International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 14 Issue 4 April 2025 || PP 21-35

Modélisation de la densité de la population dans le district de la Vallée du Bandama (Centre de la Côte d'Ivoire) : apport de la géomatique et l'indice de « Vegetation - Temperature-Light-Population (VTLPI) »

Vafoungbé BAMBA [1], Éric M'moi Valère DJAGOUA [1] ,Vassiriki CISSÉ [1], Egomli -Stanislas ASSOUHOUN [2] , Mamadou DIARRA [3]

[1](*Centre Universitaire de Recherche et d'Application en Télédétection (CURAT), Université Félix Houphouët Boigny d'Abidjan, Côte d'Ivoire, bamba.vafoungbe@ufhb.edu.ci)*

[2](*Laboratoire des Sciences et Techniques de l'environnement (LSTE), Université Jean Lorougnon Guédé de Daloa (UJLoG), Côte d'Ivoire*)

[3](*Laboratoire de Mécanique et Informatique, Université Félix HOUPHOUËT-BOIGNY de Cocody 22 BP 801 Abidjan, Côte d'Ivoire)*

RESUME : *La taille de la population dans les zones géographiques est un indicateur crucial pour la planification urbaine et l'évaluation de l'accès aux différents services sociaux de base. Dans la plupart des pays et particulièrement en Côte d'Ivoire, les recensements de population classique ont lieu tous les dix ans, ce qui limite leur utilité pour l'établissement de statistiques régulières. L'imagerie satellitaire offre une alternative aux données d'enquête traditionnelles pour générer des indicateurs socio-économiques et socio-démographiques, offrant ainsi une nouvelle perspective pour les statistiques régulières. Un modèle statistique a été construit pour prévoir la densité de population dans la Vallée du Bandama en comparant les performances des algorithmes du Random Forest et du XGBoost. Ce modèle a utilisé des indices dérivés de l'imagerie nocturne (NTL), la NDVI et la LST, pour construire un nouvel indice VTLPI afin de modéliser la densité de la population.*

Le modèle a bien fonctionné, prédisant avec précision fort appréciable, la densité de population au niveau du district de la Vallée du Bandama. Les résultats indiquent que la densité de population estimée par le VTLPI se rapproche des données d'enquête de terrain. En utilisant les données de recensement comme référence, l'erreur absolue moyenne au niveau du District est respectivement de 0,54 et l'erreur quadratique moyenne est de 264 personnes/km2. De plus, le coefficient de détermination R^2 pour les algorithmes du Random Forest et du XGBoost sont respectivement de 0.84 et 0.87. Les résultats montrent que notre modèle basé sur le VTLPI permet d'obtenir une meilleure estimation de la densité de population dans le district et permets de caractériser davantage de variations spatiales au niveau de la grille de 250 mètres au niveau local. La densité de population obtenue offre de meilleures données sur l'exposition de la population pour l'évaluation des risques et des pertes liés aux catastrophes naturelles ainsi que pour d'autres applications connexes.

Mots Clés : *Télédétection, densité de population, imagerie satellitaire, apprentissage automatique, Vallée du Bandama, Côte d'Ivoire.*

ABSTRACT: *Population size in geographical areas is a crucial indicator for urban planning and assessing access to various basic social services. In most countries, and particularly in Côte d'Ivoire, conventional population censuses take place every ten years, limiting their usefulness for regular statistics. Satellite imagery offers an alternative to traditional survey data for generating socio-economic and socio-demographic indicators, providing a new perspective for regular statistics. A statistical model was built to forecast population density in Côte d'Ivoire and the Bandama Valley in particular, by comparing the performance of the Random Forest and XGBoost algorithms. This model used indices derived from night-time imagery (NTL), NDVI and LST, to construct a new VTLPI index to model population density. The model performed well, accurately predicting population density at the Bandama Valley district level. The results indicate that the population density estimated by the VTLPI is close to the field survey data. Using the census data as a reference, the mean absolute error at District level is 0.54 and the mean square error is 264 persons/km2 respectively. In addition, the coefficient of determination R^2 for the Random Forest and XGBoost algorithms are 0.84 and 0.87 respectively. The results show that our VTLPI-based model provides a better estimate of population density in the district, and allows us to characterize more spatial variation at the 250-meter grid level at the local level. The resulting population density*

provides better data on population exposure for natural disaster risk and loss assessment and other related applications.

KEYWORDS: *Remote sensing, population density, satellite imagery, machine learning, Bandama Valley, Ivory Coast.*

Date of Submission: 27-03-2025 Date of Acceptance: 06-04-2025

I. INTRODUCTION

La connaissance précise de la répartition de la population est essentielle pour une bonne gouvernance et une planification efficace. Les recensements traditionnels, bien que fondamentaux, ne permettent pas une mise à jour régulière des données. Pour remédier à cette situation, cette étude explore des outils de la géomatique et de l'apprentissage automatique pour estimer la densité de population dans le District de la Vallée du Bandama.

En exploitant les corrélations entre les données satellitaires et les indicateurs socio-économiques, il est possible d'obtenir des informations plus fréquentes et plus détaillées sur la répartition de la population. Cette approche a l'avantage de permettre d'étudier et comprendre les dynamiques démographiques et socio-économiques en vue de mieux appréhender les mutations socio-économiques d'un territoire.

II. CADRE DE L'ETUDE

1. Présentation de la zone d'étude

La Vallée du Bandama est un district de Côte d'Ivoire, en Afrique de l'ouest, qui a pour chef-lieu la ville de Bouaké. Elle a une superficie de 28 530 km2 et une population estimée à 1 964 929 habitants (RGPH, 2021). Ce district est situé au centre du pays, entre les districts du Woroba à l'ouest, des Savanes au nord, du Zanzan à l'est, des Lacs au sud et du Sassandra-Marahoué au sud-ouest.

Depuis le redécoupage, ce district regroupe deux régions distinctes : le Hambol et le Gbêkê.
Le district est peuplé en majorité par les Baoulés (région de Gbêkê), les djiminis et les Tagbanas (région du Hambol). Le district tire son nom du fleuve Bandama alors que le fleuve ne le traverse pas mais marque sa frontière ouest avec le district du Woroba. (Figure 1).

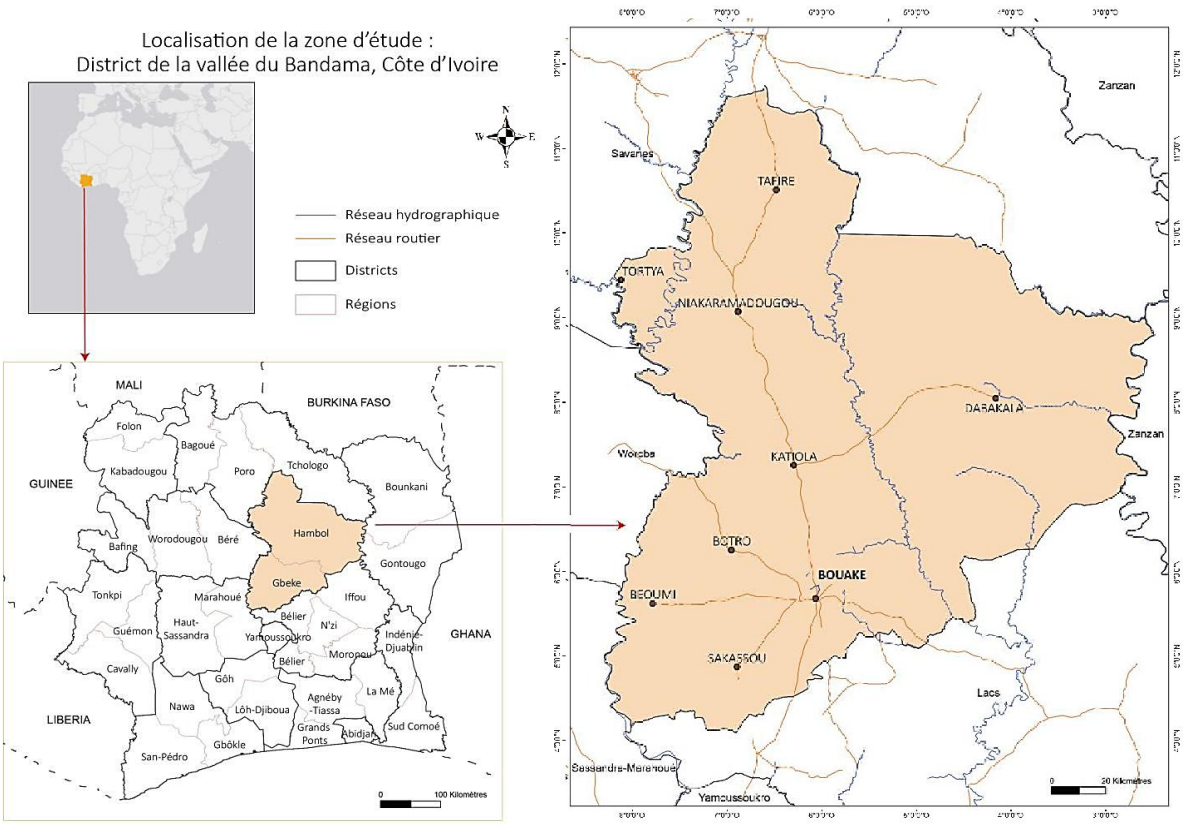

Figure 1: Localisation de la zone d'étude : District de la Vallée du Bandama

III. MATERIELS ET METHODES

1. Description des données

Les données utilisées pour mettre en œuvre la présente étude sont regroupées en deux catégories ou types que sont les variables dépendantes et les variables indépendantes.

Les variables dépendantes renferment les données géospatiales ou satellitaires tandis que les variables indépendantes concernent les données additionnelles qui sont de types tabulaires issues du RGPH 2021.

1.1. Les données satellitaires (Variables dépendantes)

1.1.1. Données collectées à partir de Google Earth Engine

Google Earth Engine est une plate-forme informatique et de stockage de données ouvertes basées sur le cloud. Cette plateforme permet d'accéder gratuitement à des images satellitaires. Dans cette étude, Google Earth Engine (GEE) est utilisé pour extraire et analyser les variables dépendantes que sont :

- L'intensité de l'éclairage nocturne (NTL)
- Les données de Température de la surface du sol (LST)
- L'indice de végétation par différence normalisée (NDVI)

Tableau 1 : Spécifications techniques des données obtenues à partir de Google Earth Engine

Données	Nom du satellite	Nom complet de la donnée	Résolution	Superficie (Approximative)	Disponibilité de la données	Durée	Fréquence
Lumière nocturne (Ancien)	DMSP OLS	DMSP OLS: Nighttime Lights Time Series Version 4, Defense Meteorological Program Operational Linescan System	30 arc seconds	1 km2/Pixel	1er Jan 1992–1er Jan 2014	27 years	Annuel
Lumière nocturne (Nouveau)	VIIRS/DNB (Day/Night Band)	VIIRS Nighttime Day/Night Band Composites Version 1	15 arc seconds	375 m2/Pixel	1er Avril 2012-1er Juin 2021	10 years	Mensuelle
Température de la surface du sol (Jour)	MODIS	MOD11A1.006 Terra Land Surface Temperature and Emissivity Daily Global 1 km	30 arc seconds	1 km2/Pixel	5 Mar 2000–22 Aou 2021	22 years	Journalière
Température de la surface du sol (Nuit)	MODIS	MOD11A2.006 Terra Land Surface Temperature and Emissivity 8-Day Global 1 km	30 arc seconds	1 km2/Pixel	5 Mar 2000–13 Aou 2021	22 years	Journalière
NDVI	Sentinel-2	MODIS Combined 16-Day NDVI	15 arc seconds	375 m2/Pixel	18 Feb 2000–14 Mar 2021	20 years	16 Jours

1.1.2. Couche mondiale sur les établissements humains (GHSL)

Les informations sur la hauteur des bâtiments ou leur utilisation (résidentiel ou non) pourraient améliorer les prévisions en matière d'estimation de la densité dans une région définie. Le projet Global Human Settlement Layer contient ces informations. Cela a été confirmé par Wardrop et *al* (2018). Pour cette étude, trois (03) couches raster ont été utilisées :

- **GHS_Built_H** : Ce jeu de données rasters spatiaux représente la distribution spatiale des hauteurs des bâtiments. Les hauteurs moyennes des bâtiments (AGBH) extraites de ces sources sont mises à jour à l'aide des marqueurs d'ombre extraits du composite de données d'images Sentinel-2 de l'année 2021 (GHS-composite-S2 R2020A). Ces hauteurs moyennes ont une résolution de 100 mètres - World Mollweide (EPSG :54009)
- **GHS_Built_S** : Ce jeu de données raster illustre la distribution des surfaces bâties, exprimée en mètres carrés. La résolution est de 10 mètres. Les valeurs allant de 0 à 100 indiquent le pourcentage urbanisé ou construit.
- **GHS_Built_C** : Cet ensemble de données raster délimite les limites des établissements humains à une résolution de 10 m et décrit leurs caractéristiques internes en termes de morphologie de l'environnement bâti et d'utilisation fonctionnelle. La zone de peuplement morphologique (MSZ) délimite le domaine spatial de tous les établissements humains à l'échelle voisine d'environ 100 m. Cette source de donnée fournit des informations sur l'occupation du sol, la densité de population et d'autres caractéristiques liées aux établissements humains.

1.1.3. Données sur la couverture terrestre

Le produit WorldCover de l'Agence Spatiale Européenne (ESA en anglais) fournit une carte de la couverture terrestre mondiale pour 2020 et 2021 à une résolution de 10 mètres. Cette couverture terrestre est basée sur les données de Copernicus Sentinel-1 et de Sentinel-2. Le produit WorldCover comprend onze (11) classes de couverture terrestre et a été généré dans le cadre du projet WorldCover, qui fait partie du 5ème programme d'observation de la Terre (EOEP-5) de l'Agence Spatiale Européenne. Les données FROM-GLC 2015, qui offrent une classification détaillée de la couverture terrestre, ont joué un rôle essentiel dans notre étude. En particulier, la classe « surface imperméable » de ces données ont été d'un apport considérable qui a servi à valider notre indice crée et à analyser la relation entre la végétation, la température et l'urbanisation.

1.2. Les Données in-situ (Variables indépendantes)

- L'Enquête Démographique et de Santé réalisée en 2021 en Côte d'Ivoire via le portail de la banque mondiale avec le lien https://dhsprogramm.com.
- Les données sur l'EHCVM (Enquête Harmonisée sur les Conditions de Vie des Ménages) de 2021 obtenues du Ministère de la Solidarité et de la Lutte contre la Pauvreté.

Tableau 2 : La répartition de la population dans le district de la Vallée du Bandama selon le RGPH 2021

Régions	Localités	Population totale	Nombre de ménages
Gbêkê	Béoumi	192.015	31.416
	Botro	117.924	18.309
	Bouaké	931.851	167.480
	Sakassou	108.110	20.817
Hambol	Dabakala	254.430	52.304
	Katiola	162.472	30.073
	Niakaramadougou	195.127	35.309
Total Vallée du Bandama		**1.964.929**	**355.708**

2. Cadre méthodologique

2.1. Méthodes d'estimation de la population et approche de l'étude

La littérature scientifique présente diverses méthodes pour estimer la population. Ces méthodes se répartissent principalement en deux catégories : l'interpolation surfacique et la modélisation statistique.

Cette étude se concentre sur la modélisation statistique (figure 2) pour estimer la densité de population à l'horizon 2030. Elle utilise une approche ascendante, combinant des données de micro-recensement avec des images satellites pour construire un modèle statistique.

La modélisation statistique vise à établir des liens entre la population et d'autres variables afin d'estimer la population totale de notre zone ; et les données de recensement (RGPH 2021) servent à élaborer le modèle. Cette approche permet également de stratifier un plan de sondage aréolaire à partir d'images satellites, facilitant ainsi la sélection d'un échantillon pour des enquêtes démographiques ou socio-économiques.

L'utilisation d'images satellites repose sur la corrélation entre les caractéristiques démographiques et socio-économiques des habitants et les caractéristiques morphologiques du milieu urbain. L'étude part du principe qu'une meilleure compréhension de l'espace intra-urbain peut améliorer les techniques d'enquête par sondage.

2.2. Méthodologie de sondage aréolaire et traitement des images

La méthode de sondage aréolaire que nous avons adoptée se déroule en deux étapes : sélection d'îlots au premier degré, puis de ménages au second degré. Les images satellites sont utilisées pour :

- Définir la base de sondage à partir de la limite de la trace urbaine.
- Stratifier la base de sondage en fonction de la densité du bâti.
- Sélectionner un échantillon d'îlots répartis géographiquement dans toute la zone.

La zone urbaine est principalement impactée par l'occupation du sol à cause de la présence humaine, ce qui rend aisé son identification. La stratification selon la densité du bâti est réalisée par classification supervisée de l'image. Cette classification repose sur la recherche de la meilleure régression entre la densité observée sur le terrain et l'indice de végétation calculé à partir d'images satellite. Les résultats de la régression sont ensuite appliqués à l'image, qui est finalement divisée en classes en fonction de la densité du bâti.

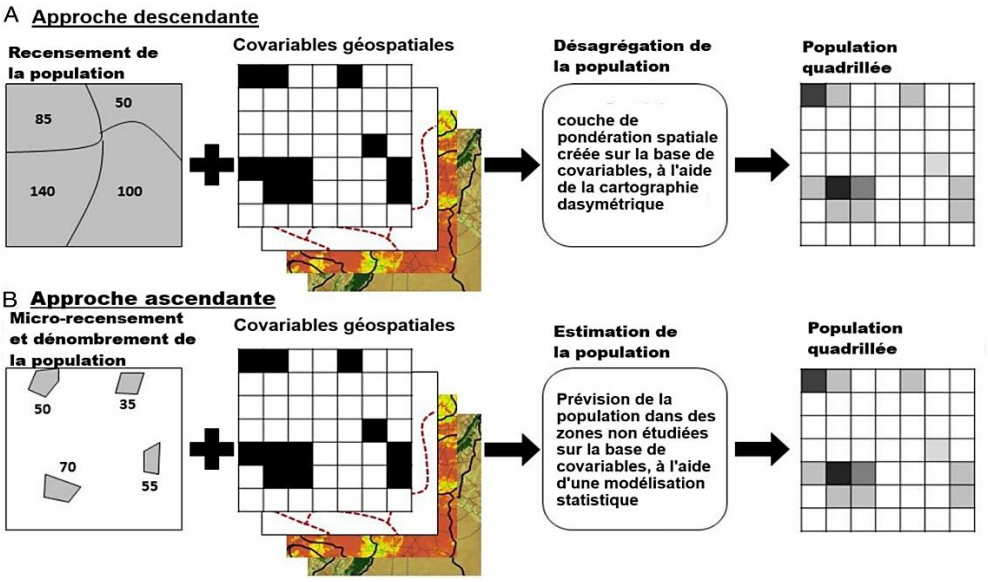

Figure 2: Schéma des approches de cartographie descendantes (A) et ascendantes (B).
Source: Wardrop, N. A., & al.

La méthodologie employée se divise en plusieurs étapes présentée dans le schéma à suivre (figure 3) :

- Etape 1 : extraction des indices susceptibles de pouvoir servir à l'estimation de la densité de population. Ce sont les variables dépendantes et les variables indépendantes. Ces données ont servi au prétraitement des données, y compris la transformation de la résolution spatiale, la normalisation et l'intégration des jeux de données.
- Etape 2 : Mise en place des deux algorithmes d'apprentissage automatique que sont le Random Forest et le XGBoost. Suivant la technique d'approche de McBride et Nicolas (2018) et Hu et *al.* (2022), 80% des données ont servis à l'entrainement des modèles et 20 % aux tests.
- Etape 3 : Validation et choix du modèle le plus pertinent basée sur le $R^{2.}$

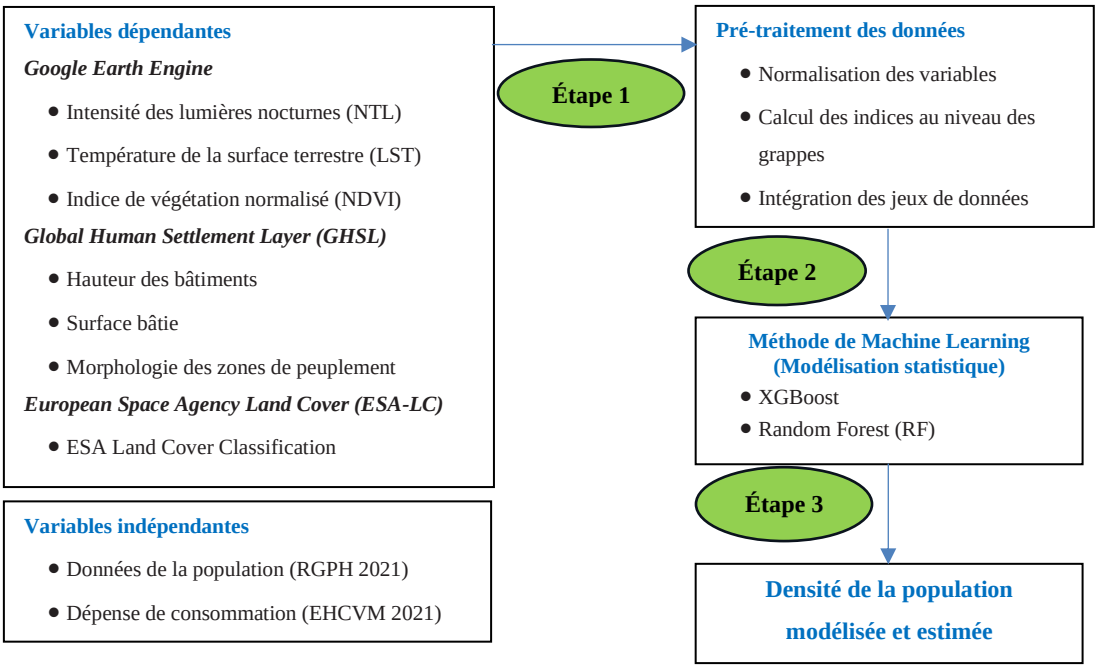

Figure 3: Cadre méthodologique de l'étude

3. Elaboration de l'indice de l'occupation humaine

Afin d'estimer la population, nous avons mis en place un nouvel indicateur en utilisant des images satellitaires multispectrales. Cette approche, inspirée des travaux de Peng Luo et *al.* (2019), combine les informations de la lumière nocturne et diurne (VIIRS-DNB), de la végétation (Sentinel-2 NDVI) et de la température de surface (Landsat-8 LST). En intégrant ces différentes sources de données, nous visons à surmonter les limitations des méthodes traditionnelles basées uniquement sur la lumière nocturne, qui peuvent sous-estimer la population dans certaines zones, notamment rurales.

Pour obtenir une estimation plus robuste de la luminosité nocturne annuelle, nous avons calculé la moyenne des images VIIRS-DNB sur une période de 12 mois (01/01/2021 à 31/12/2021). Ensuite, nous avons appliqué une transformation racine carrée suivie d'une normalisation pour améliorer la qualité des données et atténuer les effets de saturation. Ces traitements préalables permettent de mieux représenter les variations de densité de population, en particulier dans les zones rurales.

De plus, afin de réduire l'impact des variations saisonnières, nous avons calculé la moyenne annuelle de la température pour chaque pixel à partir des données LST de Landsat 8. Cette image moyenne a été normalisée pour faciliter les comparaisons. Étant donné que les zones urbaines sont généralement plus chaudes que les zones rurales, nous avons utilisé cette différence de température pour identifier les zones urbanisées.

Par ailleurs, le NDVI est largement utilisé pour étudier l'expansion urbaine. En effet, la diminution du NDVI au fil du temps est souvent liée à l'augmentation des surfaces imperméables. En sélectionnant les valeurs maximales de NDVI, nous améliorons la détection des zones urbanisées, même en présence de nuages ou de variations saisonnières. Les études de Puttanapong et *al*. (2022) ont confirmé cette relation négative entre le NDVI et l'urbanisation.

3.1. Calcul de l'indice « Végétation -Température-Light-Population » (VTLPI)

Nous avons donc mis au point un nouvel indice appelé indice de population lumineuse de la température de la végétation (VTLPI) en combinant le rapport entre le LST et le NDVI avec les données VIIRS/DNB dans le cadre de notre étude. L'équation (1), nous donne la formule de calcul de l'indice VTLPI.

$$\mathrm{VTLPI} = \frac{\sqrt{\mathrm{DNB_{norm}}}}{\mathrm{NDVI_{max}}} * \mathrm{LST_{norm}} \qquad (1)$$

*Où **DNBnorm** désigne la Lumière Nocturne/Diurne (DNB) normalisée (NTL) ;*
***NDVImax** et **LSTnorm** sont respectivement l'Indice de Végétation Normalisée (NDVI en anglais) composite maximal et la Température de surface de la Terre (LST en anglais) normalisée.*

Il convient de mentionner que l'utilisation du NDVI n'est pas une exigence générale et que d'autres indices de végétation (par exemple, l'indice de végétation amélioré, EVI) peuvent être utilisés dans le cadre de l'étude.
Pour que toutes les sources de données se situent dans la même plage entre 0 et 1, l'image LST a été normalisée à l'aide de l'équation (2).

$$\mathrm{LST_{norm}} = \frac{\mathrm{LST} - \mathrm{LST_{min}}}{\mathrm{LST_{max}} - \mathrm{LST_{min}}} \qquad (2)$$

*Où **LSTnorm** est la LST normalisée avec un intervalle de [0,1], et **LST** est la LST originale. **LSTmin**, **LSTmax** est la LST minimale et la LST maximale, respectivement.*

De même, l'image DNB est également normalisée à l'aide de l'équation (3).

$$\sqrt{\mathrm{DNB_{norm}}} = \frac{\sqrt{\mathrm{DNB}} - \sqrt{\mathrm{DNB}}_{\mathrm{min}}}{\sqrt{\mathrm{DNB}}_{\mathrm{max}} - \sqrt{\mathrm{DNB}}_{\mathrm{min}}} \qquad (3)$$

*Où **DNBnorm** est le DNB normalisé avec un intervalle de [0,1], et **DNB** est le DNB original.*
***DNBmin**, **DNBmax** est le DNB minimum et maximum, respectivement.*

3.2. Validation du modèle

L'évaluation de la précision est une étape critique pour l'estimation de la population. Outre le coefficient de Détermination R^2 et l'Erreur Moyenne des Valeurs observées (RMSE), l'Erreur Relative Moyenne (ERM) est un bon indicateur pour quantifier la performance du modèle et se calcule à l'aide de l'équation (4).

$$MRE = \frac{\sum_{i=1}^{n} |(RE_i)|}{n} \qquad (4)$$

L'ER est le rapport de la différence entre la population estimée et la population recensée et la population recensée dans la localité **i,** et **n** est le nombre total de localités dans la zone d'étude. L'erreur relative absolue médiane (ERM) est également utilisée dans cette étude car l'ERM est sensible aux valeurs extrêmes. Pour évaluer

l'indice VTLPI calculé, les deux modèles de régression de pointe pour l'estimation de la population, basés sur le DNB (ou NTL) et sur l'IDH (Indice de Développement Humain), ainsi que d'autres produits démographiques rendus publics sont choisis à des fins de comparaison, et les données de l'enquête par échantillonnage au niveau du district sont utilisées comme référence pour valider les résultats de la modélisation.

IV. Résultats et discussion

4.1. Distribution spatiale des caractéristiques environnementales

4.1.1. Carte de l'intensité de la lumière nocturne (NTL)

Afin d'obtenir une représentation de la luminosité nocturne moyenne sur une année, les images VIIRS-DNB ont été traitées (figure 4). Cette étape consiste à calculer la moyenne des valeurs de luminosité pour chaque pixel sur une période de 12 mois. Ce calcul permet de lisser les variations brusques de luminosité qui peuvent apparaître sur des images individuelles, et ainsi obtenir une image plus stable et représentative de la luminosité nocturne moyenne annuelle.

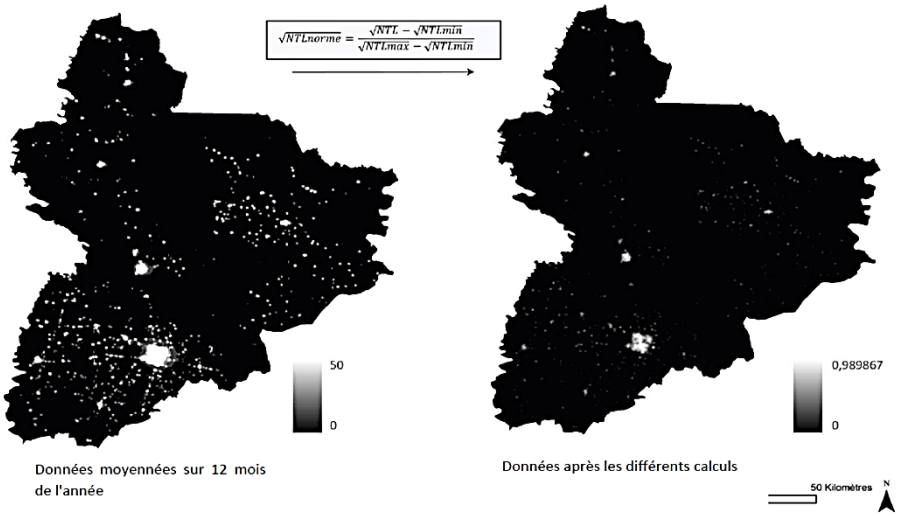

Figure 4 : Distribution spatiale de l'intensité de la lumière nocturne en 2021 dans la Vallée du Bandama

Pour améliorer la qualité de l'image obtenue, deux traitements supplémentaires ont été appliqués :

- La transformation racine carrée a été appliquée aux valeurs de luminosité de chaque pixel. Elle permet de réduire l'écart entre les zones de forte et de faible luminosité. En effet, sans cette transformation, les zones urbaines très lumineuses pourraient apparaître comme excessivement denses en population, tandis que les zones rurales peu lumineuses pourraient sembler sous-estimées.
- La normalisation des données : Les valeurs de luminosité ont ensuite été normalisées, c'est-à-dire ramenées à une échelle comprise entre 0 et 1. Cette étape permet de faciliter la comparaison des valeurs de luminosité entre différents pixels ou différentes images, en les exprimant sur une échelle commune.

Ces traitements successifs permettent d'obtenir une image de la luminosité nocturne moyenne annuelle plus précise et plus représentative de la réalité, en corrigeant les biais liés aux variations brusques de luminosité et aux différences de densité de population. Nous avons à gauche, l'image brute de la luminosité et à droite l'image traité et normalisée.

4.1.2. Carte de la distribution spatiale moyenne de la Température de la Surface de la Terre (LST)

Les données LST utilisées dans cette étude variaient initialement entre 24 et 38°C. Pour faciliter la comparaison des températures entre différentes zones géographiques et avec d'autres indicateurs, les données ont été normalisées. Cette normalisation consiste à transformer les valeurs de température pour qu'elles se situent dans une plage comprise entre 0 et 1 (figure 5) et permet de faciliter l'analyse. Le LST est un indicateur précieux pour étudier l'urbanisation et ses impacts thermiques. La normalisation des données LST permet de faciliter les analyses comparatives entre différentes zones géographiques et avec d'autres indicateur. Au regard de ce résultat, les zones moins urbanisées se situent au sud du district notamment les localités de Sakassou et Béoumi, tandis que les zones

les plus urbanisées se situent plus au nord du district et au centre avec les localités de Bouaké, Dabakala, Niakaramadougou.

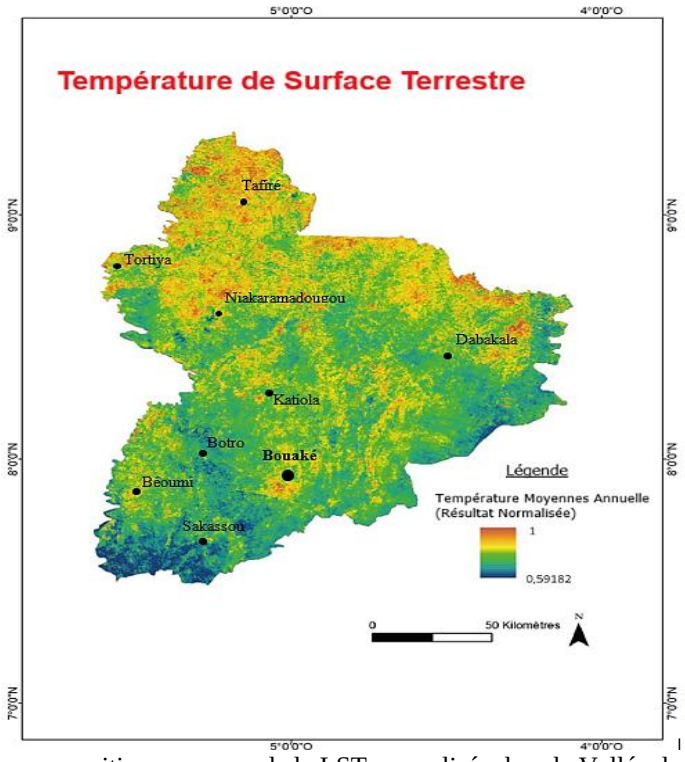

Figure 5 : composition moyenne de la LST normalisée dans la Vallée du Bandama en 2021

4.1.3. Carte de l'Indice de Végétation Normalisée (NDVI) maximal annuel

La carte du NDVI (figure 6) met en exergue la densité des activités humaines autours des principales villes que sont Bouaké, Katiola, Niakaramadougou et Tafiré.

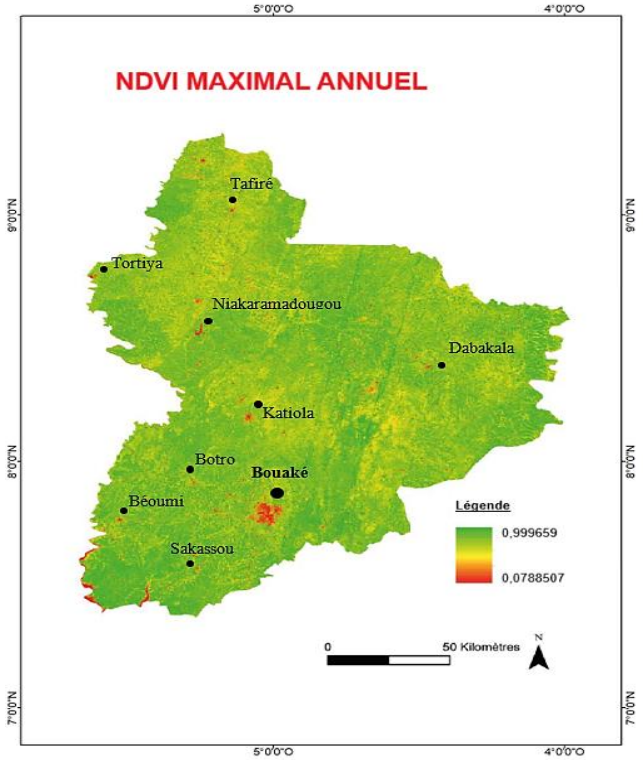

Figure 6 : Composition annuelle maximale du NDVI dans la Vallée du Bandama en 2021

4.1.4. Carte de l'indice Vegetation -Temperature-Light-Population (VTLPI)

L'indice VTLPI reflète les caractéristiques de distribution spatiale de la population dans différentes régions. La combinaison du NDVI et de la NTL permet d'extraire les couches imperméables et les aires de répartition urbaines. Les habitats humains dans les zones rurales ont souvent des valeurs NDVI élevées mais, en raison des différences d'échange de chaleur, le LST dans l'environnement urbain varie considérablement. Par conséquent, les informations LST améliore la distinction de la couche imperméable à la fois dans les zones urbaines et rurales. Il permet donc refléter la différence de répartition de la population entre les zones urbaines et rurales. L'association du NDVI, du LST et des images nocturnes a permis de différencier les noyaux urbains et de donner des informations sur la qualité de vie des habitants (par exemple, les lumières nocturnes informent sur la présence ou non d'électricité, sur la densité de bâti à travers le taux de végétation et la température du sol).

La carte de l'indice VTLPI de la figure 7 nous confirme une forte densité dans les zones urbaines, notamment les grandes agglomérations telles que Bouaké, Niakaramadou, Katiola etc.

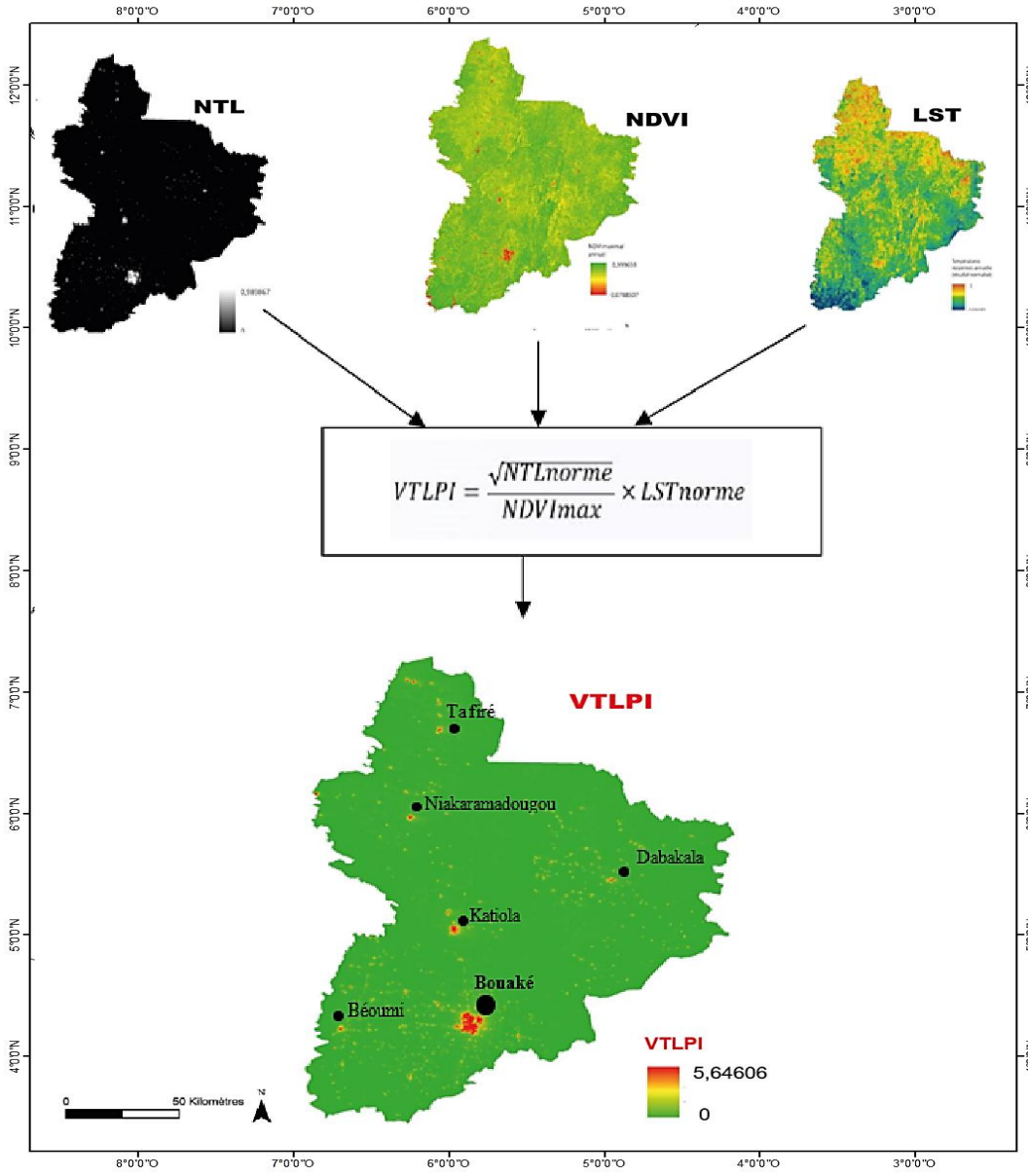

Figure 7 : Construction de l'indice VTLPI pour la Vallée du Bandama en 2021

4.1.5. Evaluation de l'indice VTLPI

L'objectif de l'indice VTLPI est double, d'une part, il doit permettre d'extraire avec précision les zones d'occupation humaine (densité de la population) et d'autre part, il doit pouvoir améliorer le signal des lumières nocturnes (NTL) en augmentant la variabilité intra-urbaine.

Pour valider la capacité de l'indice à extraire les zones d'occupation humaine, nous avons superposé une image aérienne issue de Google Earth à l'indice VTLPI. Ici, la localité est un petit village situé en milieu rural.

On peut voir que là où les valeurs de l'indice sont le plus élevées, il y a des constructions représentées par la couleur rouge sur la figure 8. Les routes et terrains nus qui ont des valeurs intermédiaires sont représentés en marron. Les espaces verts, non urbanisés sont en vert.

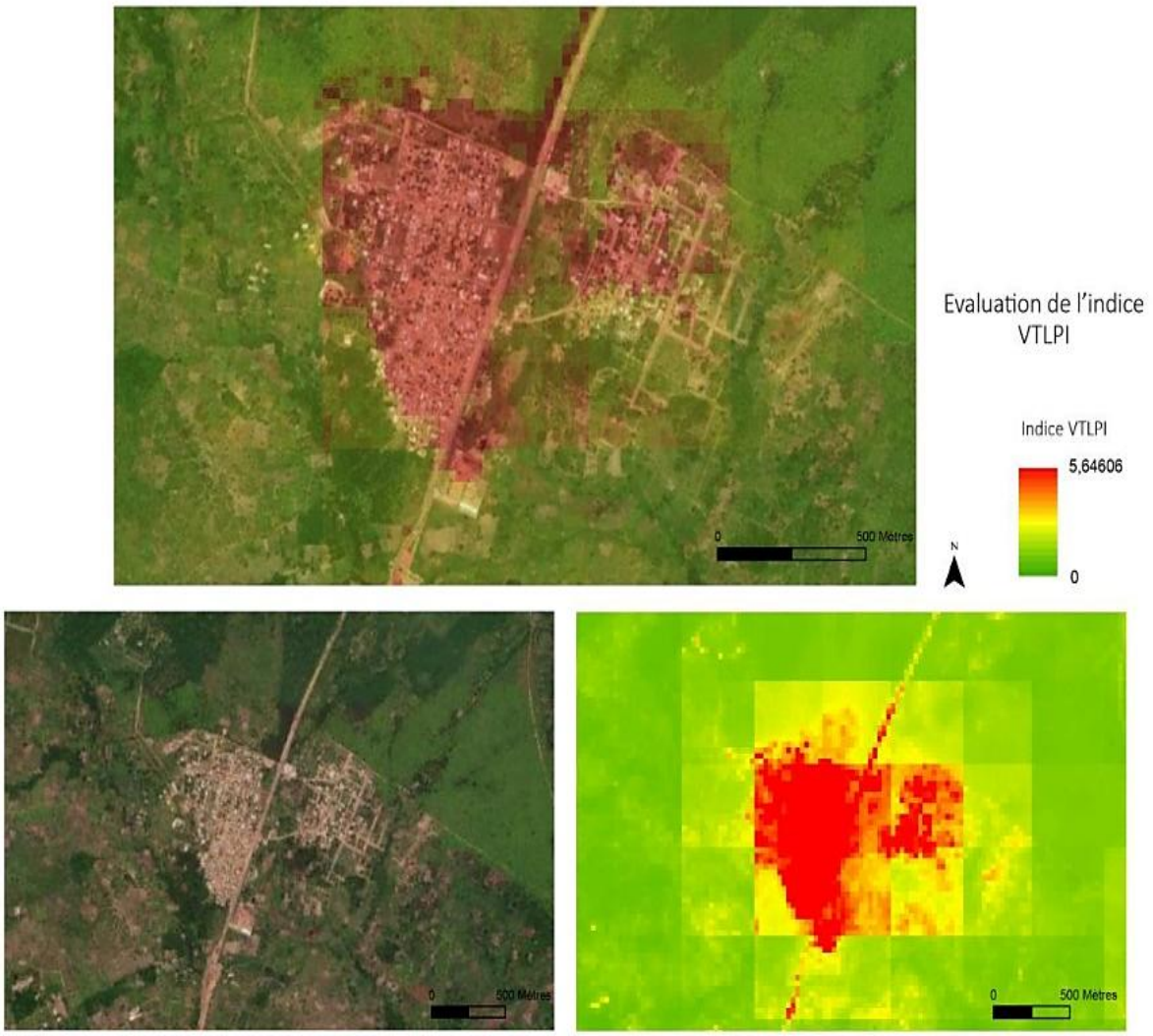

Figure 8 : Evaluation de la capacité de l'indice VTLPI à extraire les zones d'occupation humaine

Dans un second temps, nous avons cherché à vérifier si le fait de combiner les images nocturnes à d'autres variables augmentait la variabilité intra-urbaine. On peut voir sur le profil ci-dessous (figure 9) que l'indice VTLPI augmente bien la variabilité du signal NTL.

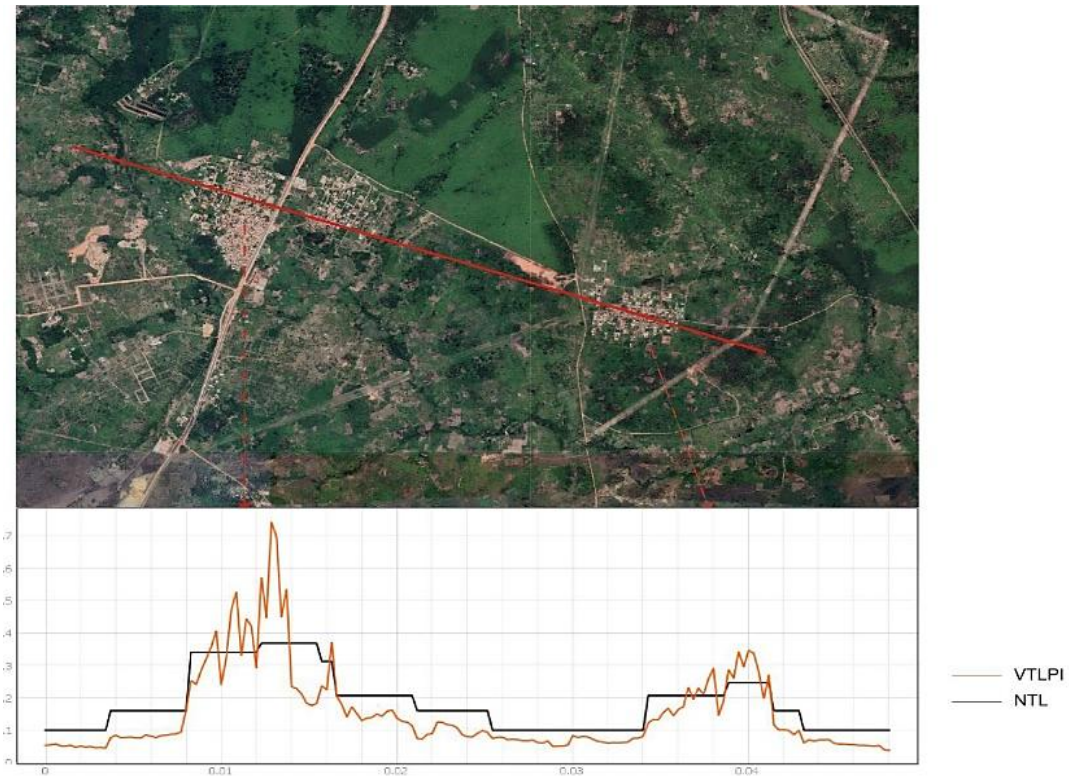

Figure 9 : Création d'un transect permettant de comparer les variations de l'indice VTLPI par rapport au données NTL

La figure 9 montre que l'indice VTLPI proposé est plus sensible aux changements des activités humaines et des composantes urbaines que la lumière nocturne (DNB). Nous constatons que lorsque l'indice VTLPI augmente, cela correspond à une augmentation des valeurs du signal de la lumière nocturne (NTL) et vice versa. Par conséquent, l'indice VTLPI peut refléter les caractéristiques de la distribution spatiale de la population dans différentes régions. Elle permet de mieux modéliser la répartition de la population, en augmentant le nombre de détails sur les changements démographiques.

4.1.6. Analyse des corrélations pour l'estimation de la densité de population

La matrice de corrélation (représentée à la figure 10) est un outil statistique qui permet de visualiser et d'analyser les relations linéaires entre les différentes variables d'un ensemble de données. Elle permet de déterminer si les variables sont indépendantes les unes des autres, ou si elles sont liées par une relation linéaire.

Dans cette matrice, chaque case représente la corrélation entre deux variables. La couleur de la case indique la force et la direction de la corrélation. Une case rouge indique une corrélation positive forte, c'est-à-dire que les deux variables varient dans le même sens. Plus la valeur de l'une augmente, plus la valeur de l'autre augmente également. Les valeurs proches de 1 indiquent une corrélation positive très forte.

L'analyse de la matrice de corrélation révèle plusieurs relations importantes :

- Corrélation entre la population et les infrastructures **:** La variable "SUM_TotalP" (population totale) présente une forte corrélation positive avec les variables "SUM_GHS_H", "SUM_GHS_S" et "Infrastructures", avec des valeurs de corrélation supérieures à 0.90. Cela signifie que les zones à forte population ont tendance à avoir une densité plus élevée d'infrastructures et de bâtiments (représentés par "SUM_GHS_H" et "SUM_GHS_S").
- Corrélation entre la population, l'indice VTLPI et le nombre de routes **:** La population est également fortement corrélée avec l'indice VTLPI (un indice de vulnérabilité) et le nombre de routes, avec des corrélations positives de 0.88. Cela suggère que les zones à forte population ont tendance à avoir un indice de vulnérabilité plus élevé et un réseau routier plus dense.

En somme, la matrice de corrélation permet de mettre en évidence les relations fortes entre la population, les infrastructures, la vulnérabilité et le réseau routier. Ces relations sont utiles pour comprendre les dynamiques spatiales et les interactions entre les différentes variables étudiées.

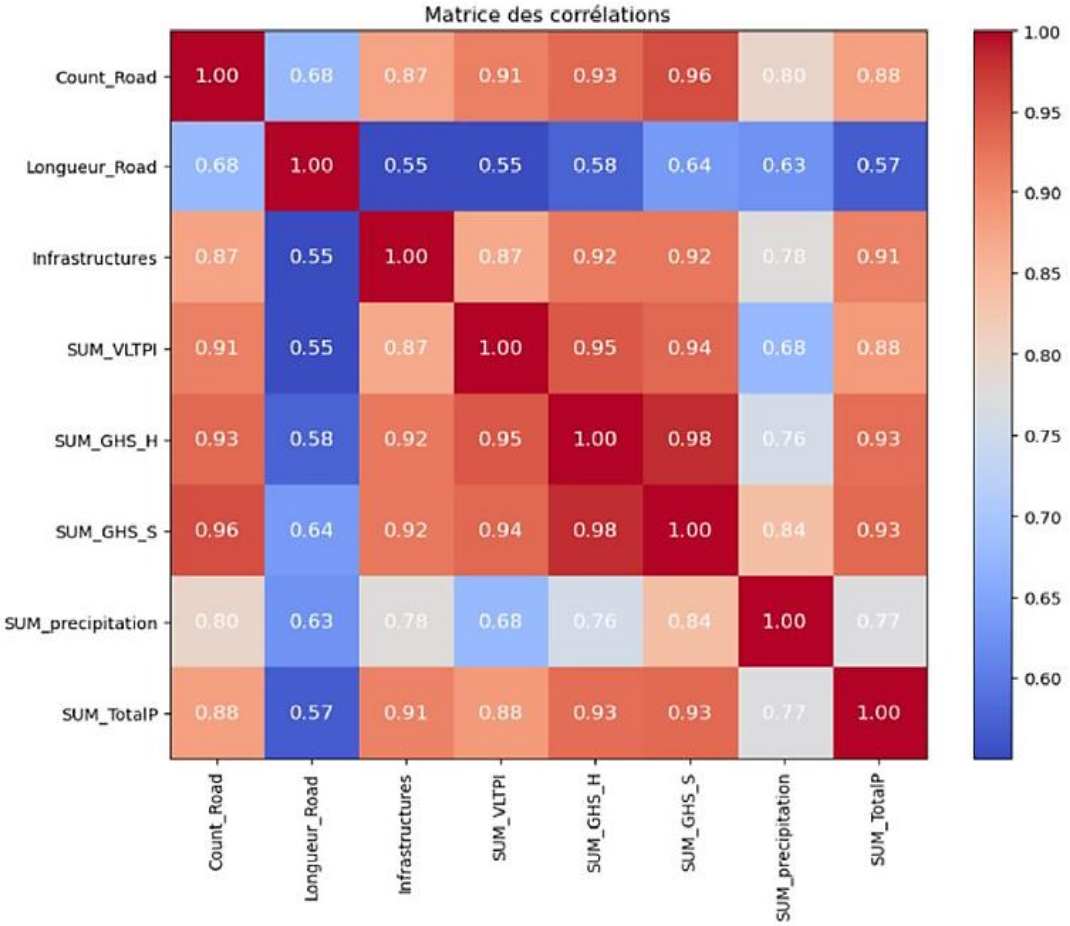

Figure 10 : Matrice représentant les corrélations entre les données de population *(Population totale)* et les différents indices GHSL

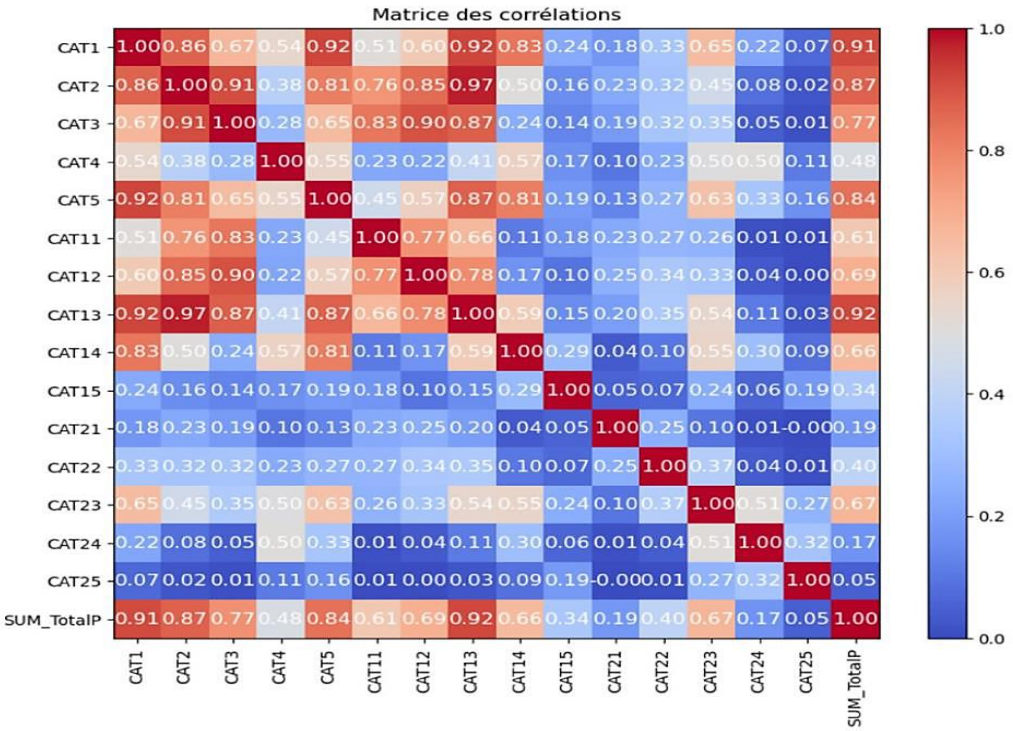

Figure 11 : Matrice représentant les corrélations entre les données de population en abscisse et les différents indices liés au LAND COVER en ordonnés

La figure 11 présente les corrélations entre les données de population et les différentes catégories de morphologie des zones de peuplement, telles que définies par le produit GHS_Built_C. Cette analyse permet de comprendre comment la répartition de la population est liée aux différents types d'environnements bâtis.
Nous observons principalement une:

- ✓ Corrélation avec les espaces ouverts et les infrastructures routières :
  - Les données de population montrent une forte corrélation avec certaines catégories de zones bâties, notamment les espaces ouverts avec une végétation basse à moyenne et les surfaces correspondant aux routes. Cela suggère que les zones habitées sont souvent associées à ces types d'environnements.
- ✓ Corrélation maximale avec les zones résidentielles de moyenne hauteur :
  - La corrélation la plus forte est observée entre la population et la catégorie 13 du GHS_Built_C, qui représente les zones résidentielles avec des bâtiments d'une hauteur comprise entre 6 et 15 mètres. Cela indique que les zones résidentielles de cette hauteur sont particulièrement associées à une forte densité de population.

En gros, cette analyse montre que les endroits où il y a beaucoup une population nombreuse ont tendance à être des endroits avec des espaces ouverts avec un peu de végétation, beaucoup de routes et surtout, des zones où les maisons ont entre 6 et 15 mètres de hauteur.
Cela nous aide à comprendre comment les gens se répartissent dans les différents types de zones bâties.

4.2. Modèle de prédiction de la densité de la population basé sur l'indice VTLPI

La figure 12 montre la distribution spatiale de la densité de la population à partir de l'indice VTLPI de 2021 à 2030.

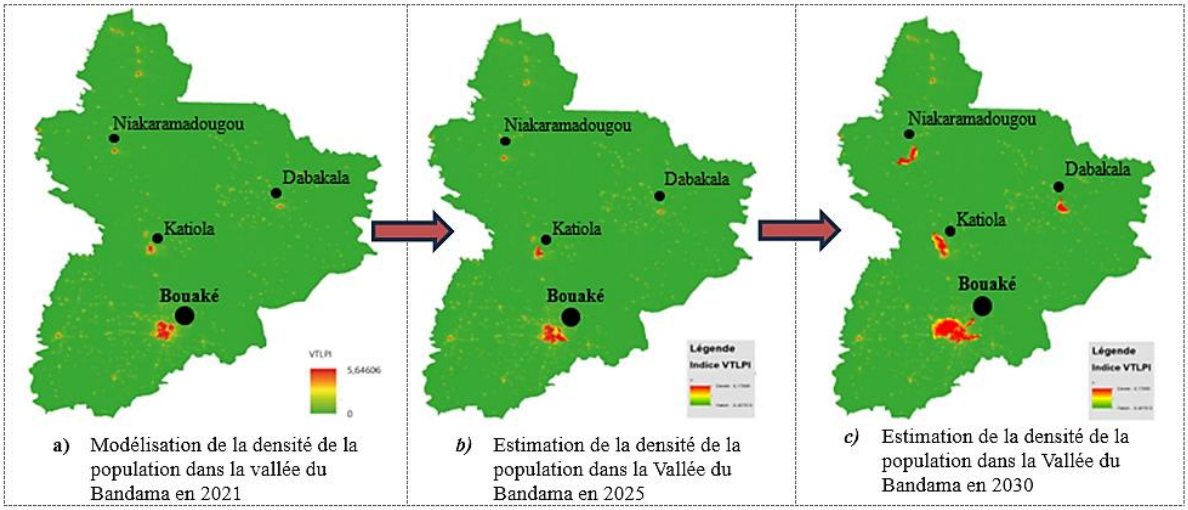

Figure 12 : Evolution de la densité de la population dans la vallée du Bandama de 2021 à 2030

La modélisation de la densité et la projection faite à travers les méthodes de Machine Learning, de 2021 à 2030 nous donne la figure 12 ci-dessus. Les résultats viennent confirmer la tendance d'urbanisation des grandes agglomérations telles que Bouaké, Katiola Dabakala etc. Cela corrobore les résultats du RGPH 2021 qui montrent une forte expansion de la population dans les zones urbaines au détriment des zones rurales.

La figure 13 ci-dessous, nous montrent à quel point nos modèles (Random Forest et XGBoost) sont bons pour l'estimation de la densité. Les points de couleur sont leurs suppositions, et la ligne noire est la réponse parfaite. Plus les points sont proches de la ligne, meilleures sont les suppositions. Le R^2 nous dit à quel point les modèles sont meilleurs. Ce qui nous donne comme valeur du R2 et moyenne des scores de validation croisée suivantes :

Random Forest :

- $R^2 = 0.85$, ce qui signifie que le modèle explique 85 % de la variance des valeurs réelles à l'échelle du district.
- Moyenne des scores de validation croisée : 0.60, indiquant une certaine variabilité dans les performances du modèle sur différents ensembles de données de validation.

XGBoost:

- R^2 = 0.87, ce qui signifie que le modèle explique 87 % de la variance des valeurs réelles à l'échelle du district.
- Moyenne des scores de validation croisée : 0.59, montrant une légère diminution par rapport au R^2. En d'autres termes, le modèle peut être moins stable lorsqu'il est appliqué à de nouvelles données par rapport à son ajustement aux données d'entraînement.

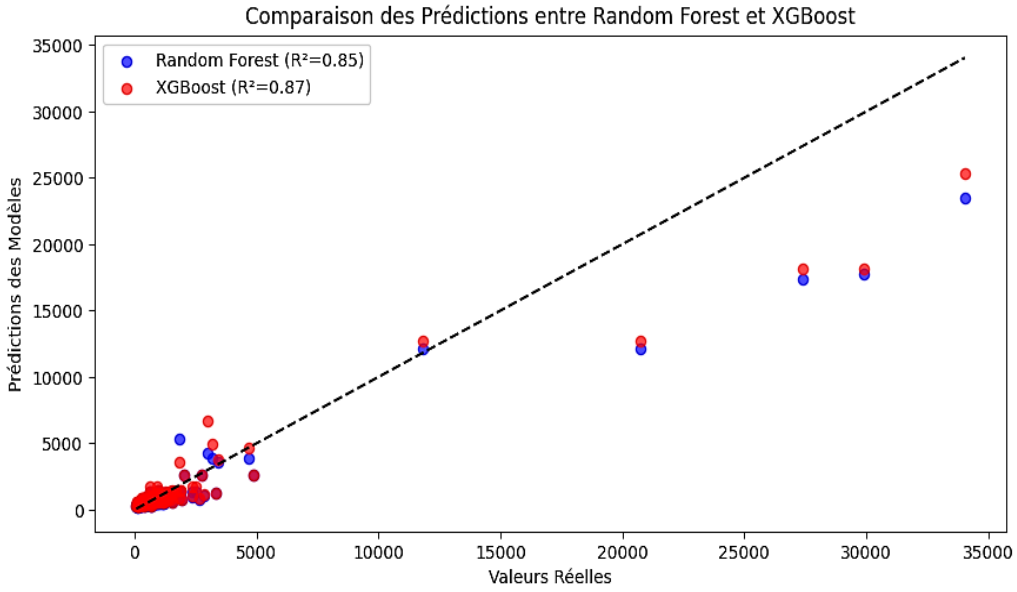

Figure 13 : Graphique de comparaison des prédictions vs valeurs réelles pour la densité de population à partir de l'indice VTLPI à l'échelle de la Vallée du Bandama

À cette échelle, XGBoost semble légèrement meilleur en termes de R^2. Cela indique qu'il est capable d'expliquer une plus grande partie de la variance des données spécifiques à ce district. Cependant, il est important de noter que le score de validation croisée est légèrement inférieur, suggérant une certaine variabilité dans les performances de XGBoost sur différents ensembles de données de validation.

V. Discussion

La modélisation basée sur le VTLPI a montré des résultats plus précis au niveau du district, notamment dans les grandes agglomérations, où la répartition de la population est plus variée. L'approche utilisant le Machine Learning a permis de corriger certaines limites des méthodes reposant uniquement sur le DNB, en réduisant l'effet de boom des lumières nocturnes (NTL) ainsi que les biais liés au niveau de développement économique et aux infrastructures lumineuses, comme les aéroports, qui faussent parfois l'évaluation de la densité de population.

Le principal atout du VTLPI réside dans l'intégration des données LST et la transformation en racine carrée des données lumineuses. Par rapport au DNB, les modèles de régression basés sur Random Forest et XGBoost minimisent la surestimation observée autour des aéroports et des espaces verts urbains. Toutefois, le VTLPI repose sur NDVI comme indicateur clé, en supposant ce dernier est généralement plus faible en zone urbaine. Cela peut entraîner une surestimation dans certaines régions arides, comme la savane, où le NDVI est souvent plus élevé que sur des sols nus. De plus, notre modèle basé sur le VTLPI tend à surestimer la densité de population dans certaines zones industrielles, comme à Bouaké, où les fortes températures, influencent les estimations.

Par ailleurs, l'intégration de la relation entre couverture terrestre et zones bâties a considérablement amélioré la précision des estimations. Une fois ces paramètres appliqués à l'échelle régionale, le coefficient de détermination R^2 du modèle VTLPI est passé de 0,74 à 0,85 pour Random Forest et de 0,69 à 0,87 pour XGBoost. La moyenne des scores de validation croisée a également progressé, atteignant 0,60 pour les deux modèles. Comme la majorité des populations sont concentrées en milieu urbain, les recherches futures pourraient exploiter des données issues de la détection sociale, comme les points d'intérêt, afin de modéliser la densité de population de manière dynamique. Cette approche permettrait d'affiner la distribution des populations en fonction des besoins, notamment pour évaluer les impacts des crises et catastrophes.

VI. Conclusion

L'objectif de ce travail de recherche était d'utiliser le potentiel des images satellites comme alternative aux données de recensements et d'enquêtes afin d'estimer la densité de population.
Afin de construire des modèles de prédictions robustes, une grande diversité de variables a été sélectionnée : NTL, LST, NDVI, indice indiquant la présence humaine par exemple.

La modélisation de la densité de population à donner de très bon résultat avec des coefficients de détermination allant de jusqu'à 0,87 pour les estimations à l'échelle de notre région d'étude avec le modèle XGBoost. Les résultats obtenus sont cohérents avec les résultats observés dans la littérature. En raison de limitations matérielles (temps et puissance de calcul), seulement deux modèles d'apprentissage automatique ont été élaborés.

Les résultats des modèles de prédictions présentés dans ce travail mériteraient d'être validés grâce à des données externes. En effet, il serait intéressant de comparer les résultats de cette étude avec des modèles qui estiment la densité de population issue d'autres sources de données comme par exemple les estimations de LandScan, WorldPop ou GPW.

L'analyse des données environnementales dans Google Earth Engine s'avère être un outil puissant pour comprendre et prédire la densité de la population.

Références bibliographiques

[1]. Wardrop, N. A., W. C. Jochem, T. J. Oiseau, H. R. Chamberlain, D. Clarke, D. Kerr, L. Bengtsson, S. Juran, V. Seaman, et A. J. Tatem. "Spatially disaggregated population estimates in the absence of national population and housing census data ". Proceedings of the National Academy of Sciences 115, no 14 (3 avril 2018): 3529 37. https://doi.org/10.1073/pnas.1715305115.

[2]. McBride, L.; Nichols, A. (2018). "Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning." World Bank Econ. Rev. 2018, 32, 531–550.

[3]. Hu, S. ; Ge, Y. ; Liu, M. ; Ren, Z. ; Zhang, X. (2022). « Village-level poverty identification using machine learning, high-resolution images, and geospatial data ». Int. J. Appl. Earth Obs. Geoinf. 2022, 107, 102694.

[4]. Luo Peng, Xianfeng Zhang, Junyi Cheng, and Quan Sun. 2019. "Modeling Population Density Using a New Index Derived from Multi-Sensor Image Data" Remote Sensing 11, no. 22: 2620. https://doi.org/10.3390/rs11222620

[5]. Puttanapong, Nattapong, Arturo Martinez, Jr., Joseph Albert Nino Bulan, Mildred Addawe, Ron Lester Durante et Marymell Martillan. 2022. "Predicting Poverty Using Geospatial Data in Thailand" ISPRS International Journal of Geo-Information 11, no. 5: 293. https://doi.org/10.3390/ijgi11050293

[6]. Recensement général de la population et de l'habitat, 2021 – Résultats globaux (octobre 2022) RESULTATS DEFINITIFSRP21.pdf (ins.ci)

[7]. « Gridded Population of the World (GPW), v4 | SEDAC ». Consulté le 7 mars 2023. https://sedac.ciesin.columbia.edu/data/collection/gpw-v4.

[8]. « High-resolution population estimates - GRID3 ». Consulté le 10 Mai 2023. https://grid3.org/solution/high-resolution-population-estimates.

[9]. « ORNL LandScan Viewer - Oak Ridge National Laboratory ». Consulté le 20 Avril 2023. https://landscan.ornl.gov/.

[10]. WorldPop. « Population Estimation for Sustainable Development ». Consulté le 15 Mai 2023. https://www.worldpop.org/methods/population-estimation-for-sustainable-development/.

[11]. « Mesure de la densité de population ». Consulté le 26 mai 2023. https://e-cours.univ-paris1.fr/modules/uved/envcal/html/milieux-anthropises/teledetection-territoires-urbain/2-2- mesuresdensitepopulation.html.

[12]. Géoconfluences. « Indicateurs ». Terme. École normale supérieure de Lyon, janvier 2013. ISSN : 2492-7775. http://geoconfluences.ens-lyon.fr/glossaire/indicateurs. Consulté le 15 Mai 2023