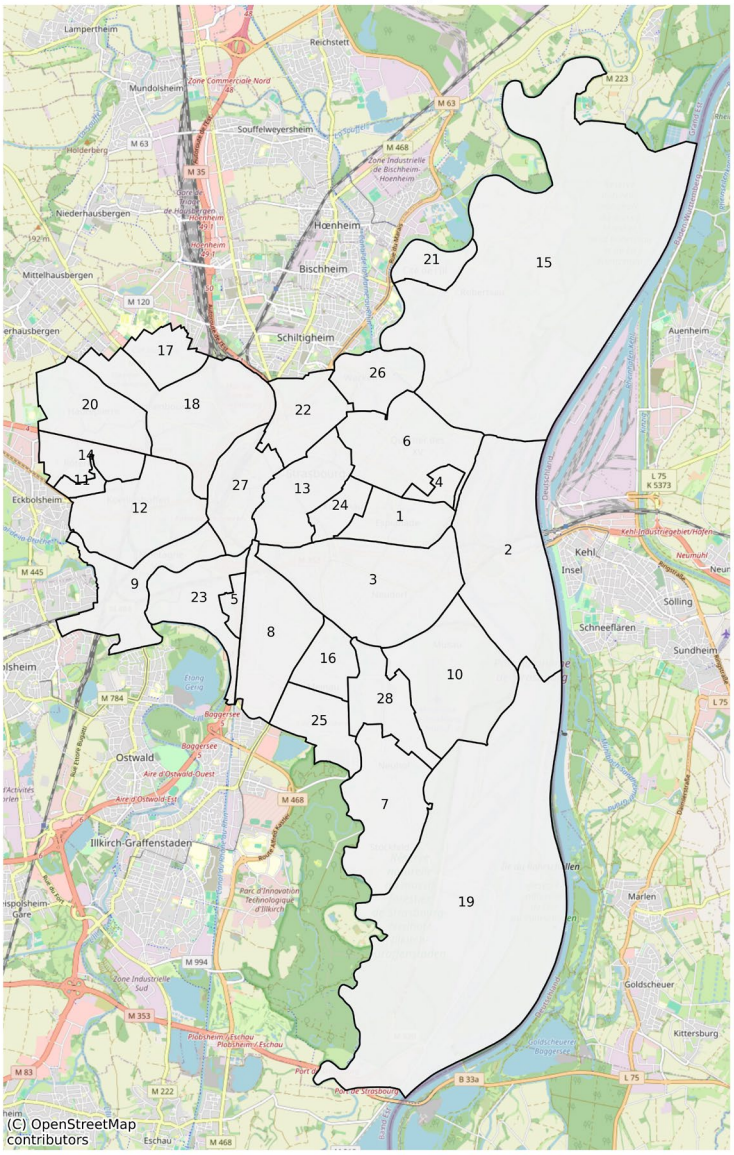
(Ledermann, 2025)

id	nom
1	Lampertheim
2	Souffelweyersheim
3	Eschau
4	Hoenheim
5	Vendenheim
6	Mundolsheim
7	Kolbsheim
8	Illkirch-Graffenstaden
9	Achenheim
10	Schiltigheim
11	Plobsheim
12	Eckbolsheim
13	Fegersheim
14	La Wantzenau
15	Hangenbieten
16	Eckwersheim
17	Geispolsheim
18	Lipsheim
19	Oberhausbergen
20	Strasbourg
21	Reichstett
22	Bischheim
23	Holtzheim
24	Wolfisheim
25	Osthoffen
26	Breuschwickersheim
27	Blaesheim
28	Entzheim
29	Lingolsheim
30	Mittelhausbergen
31	Niederhausbergen
32	Oberschaeffolsheim
33	Ostwald

(Ledermann, 2025)

Annexe 3 : Quartiers EMS, carte et tableau

Quartiers de Strasbourg (numérotés)



(Ledermann, 2025)

id	nom
1	Esplanade
2	Port du Rhin
3	Neudorf
4	Spach-Rotterdam
5	Elsau-Maison d'arrêt
6	Orangerie-Conseil des XV
7	Neuhof-Village
8	Meinau-ZA plaine des Bouchers
9	Montagne-Verte
10	Musau
11	Hohberg
12	Koenigshoffen
13	Centre-Ville
14	Poteries
15	Robertsau
16	Meinau-Villas
17	Cronenbourg - Cité Nucléaire
18	Vieux-Cronenbourg
19	Neuhof-Port autonome-forêt
20	Hautepierre
21	Cité de l'Ill
22	Tribunal
23	Elsau
24	Bourse-Krutenau
25	Meinau-Canardière
26	Wacken
27	Gare
28	Neuhof-Cités

(Ledermann, 2025)

Résumé

Ce mémoire propose une méthode de cartographie dynamique de la population présente, à partir de données ouvertes et de techniques d'apprentissage automatique. Il s'appuie sur un pipeline Python entièrement reproductible, mobilisant des variables morphologiques, fonctionnelles et économiques issues de bases telles que OSM, SIRENE ou la BD TOPO, croisées avec les données du Mobiliscope. Trois modèles sont testés (régression linéaire, Random Forest, XGBoost) pour estimer la population moyenne présente de jour et de nuit à une maille de 200 mètres. Les résultats révèlent une forte structuration spatiale de la population diurne par l'activité économique et une redistribution résidentielle nocturne. Le modèle XGBoost obtient les meilleures performances (R^2 ≈ 0,99). Si cette approche présente de nombreuses limites méthodologiques, elle offre néanmoins une lecture originale des dynamiques urbaines et constitue un outil exploratoire pertinent pour les chercheurs comme pour les collectivités territoriales.

Mots-clés : cartographie dynamique, apprentissage automatique, population présente, Mobiliscope, proxies spatiaux, géographie quantitative, modèle explicatif.

Abstract

This thesis develops a method for dynamic mapping of the present population using open data and machine learning techniques. A fully reproducible Python pipeline was implemented, combining morphological, functional and economic indicators from sources such as OSM, SIRENE and BD TOPO, with data from the Mobiliscope. Three models were tested (linear regression, Random Forest, XGBoost) to estimate daytime and nighttime average population at a 200-meter grid resolution. Results show a strong daytime spatial structure driven by economic activity, and a nocturnal redistribution toward residential areas. The XGBoost model achieved the highest performance (R^2 ≈ 0.99). Despite methodological limitations, this approach offers a novel perspective on urban dynamics and serves as a relevant exploratory tool for both academic research and local governance.

Keywords: dynamic mapping, machine learning, present population, Mobiliscope, spatial proxies, quantitative geography, explanatory model.