Amélioration de la précision d'un estimateur par la prise en compte l'hétérogénéité spatiale

Jean-Michel Floch et Eric Lesage

Département de l'action régionale
INSEE

18 mai 2017 / Colloque Théoquant

Plan

Introduction

Dans le cadre de l'estimation sur des domaines géographiques, et a fortiori sur des petits domaines, à partir de données d'enquêtes, il est maintenant habituel de recourir à une approche basée sur un modèle et d'utiliser des estimateurs EBP (Empirical Best Predictors) ou BLUP (Best Linear Unbiased Predictors) (voir par exemple Chambers et Clark (2012)).

Ainsi, les valeurs des unités non-échantillonnées sont remplacées par des valeurs prédites à partir d'un modèle dont les paramètres sont estimés au moyen des valeurs des unités échantillonnées.

Dans cette étude, on emploie un modèle de régression géographique pondérée (RGP) qui est une variante du modèle classique de régression linéaire qui permet d'utiliser plus efficacement l'information spatiale et offre un meilleur ajustement local des données.

La régression géographique pondérée

Lorsqu'on écrit un modèle linéaire sur une population U, on fait l'hypothèse que le phénomène modélisé est homogène spatialement. Si on présume que les mécanismes étudiés présentent une certaine hétérogénéité spatiale, on peut chercher à calibrer plusieurs modèles de régression linéaires à partir de sous-populations présentant une cohérence géographique.

La geographically weighted regression (GWR) que l'on pourrait traduire par régression géographiquement pondérée (RGP) est une des méthodes les plus utilisées pour prendre en compte l'information spatiale.

Elle présente l'avantage de ne pas nécessiter un partitionnement géographique fixe de la population initiale. Elle a été proposée par les géographes Brunsdon, Fotheringham et Charlton (voir par exemple Fotheringham et al (2003)).

Considérons le modèle de régression linéaire suivant pour la population U :

$$y_i = \beta^\top \mathbf{x}_i + \varepsilon_i, \tag{1}$$

où

- y est la variable d'intérêt et x le vecteur des covariables,
- β est un vecteur de paramètres
- et les ε_i sont i.i.d.

Définition : Le principe de la RGP

on estime localement le modèle spécifié sur le territoire d'étude. L'estimation locale des coefficients se fait en utilisant une version pondérée des moindres carrés.

Selon la pondération utilisée, tout ou partie des observations recueillies rentrent dans l'estimation locale.

Définition : Moindres carrés géographiquement pondérés

Pour chaque point s de l'espace géographique étudié, on obtient l'estimateur pondéré suivant :

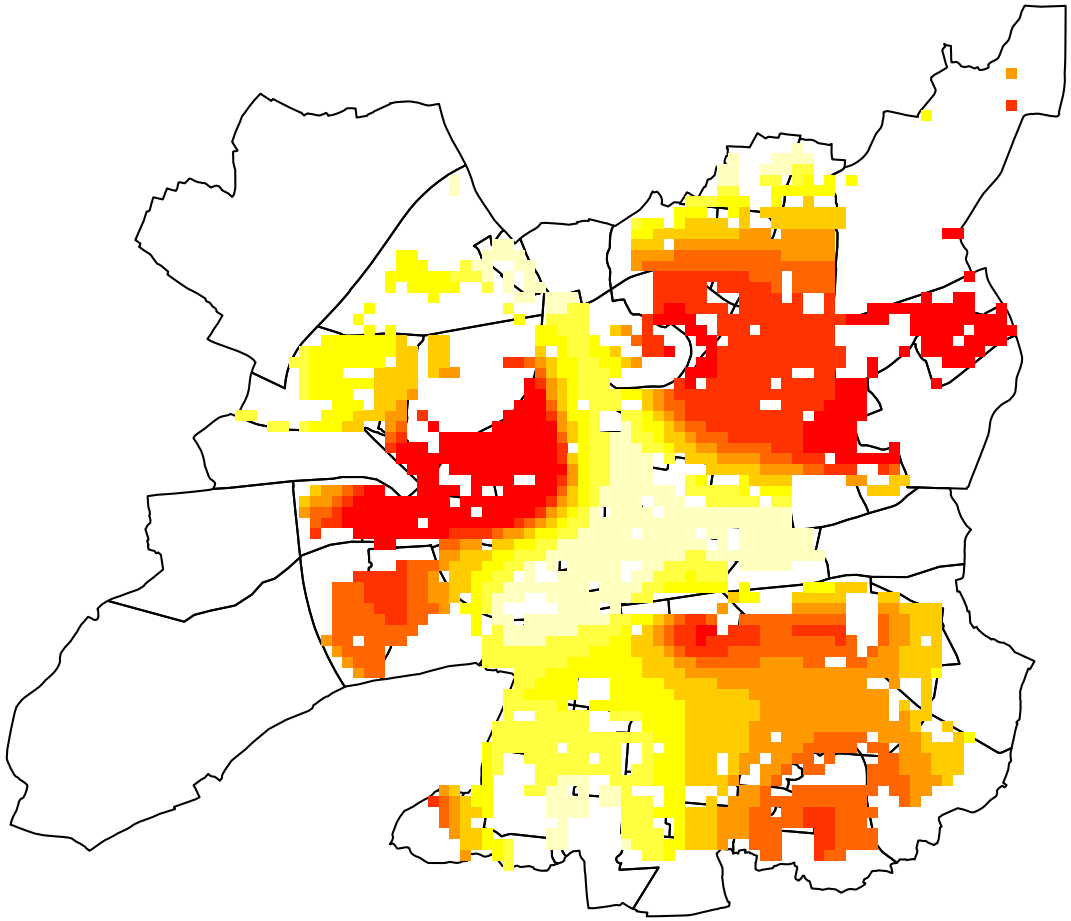
$$\hat{\beta}_s = \left(\sum_{i \in U} w_i(s) \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left(\sum_{i \in U} w_i(s) \mathbf{x}_i y_i \right),$$

où $w_i(s)$ est le poids de l'individu i au point d'estimation s.

Les raisons qui conduisent à estimer l'ensemble des paramètres par les moindres carrés pondérés sur un sous-ensemble de points situés dans le voisinage du point d'estimation sont discutées dans Fotheringham et al (2003).

Les poids $w_i(s)$ utilisés sont construits à partir de fonctions décroissantes de la distance entre le point d'estimation s et les individus i, comme la fonction gaussienne et la fonction biweight.

Pente des régressions géograpiques pondérées



Etude par simulations

On réalise une étude par simulations à partir de données réelles dans laquelle on compare les propriétés Monte Carlo de trois estimateurs.

Les données :

- données au carreau de la CNAM et de la DGFIP pour l'année 2009,
- carreaux de 100 mètres de côté.
- y : nombre de personnes à bas revenu (revenu fiscal par unité de consommation inférieur à 60% du revenu médian des habitants de Rennes)
- x : nombre de personnes bénéficiant de la CMUC (couverture médicale universelle complémentaire).

On cherche à estimer le nombre total des personnes à bas revenus dans les Iris de la ville de Rennes. Le nombre de personnes bénéficiant de la CMUC sera notre variable auxiliaire, connue pour tous les carreaux de la ville de Rennes.

On reconstitue un fichier d'étude correspondant aux carreaux de la ville de Rennes, U, qui contient :

- les coordonnées des carreaux,
- le nombre de personnes sous le seuil de pauvreté,
- le nombre de personnes bénéficiant de la CMUC,
- l'Iris auquel appartient le carreau.

On élimine du fichier initial les 8 plus petits Iris, i.e; ceux qui comportent moins de 40 personnes à bas revenu.

On a finalement $N = 2141$ carreaux et 84 Iris.

Plan de sondage

On sélectionne un échantillon s de carreaux :

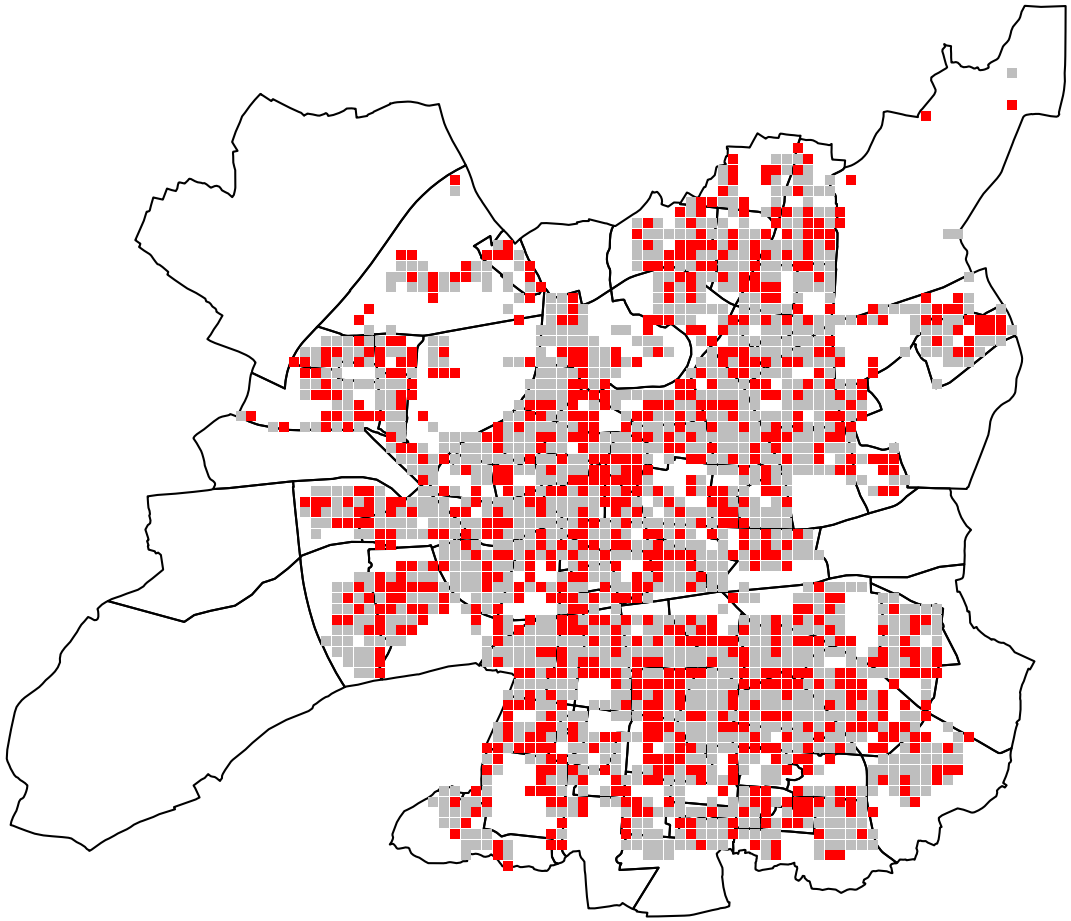
- tirage aléatoire simple sans remise
- taille $n = 856$
- faction de sondage $n/N = 40\%$

On note r le complémentaire de s dans U, i.e. le sous-ensemble de carreaux qui ne sont pas échantillonnés.

- y_i connus pour $i \in s$
- x_i connus pour tous les carreaux $i \in U$

L'estimation du nombre de personnes à bas revenus par Iris est un problème classique d'estimation sur domaine.

Rennes découpée en Iris



Model based estimation

- Il existe une relation linéaire forte entre le nombre de personnes à bas revenus et le nombre de bénéficiaires de la CMUC par carreau.
- On adopte dans cet étude une approche dite "basée sur le modèle" ;
  on prédit les valeurs y_i des carreaux non échantillonnés grâce à :
  1. un modèle estimé avec les données de l'échantillon
  2. une information auxiliaire x disponible pour les carreaux non-échantillonnés.
- On emploie deux types de modèles :
  (1) un modèle classique de régression linéaire
  (2) un modèle de régression géographique pondérée

Modèle de régression linéaire : $\hat{\beta}$

- On estime d'abord le vecteur des coefficients d'un modèle de régression linéaire pour les carreaux i de s :

$$y_i = \beta^\top \mathbf{x}_i + \varepsilon_i, \tag{2}$$

  - ε_i sont i.i.d.
  - $\beta^\top = (\beta_1, \beta_2)$.

- $\hat{\beta}$: estimateur des moindres carrés ordinaires

$$\widehat{\beta} = \left(\sum_{i \in s} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left(\sum_{i \in s} \mathbf{x}_i y_i \right).$$

- L'ajustement est très bon et le $R^2 = 90\%$.

Régression géographique pondérée : $\widehat{\beta}_l$, $l \in r$

On obtient l'estimateur pondéré suivant pour chaque carreau $l \in r$:

$$\widehat{\beta}_l = \left(\sum_{i \in s} w_i(l) \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left(\sum_{i \in s} w_i(l) \mathbf{x}_i y_i \right).$$

- $w_i(l)$: poids du carreau i pour l'estimation au point l

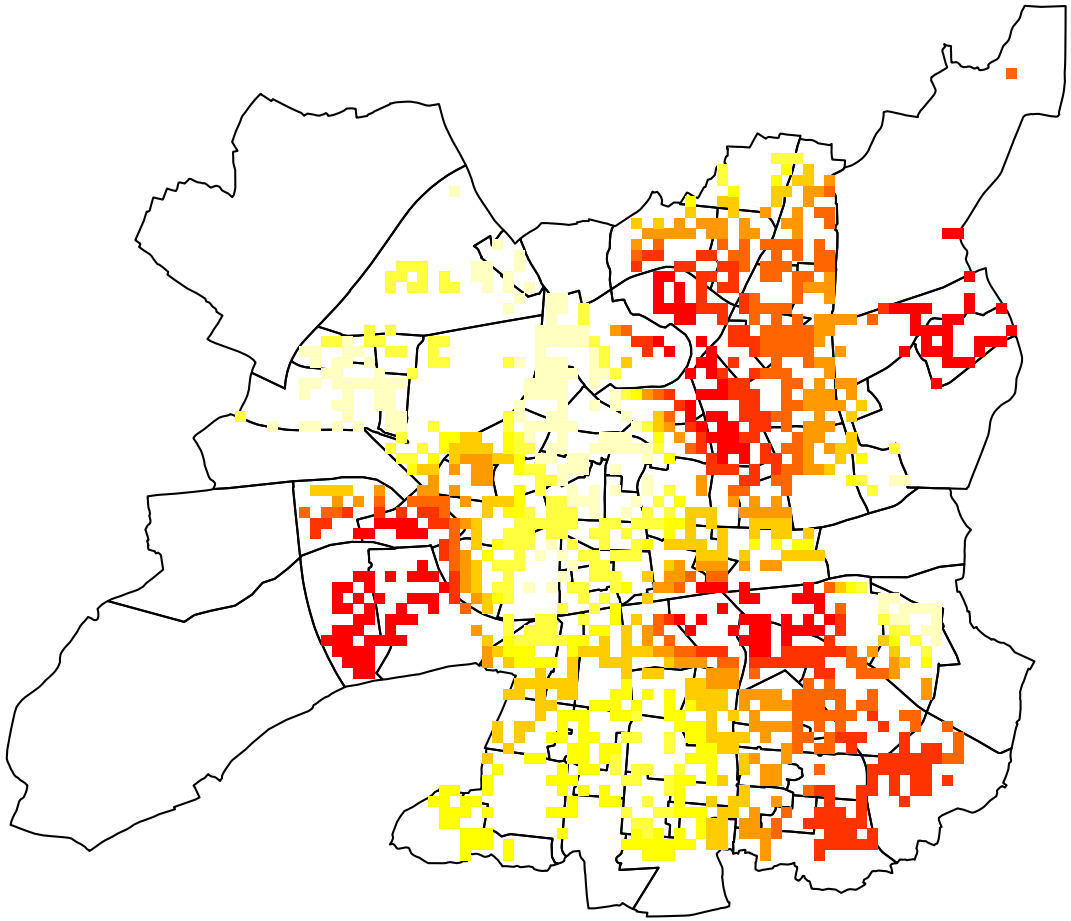
$$w_i(l) = \exp\left(-\frac{1}{2} \left[\frac{d(i,l)}{h} \right]^2 \right),$$

  - $d(i,l)$ est la distance euclidienne entre les carreaux i et l,
  - h la largeur d'une fenêtre paramétrable fixée à 500 mètres,
  - les poids des carreaux $i \in s$ décroissent en fonction de l'éloignement entre les carreaux i et l,
  - fonction gaussienne.

Régression géographique pondérée

- On utilise la fonction *gwr( )* du package spgwr du logiciel R pour réaliser nos calculs.
- Les ordonnées à l'origine $\hat{\beta}_{j,1}$ varient peu d'un carreau à l'autre.
- Les pentes $\hat{\beta}_{j,2}$ varient sensiblement de 1.6 à 3.3.
- La carte de la Figure (3) illustre les variations spatiales des pentes.

Pente des régressions géograpiques pondérées



Paramètre d'intérêt : $t_y(j)$

On note :

- U_j : ensemble des carreaux de l'Iris j, $j = 1, \ldots, 84$,
- s_j : sous-ensemble des carreaux échantillonnés qui appartiennent à l'Iris j
- r_j : sous-ensemble des carreaux non échantillonnés de l'Iris j,
- $t_y(j)$: nombre de personnes à bas revenus de l'Iris j.

Trois estimateurs de $t_y(j)$

On calcule pour chaque Iris j, $j = 1, \ldots, 84$, trois estimateurs de $t_y(j)$:

1. Le premier estimateur est **l' estimateur Horvitz-Thompson** :

$$\hat{t}_y(j) = \frac{N}{n} \sum_{i \in s_j} y_i.$$

2. Le second estimateur est **l'estimateur basé sur le modèle de régression** :

$$\hat{t}_{y,reg}(j) = \sum_{i \in s_j} y_i + \sum_{l \in r_j} \tilde{y}_l, \qquad \text{où } \tilde{y}_l = \widehat{\beta}^\top x_l.$$

3. Le troisième estimateur est **l' estimateur basé sur les modèles de régression géographique pondérée** :

$$\hat{t}_{y,RGP}(j) = \sum_{i \in s_j} y_i + \sum_{l \in r_j} \check{y}_l, \qquad \text{où } \check{y}_l = \hat{\beta}_l^\top x_l.$$

- On répète $K = 1000$ fois ce processus (échantillonnage et estimation).
- On obtient alors pour chaque Iris, 1000 valeurs pour chacun des trois estimateurs.
- On construit des estimations Monte Carlo des erreurs quadratiques moyennes des estimateurs (EQM).

Si on note $\hat{t}_y(j)^{(k)}$ l'estimateur du total de la variable y pour l'Iris j obtenu à la k^e répétition, alors

- L'erreur quadratique moyenne Monte Carlo est définie par :

$$EQM\left(\hat{t}_y(j)\right) = K^{-1}\sum_{k=1}^{K}\left\{\hat{t}_y(j)^{(k)} - t_y(j)\right\}^2.$$

- La racine carrée de l'erreur quadratique moyenne relative (RRMSE):

$$RCEQMR\left(\hat{t}_y(j)\right) = \sqrt{EQM\left(\hat{t}_y(j)\right)}/t_y(j).$$

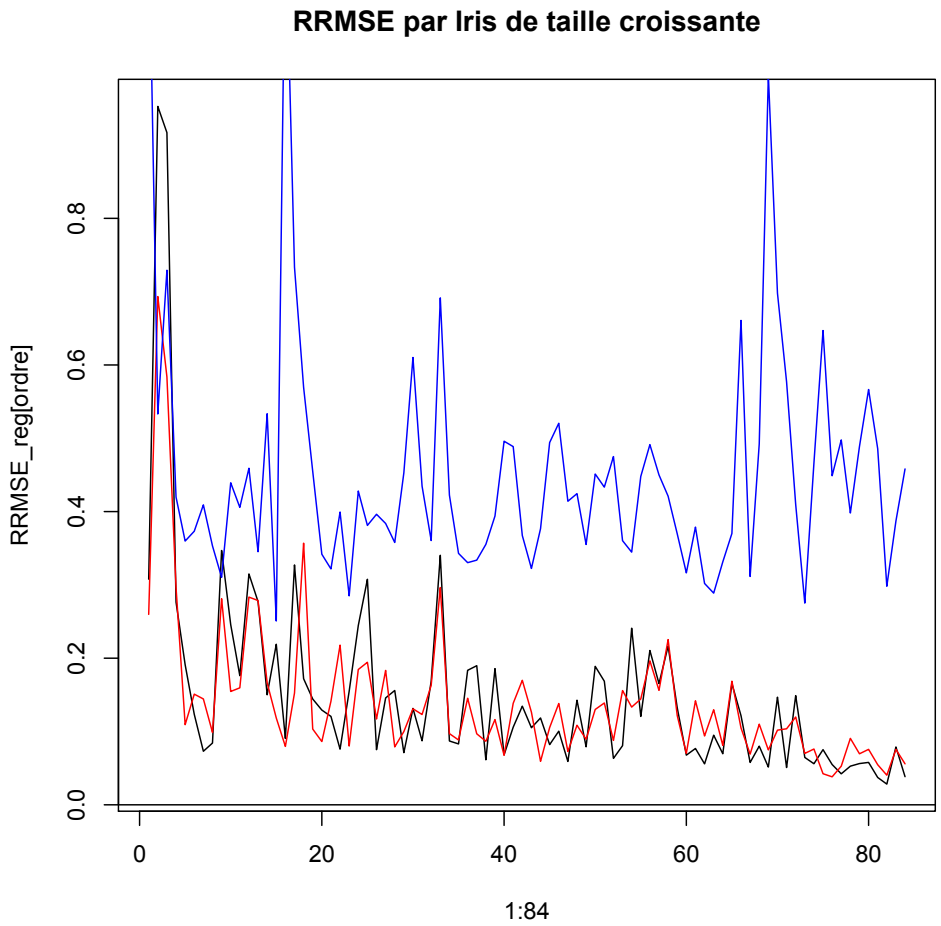


Figure: RCEQMR de l'estimateur Horvitz-Thompson (en bleu), de l'estimateur par la régression (en noir) et de l'estimateur RGP (en rouge) pour les 84 Iris classés par ordre croissant de taille, pour une fenêtre de largeur $h = 500$ mètres.

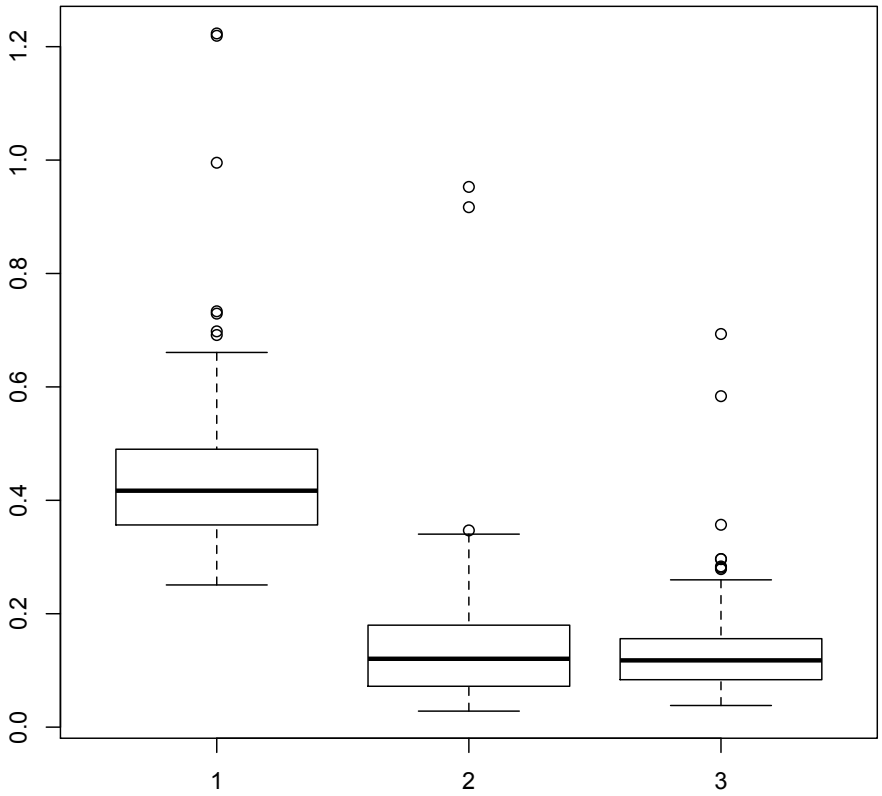


Figure: Boxplot des RCEQMR de l'estimateur Horvitz-Thompson (en 1), de l'estimateur par la régression (en 2) et de l'estimateur RGP (en 3).

- L'estimateur Horvitz-Thompson est moins précis que les deux estimateurs "modèles". Sa RCEQMR médiane pour les estimations par Iris est de l'ordre de 42% (voir la Figure (5)).
- L'estimateur par la régression est nettement plus précis que l'estimateur Horvitz-Thompson : la RCEQMR médiane est de l'ordre de 12%.
  Cette nette amélioration de performance repose sur le fait que la relation linéaire entre la variable y (les ménages à bas revenus) et la variable x (les ménages qui ont la CMUC) est forte.
  On rappelle que le R^2 de ce modèle est d'environ 90%.
- Avec les données réelles employées ici, on ne gagne pas beaucoup en précision lorsqu'on utilise un modèle de régression géographique pondérée.
- L'estimateur RGP est tout de même meilleur que l'estimateur par la régression :
  - pour 75% des Iris la RCMSER de l'estimateur RGP est inférieure à 15.6%.
  - Pour l'estimateur par la régression ce seuil est à 17.8%.

Conclusion

- Cette étude a permis d'illustrer l'apport de la régression géographique pondérée dans le cadre d'un approche modèle en sondage.
  Utilisation à l'Insee pour estimer des statistiques à partir du recensement sur les quartiers de la politique de la ville (Estimation mixte de données du recensement de 2006, site de l'Insee).
- On n'a pas mis en avant un inconvénient majeur de la RGP qui est le temps de calcul. Cette méthode nécessite d'estimer $N - n$ régressions pour chaque variable d'intérêt.
- travaux en cours :
  - Application de cet méthode d'estimation à des jeux de données où le caractère spatial jouerait un plus grand rôle,
  - par exemple, des données disponibles par adresses et non plus par carreaux.
  - Traitement de l'hétérogénéité spatiale à l'étape d'échantillonnage : échantillonnage équilibré spatialement (Grafström et Al, 2012).

Bibliographie partielle

Chambers, Ray and Clark, Robert (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Fotheringham, A. S., Brunsdon, C. et Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Merci de votre attention.