Méthodologie

Dans le cadre de cette étude, nous cherchons à modéliser le nombre de consultations médicales en prenant en compte les caractéristiques socio-économiques et spatiales des communes. La littérature suggère que les modèles de régression de Poisson et ses variantes avec effets aléatoires sont particulièrement adaptés aux données de comptage, tout en tenant compte de l'hétérogénéité locale et des dépendances spatiales. Ainsi, nous proposons une approche combinant les modèles de régression de Poisson standard et les modèles mixtes généralisés pour mieux comprendre les dynamiques d'accès aux soins.

1. Modélisation du nombre de visites de médécins généralistes

1.1. Modèle de régression de Poisson

La régression de Poisson est un type de modèle linéaire généralisé (GLM) utilisé pour modéliser des données de comptage. La distribution de Poisson suppose que la variance est égale à la moyenne, ce qui est souvent approprié pour des données de comptage. Le modèle de base est :

$$\log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} = (X\beta)_i$$

où μ_i est le nombre attendu de cas pour l'observation i, X_{ij} sont les variables explicatives, et β_j sont les coefficients à estimer. Toutefois, ce modèle suppose une indépendance entre observations, ce qui peut ne pas être réaliste en présence de corrélation spatiale ou d'hétérogénéité non observée.

1.2. Modèles de Régression de Poisson avec effets aléatoires

Afin de mieux capter l'hétérogénéité entre les communes, nous introduisons des modèles avec effets aléatoires. Ces modèles permettent de prendre en compte des facteurs non observés qui influencent le nombre de consultations.

Un effet aléatoire v_i capture l'effet d'hétérogénéité non mesurable qui influence les résultats dans chaque agglomération. Il peut être considéré comme une variation spécifique à l'agglomération qui n'est pas expliquée par les covariables observées dans le modèle (comme les facteurs socio-économiques).

1.2.1. Régression de Poisson avec effets aléatoires non spatiaux La formulation générale de ce modèle peut être écrite comme suit :

$$\log(\mu_i) = \log(E_i) + (X\beta)_i + u_i$$

où :

- μ_i est la moyenne du nombre de visites pour l'agglomération i,
- E_i est la taille de la population de l'agglomération i (terme d'ajustement de l'exposition),
- X est la matrice de conception pour les facteurs socio-économiques,
- β est le vecteur des coefficients à estimer,
- u_i est l'effet aléatoire spécifique à l'agglomération i, supposé indépendant et suivant une distribution normale.

1.2.2. Régression Poisson avec effets aléatoires spatiaux

Deux approches sont utilisées pour modéliser la structure spatiale des données :

- Effets aléatoires basés sur la distance : ajout d'un effet spatial v_i capturant la corrélation entre communes distantes.
- Effets aléatoires basés sur le voisinage : utilisation d'un terme w_i représentant les influences des communes voisines à travers un modèle autorégressif conditionnel (CAR).

Ce modèle prend en compte l'autocorrélation spatiale par l'introduction d'effets aléatoires spatiaux :

$$\log(\mu_i) = \log(E_i) + (X\beta)_i + u_i + w_i$$

où w_i suit une distribution conditionnelle dépendant des voisins.

Le modèle CAR peut être spécifié comme suit :

$$w_i \mid w_j, j \neq i \sim N\left(\frac{1}{n_i}\sum_{j\in\delta_i} w_j, \sigma^2\right)$$

où :

- n_i est le nombre de voisins pour l'agglomération i,
- δ_i est l'ensemble des voisins de l'agglomération i,
- σ^2 est la variance de l'effet aléatoire.

Ce modèle suppose que la valeur de w_i dépend en moyenne des valeurs observées dans les zones voisines, ce qui permet de modéliser la corrélation spatiale.

2. Estimation des paramètres

2.1. Structure de variance des effets aléatoires

Dans le cadre d'un modèle mixte, les effets aléatoires v_i et w_i sont supposés suivre une distribution normale centrée avec une variance σ_v^2 et σ_w^2 respectivement.

2.2. Méthodes d'estimation

Deux principales méthodes sont utilisées pour estimer les paramètres des modèles

Approche de Maximum de Vraisemblance (ML)

Une fois que le modèle est ajusté, les effets aléatoires sont souvent estimés comme partie intégrante du processus d'optimisation de la vraisemblance. Cela se fait généralement en maximisant la vraisemblance marginale des données en tenant compte des effets aléatoires.

Après avoir obtenu les estimations des paramètres du modèle, on peut dériver les estimations de v_i à partir des résidus du modèle.

Approche Empirique

Après l'ajustement du modèle, les valeurs de v_i peuvent être calculées à partir des résidus du modèle pour chaque agglomération. Cela implique de comparer les cas observés dans chaque agglomération à ceux prédits par le modèle, en tenant compte de l'effet des covariables.

2.3. Étapes d'Estimation

Pour estimer les paramètres de ces modèles (y compris v_i et w_i), une approche courante consiste à utiliser des méthodes spécifiques dans le cadre de la **maximum de vraisemblance** ou des **approches bayésiennes**.

1. **Pseudo-vraisemblance**
   - On utilise la vraisemblance conditionnelle de Y étant donné les effets aléatoires, pour améliorer l'estimation des paramètres.
   - Cette approche permet de simplifier l'estimation en contournant certaines intégrations complexes dans les modèles à effets aléatoires.
2. **Optimisation**
   - Les algorithmes d'optimisation, tels que **Newton-Raphson**, sont utilisés pour maximiser la vraisemblance.
   - Pour des approches bayésiennes, des méthodes de simulation comme **MCMC (Markov Chain Monte Carlo)** sont couramment appliquées pour estimer v_i et w_i.

Ce modèle est similaire au précédent, mais v_i est remplacé par un effet aléatoire basé sur le voisinage :

$$\log(\mu_i) = \log(E_i) + (X\beta)_i + u_i + w_i$$

où :

- w_i représente l'effet aléatoire spatial basé sur le voisinage (par exemple, un **modèle autoregressif conditionnel (CAR)** qui capture l'influence des agglomérations voisines).

3. Modèle de régression logistique

Elle modélise une réponse binaire ($y \sim B(n,p)$), où p est la probabilité de succès :

$$P(y|n,p) = \binom{n}{y} p^y (1-p)^{n-y}$$

La probabilité p est reliée au prédicteur par la fonction logistique :

$$p = \frac{1}{1+e^{-\eta}} \quad \text{où} \quad \eta = \beta_0 + \sum_{i=1}^{m} \beta_i x_i.$$

La log-vraisemblance est exprimée comme :

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log p_i + (1-y_i)\log(1-p_i) \right].$$

Note sur l'Estimation

Pour tous ces modèles, l'estimation des paramètres est réalisée à l'aide de la **méthode de pseudo-vraisemblance**, qui prend en compte les structures autocorrélées.

De plus, pour la comparaison des modèles, on utilise les critères suivants :

- **AIC (Akaike Information Criterion)** : évalue la qualité du modèle en pénalisant les modèles trop complexes.
- **BIC (Bayesian Information Criterion)** : similaire à l'AIC, mais avec une pénalisation plus forte sur la complexité du modèle.

Ces critères permettent de choisir le modèle le plus adapté en fonction des données disponibles.

Cette méthodologie propose une modélisation rigoureuse du nombre de consultations en combinant régression de Poisson et modèles mixtes. L'ajout d'effets aléatoires et de structures spatiales permet d'améliorer la précision des estimations et d'obtenir une meilleure compréhension des disparités géographiques dans l'accès aux soins. La comparaison des modèles permettra d'identifier la structure la plus adaptée aux données étudiées.