INSD

THEMES DE RECHERCHE

Estimation en petits domaines de l'accès à l'eau potable

Application au Burkina Faso

Présenté par :

KINDO P.
NATHAN

Dirigé par :

Mr. ZONGO

Août 2025

Résumé

Ce travail traite de l'estimation en petits domaines de l'accès à l'eau potable au Burkina Faso, en lien avec l'Objectif de Développement Durable 6.1. Confrontées à des tailles d'échantillon insuffisantes dans les enquêtes ménages pour produire des estimations locales précises, les autorités ont besoin de méthodes statistiques capables de combiner l'information d'enquête avec des covariables géospatiales exhaustives.

Nous proposons une approche fondée sur des modèles hiérarchiques bayésiens spatiaux (logit-binomial avec composantes CAR) intégrant des données satellites et administratives. Le travail présente la méthodologie, l'implémentation et une évaluation des performances par rapport aux estimateurs directs.

Mots-clés : estimation en petits domaines, accès à l'eau potable, modèles bayésiens hiérarchiques, données géospatiales, Burkina Faso

Table des matières

Table des figures

Liste des tableaux

Introduction générale

L'accès à l'eau potable constitue un enjeu fondamental du développement humain et durable. L'Objectif de Développement Durable 6 vise à garantir l'accès universel et équitable à une eau potable sûre et abordable d'ici 2030. Au Burkina Faso, pays sahélien confronté à des pressions climatiques et économiques, les inégalités territoriales d'accès à l'eau restent marquées. Les estimations nationales issues des enquêtes DHS/MICS masquent des disparités locales importantes.

Ce travail a pour objectif de développer et d'appliquer une méthodologie d'estimation en petits domaines (SAE) pour produire des cartes désagrégées et fiables de l'accès à l'eau potable au Burkina Faso. L'approche combine les microdonnées d'enquête avec des covariables géospatiales (WorldPop, nightlights, NDVI, CHIRPS) dans un cadre bayésien hiérarchique incluant une composante spatiale.

Problématique

Comment produire des estimations fiables de l'accès à l'eau potable au niveau communal ou départemental au Burkina Faso, alors que les enquêtes nationales ne fournissent pas d'échantillons suffisants à ces échelles ?

Objectifs

— **Objectif général :** Estimer l'accès à l'eau potable dans les petits domaines du Burkina Faso de manière précise et cartographiable.

— **Objectifs spécifiques :**

1. Produire les estimations directes à partir des microdonnées DHS ;
2. Intégrer des covariables satellites/administratives pour étayer les modèles SAE ;
3. Implémenter un modèle bayésien hiérarchique spatial (logit-binomial + CAR) ;
4. Valider et comparer les estimations obtenues aux estimateurs directs et proposer des recommandations de suivi.

Hypothèses de recherche

H1 : L'utilisation de covariables géospatiales améliore significativement la précision des estimations en petits domaines.

H2 : L'introduction d'une composante spatiale (CAR) réduit l'incertitude dans les domaines faiblement échantillonnés.

H3 : Un cadre bayésien hiérarchique permet une quantification robuste de l'incertitude pour le suivi des ODD.

Méthodologie générale

L'approche combine estimation directe, modèles Fay–Herriot (area-level), modèles unit-level (BHF) et un modèle bayésien spatial logit-binomial. L'implémentation repose sur des outils statistiques (Python, PyMC, GeoPandas), des shapefiles administratifs, et des rasters satellites agrégés aux domaines.

Structure du travail

Le travail est organisé comme suit :

Chapitre 1 : Revue de littérature et cadre théorique.
Chapitre 2 : Présentation du cadre d'étude (Burkina Faso, données).
Chapitre 3 : Méthodologie (spécification du modèle choisi).
Chapitre 4 : Résultats et analyses.
Chapitre 5 : Discussion, recommandations et conclusion.

Chapitre 1

Revue de la littérature et cadre théorique

1.1 Définitions des concepts clés

1.1.1 Accès à l'eau potable

Selon le Joint Monitoring Programme (JMP) de l'OMS et de l'UNICEF, l'indicateur ODD 6.1.1 mesure la proportion de la population utilisant des services d'eau potable gérés en toute sécurité, c.-à-d. une source améliorée, disponible sur place et exempte de contamination.

1.1.2 Estimation en petits domaines (SAE)

La *Small Area Estimation* regroupe les méthodes statistiques destinées à produire des estimations précises pour des sous-populations (petits domaines) ne disposant pas d'un échantillon suffisant pour une estimation directe fiable.

1.2 Théories et modèles existants

Dans cette section nous détaillons les principales approches permettant d'obtenir des estimations dans les petits domaines.

1.2.1 Estimation directe

L'estimation directe est la méthode la plus simple qui consiste à utiliser uniquement les données de l'échantillon du domaine d'intérêt. Pour chaque domaine i, on calcule la proportion de ménages ayant accès à l'eau potable comme le ratio entre le nombre de ménages avec accès et le nombre total de ménages enquêtés dans le domaine.

Limitation : Lorsque le nombre d'observations n_i dans un domaine est petit, la variance de l'estimateur direct est grande et les intervalles de confiance deviennent peu informatifs, rendant cette approche inadaptée pour les petits domaines.

1.2.2 Modèles area-level : Fay–Herriot

Le modèle de Fay-Herriot est une approche area-level qui combine information d'enquête et données auxiliaires. Il fonctionne en deux étapes :

1. Au premier niveau, on modélise l'estimateur direct comme la somme de la vraie valeur et d'une erreur d'échantillonnage.
2. Au deuxième niveau, on modélise la vraie valeur comme une combinaison linéaire de covariables avec un terme aléatoire.

L'idée clé est d'utiliser un facteur de shrinkage γ_i qui pondère le compromis entre l'estimateur direct (précis mais variable) et la prédiction basée sur les covariables (moins précise mais stable). Ce facteur dépend du rapport entre la variance des effets aléatoires et la variance totale.

1.2.3 Modèles unit-level : Battese–Harter–Fuller (BHF)

Contrairement aux modèles area-level qui travaillent avec des agrégats, les modèles unit-level utilisent les données individuelles. Le modèle BHF spécifie une relation entre la variable réponse et des covariables au niveau individuel, avec un effet aléatoire spécifique au domaine.

L'avantage de cette approche est qu'elle permet de capturer à la fois la variabilité entre domaines et la variabilité intra-domaines. Elle nécessite cependant des données auxiliaires au niveau individuel, ce qui n'est pas toujours disponible.

1.2.4 Approches bayésiennes hiérarchiques

Les approches bayésiennes hiérarchiques offrent un cadre flexible pour l'estimation en petits domaines. Elles permettent d'incorporer facilement des informations préalables, de modéliser des structures complexes (comme les effets spatiaux) et de fournir une quantification naturelle de l'incertitude through les distributions a posteriori.

Dans le contexte binomial qui nous intéresse, on modélise le nombre de ménages avec accès à l'eau comme une variable binomiale, et on utilise une transformation logit pour lier la probabilité d'accès à une combinaison linéaire de covariables et d'effets aléatoires.

1.2.5 Extensions spatiales : modèles CAR/BYM

Les modèles spatiaux exploitent le fait que des domaines géographiquement proches ont tendance à avoir des caractéristiques similaires. Le modèle CAR (Conditional Autoregressive) suppose que la valeur dans un domaine donné dépend des valeurs dans les domaines voisins.

Le modèle BYM (Besag-York-Mollié) combine deux types d'effets aléatoires : un effet spatial structuré (CAR) qui capture la dépendance spatiale, et un effet non structuré qui capture l'hétérogénéité résiduelle non expliquée par la structure spatiale.

1.3 Travaux antérieurs (synthèse)

Les contributions majeures dans la littérature sont :

— **fay1979** : fondements du modèle area-level.

— **battese1988** : modèle unit-level et intégration de données satellites.

— **elbers2003** : micro-cartographie de la pauvrenté en combinant enquêtes et recensement.

— **rao2015** : manuel de référence sur la SAE.

— **rue2009** : INLA pour l'approximation bayésienne rapide des modèles latents gaussiens.

— Rapports JMP (**who2017**) : définition et suivi des indicateurs d'eau liés aux ODD.

1.4 Positionnement de ce travail

Ce travail applique et combine les approches ci-dessus dans un cadre opérationnel pour le Burkina Faso :

— utilisation des microdonnées DHS pour les estimations directes ;

— intégration de covariables géospatiales (WorldPop, VIIRS, NDVI, CHIRPS, inventaires de points d'eau) ;

— spécification d'un modèle bayésien hiérarchique logit-binomial avec composantes spatiales (CAR) et non-spatiales ;

— implémentation en Python pour une solution reproductible et accessible ;

— validation par comparaisons aux estimateurs directs et diagnostics postérieurs.

Chapitre 2

Présentation du cadre d'étude (contexte empirique)

2.1 Description du milieu : le Burkina Faso

Le Burkina Faso est un pays d'Afrique de l'Ouest, enclavé, d'une superficie d'environ 273 000 km^2 et une population estimée à 22 millions (2022). Il est subdivisé administrativement en régions, provinces et communes, ce qui permet d'effectuer une estimation à des niveaux variés (ADM1/ADM2/ADM3). Le pays présente de fortes variations climatiques (Sahel au nord, zones soudaniennes au sud), influençant la disponibilité des ressources hydriques.

2.2 Justification du choix du terrain/données

Le Burkina Faso est un cas pertinent car :

— l'accès à l'eau est une priorité nationale et internationale ;

— des enquêtes DHS/MICS récentes sont disponibles ;

— l'utilisation de données satellites et d'inventaires water point est possible et pertinente ;

— la décision publique nécessite des estimations locales pour cibler les interventions.

2.3 Problèmes ou enjeux spécifiques observés

Les enjeux locaux incluent :

— inégalités urbain/rural prononcées ;

— zones à faible échantillonnage dans les enquêtes (communes peu peuplées) ;

— déplacement de populations lié à la sécurité pouvant changer rapidement la demande en eau ;

— coût élevé des enquêtes nationales régulières.

2.4 Choix méthodologique récapitulatif

Au regard des caractéristiques du pays et des objectifs d'estimation sous-nationale, nous retenons la spécification suivante pour la phase d'estimation principale :

$$Y_i \sim \text{Binomial}(n_i, p_i), \tag{2.1}$$

$$\text{logit}(p_i) = \mathbf{x}_i^\top \beta + u_i + v_i, \tag{2.2}$$

avec u_i un terme non-spatial (iid) et v_i un terme spatial CAR. Ce choix permet d'intégrer des covariables exhaustives, de modéliser la structure spatiale et de rendre compte de l'incertitude via une approche bayésienne.

Chapitre 3

Méthodologie

3.1 Sources de données et techniques de collecte

L'analyse repose sur l'intégration de plusieurs sources complémentaires :

DHS microdonnées Household Recode + GPS clusters.

Points d'eau données OSM/WPDx.

Population recensements communaux.

Précipitations bases CHIRPS/WorldClim.

Limites administratives communes (ADM3).

3.2 Construction des variables

(a) **Variable dépendante** : proportion de ménages ayant accès à une source d'eau potable.

(b) **Variables explicatives** : densité de population, densité de points d'eau, précipitations, type de milieu.

(c) **Pré-traitement** : harmonisation CRS, jointure spatiale, standardisation.

3.3 Spécification du modèle bayésien hiérarchique spatial

Nous utilisons la formulation suivante :

3.3.1 Niveau observation (likelihood)

Au niveau observationnel, nous modélisons le nombre de ménages avec accès à l'eau dans chaque domaine comme une variable binomiale :

$$Y_i \mid p_i \sim \text{Binomial}(n_i, p_i), \quad i = 1, \ldots, D. \tag{3.1}$$

Cette modélisation est naturelle car nous comptons le nombre de "succès" (ménages avec accès) sur un certain nombre d'"essais" (ménages enquêtés).

3.3.2 Lien et structure latente

Nous utilisons une transformation logit pour lier la probabilité p_i à une combinaison linéaire de covariables et d'effets aléatoires :

$$\text{logit}(p_i) = \eta_i = \mathbf{x}_i^\top \beta + u_i + v_i, \tag{3.2}$$

où

β est le vecteur des effets fixes des covariables,

u_i est un effet aléatoire non structuré qui capture l'hétérogénéité résiduelle entre domaines,

v_i est un effet aléatoire spatial structuré qui capture la dépendance spatiale entre domaines voisins.

3.3.3 Composante spatiale CAR

La composante spatiale suit un modèle Conditional Autoregressive (CAR) qui spécifie que la valeur dans un domaine donné dépend des valeurs dans les domaines voisins. Formellement, la distribution conditionnelle de v_i sachant les valeurs dans les autres domaines est :

$$v_i \mid \boldsymbol{v}_{-i} \sim \mathcal{N}\left(\frac{\sum_{j\sim i} v_j}{m_i}, \frac{1}{\tau_v m_i}\right), \tag{3.3}$$

où $j \sim i$ indique que les domaines j et i sont voisins, m_i est le nombre de voisins du domaine i, et τ_v est un paramètre de précision.

3.3.4 Priors

Nous utilisons des priors faiblement informatifs pour éviter la surinterprétation des données :

$$\beta_k \sim \mathcal{N}(0, 10^2), \quad \text{pour } k = 1, \ldots, p \tag{3.4}$$

$$\sigma_u \sim \text{Half-StudentT}(4, 0, 1) \tag{3.5}$$

$$\tau_v \sim \text{Gamma}(1, 0.01) \tag{3.6}$$

Le prior Half-StudentT sur σ_u permet une régularisation tout en étant suffisamment faiblement informatif. Le prior Gamma sur τ_v est un choix standard dans la littérature pour les modèles CAR.

3.4 Estimation numérique avec Python

3.4.1 Choix de Python

Python a été choisi comme langage d'implémentation pour plusieurs raisons :

— Large écosystème de bibliothèques scientifiques (NumPy, Pandas, SciPy)
— Bibliothèques spécialisées pour les modèles bayésiens (PyMC, PyStan)
— Outils puissants pour le géospatial (GeoPandas, Rasterio)
— Reproducibilité et accessibilité accrues
— Communauté active et support

3.4.2 Bibliothèques utilisées

PyMC pour l'inférence bayésienne avec MCMC

GeoPandas pour le traitement des données géospatiales

Rasterio pour la manipulation des données raster

NumPy et Pandas pour le traitement des données numériques

Matplotlib et Seaborn pour la visualisation

3.4.3 Implémentation MCMC avec PyMC

PyMC permet d'implémenter des modèles bayésiens complexes avec une syntaxe intuitive. L'inférence est réalisée via des algorithmes MCMC (NUTS, Metropolis-Hastings) qui échantillonnent à partir de la distribution a posteriori.

3.5 Pré-traitement et construction des covariables

(a) **Harmonisation des systèmes de référence** : Toutes les données géospatiales sont projetées dans le même système de coordonnées (WGS84).

(b) **Agrégation des rasters** : Les données raster (WorldPop, VIIRS, NDVI, CHIRPS) sont agrégées aux domaines d'intérêt en calculant des statistiques zonales (moyenne, somme).

(c) **Calcul des distances** : La distance moyenne au point d'eau le plus proche est calculée à partir des inventaires de points d'eau.

(d) **Standardisation** : Toutes les covariables sont centrées et réduites pour améliorer la convergence des algorithmes d'estimation.

(e) **Traitement des valeurs extrêmes** : Les proportions 0 et 1 sont ajustées avec une correction de type pseudocount pour éviter les problèmes numériques.

3.6 Validation et diagnostics

— **Posterior predictive checks** : Nous comparons les données observées à des données simulées à partir du modèle pour vérifier l'adéquation du modèle.

— **Comparaison des erreurs** : Nous calculons le RMSE (Root Mean Square Error) et le MAPE (Mean Absolute Percentage Error) pour comparer les performances des différents modèles.

— **Critères d'information** : Nous utilisons le WAIC (Watanabe-Akaike Information Criterion) pour comparer différents modèles.

— **Cartographie de l'incertitude** : Nous produisons des cartes des intervalles de crédibilité pour visualiser l'incertitude des estimations.

— **Diagnostics MCMC** : Nous vérifions la convergence des chaînes MCMC avec les statistiques de Gelman-Rubin et les tracés de traces.

3.7 Limites de la méthodologie

Principales limitations

— Les coordonnées GPS DHS sont perturbées (jittering) pour protéger la confidentialité, ce qui introduit une incertitude supplémentaire dans l'analyse spatiale.

— Les différences temporelles entre les sources de données (enquêtes vs données satellites) peuvent créer un désalignement.

— La complexité computationnelle peut être élevée pour les modèles spatiaux avec beaucoup de domaines.

— La spécification du voisinage spatial dans le modèle CAR peut influencer les résultats.

— Les méthodes MCMC peuvent être lentes pour de grands jeux de données.

Chapitre 4

Analyse et Résultats

4.1 Exploration initiale

Une première étape a consisté à cartographier la proportion de ménages ayant accès à l'eau potable (`prop_potable`) issue des données DHS, agrégée au niveau communal.

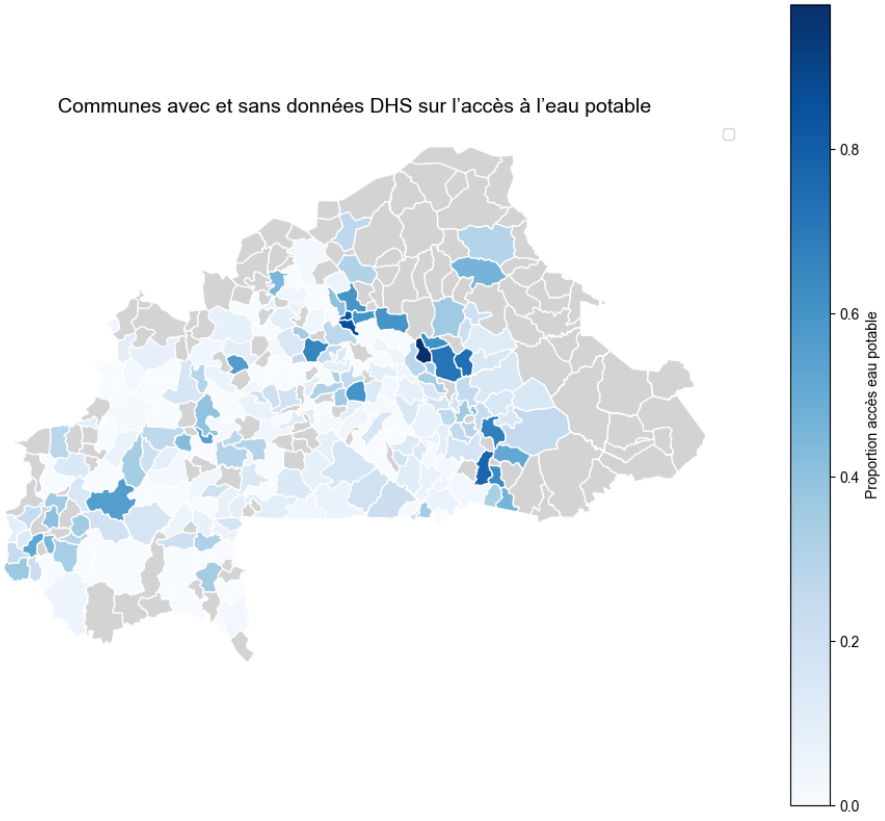

FIGURE 4.1 – Proportion de ménages ayant accès à l'eau potable par commune (estimations directes DHS).

Comme l'illustre la carte, seules certaines communes disposent d'une valeur observée de la proportion de ménages ayant accès à l'eau potable. Une grande partie du

territoire apparaît en gris, indiquant l'absence de données directes dans ces zones. Cette lacune s'explique par le plan de sondage des DHS, conçu pour fournir des estimations représentatives à l'échelle nationale ou régionale, mais non à l'échelle communale.

Cette observation justifie le recours à des méthodes d'estimation en petits domaines, permettant d'extrapoler l'indicateur d'accès à l'eau potable en mobilisant des covariables auxiliaires (densité de population, points d'eau, précipitations, etc.) et en exploitant la corrélation spatiale entre communes voisines.

4.2 Resultat

Concernant les resultats, vu la faible capacité du PC, les simulations n'ont pas abouti. Neanmoins, le notebook du code a été proposé dans le dossier 'données et code python'