Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Estimation de quantiles sur petits domaines à l'aide de la régression spline et de la vraisemblance empirique

par Zhanshou Chen, Jiahua Chen et Qiong Zhang

Date de diffusion : le 7 mai 2019

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Techniques d'enquête, 2019 (numéro spécial)
Vol. 45, nº 1, pp. 87-106
Statistique Canada, nº 12-001-X au catalogue

Estimation de quantiles sur petits domaines à l'aide de la régression spline et de la vraisemblance empirique

Zhanshou Chen, Jiahua Chen et Qiong Zhang[1]

Résumé

Le présent document étudie l'estimation de quantiles sur petits domaines selon un modèle de régression non paramétrique à erreurs emboîtées au niveau de l'unité. Nous supposons que les distributions des erreurs spécifiques sur petits domaines satisfont un modèle du rapport de densité semi-paramétrique. Nous ajustons le modèle non paramétrique à l'aide de la méthode par régression spline pénalisé d'Opsomer, Claeskens, Ranalli, Kauermann et Breidt (2008). Nous appliquons ensuite la vraisemblance empirique pour estimer les paramètres dans le modèle du rapport de densité à partir des résidus. Cela donne des estimations propres au domaine naturelles des distributions des erreurs. Puis, nous employons une méthode des noyaux pour obtenir des estimations lissées des distributions des erreurs. Ces estimations sont alors utilisées pour faire une estimation de quantiles dans deux situations : dans l'une d'elles, nous ne connaissons que les moyennes de puissances des covariables au niveau de la population; dans l'autre, nous connaissons les valeurs des covariables de toutes les unités d'échantillonnage dans la population. Selon des expériences de simulation, les méthodes proposées pour l'estimation des quantiles sur petits domaines fonctionnent bien pour des quantiles situés près de la médiane dans le premier cas et pour un large éventail de quantiles dans le second. Un estimateur de l'erreur quadratique moyenne bootstrap des estimateurs proposés est également examiné. Un exemple empirique fondé sur les données sur les revenus des Canadiens en fait partie.

Mots-clés : Quantile sur petits domaines; spline pénalisé; vraisemblance empirique; modèle du rapport de densité; modèle de régression à erreurs emboîtées.

1 Introduction

Les enquêtes-échantillons sont largement utilisées pour obtenir de l'information sur les totaux, les moyennes, les médianes et d'autres quantités de populations finies. De même, des données semblables sur des sous-populations, comme des individus dans des régions et des groupes sociodémographiques particuliers, présentent aussi un intérêt. Souvent, une enquête est conçue pour recueillir de l'information d'intérêt au niveau de la population, mais cela donne des données directes insuffisantes sur les sous-populations. Voilà pourquoi l'estimation des paramètres des sous-populations avec une précision satisfaisante et l'évaluation de leur exactitude sont de sérieux défis pour les statisticiens. Ces derniers doivent se tourner vers des modèles appropriés pour regrouper l'information des petits domaines afin de bien estimer les paramètres pour les petits domaines quand aucun échantillon ou seulement de petits échantillons dans ces domaines sont disponibles à partir de l'enquête-échantillon.

Les recherches consacrées à l'estimation sur petits domaines attirent de plus en plus l'attention des secteurs public et privé. Pour faire un petit rappel historique, mentionnons Fay et Herriot (1979), Battese, Harter et Fuller (1988), Prasad et Rao (1990), et Lahiri et Rao (1995), entre autres. Pour un examen général de l'évolution de l'estimation sur petits domaines, mentionnons Pfeffermann (2002) et Pfeffermann (2013) et les ouvrages de Rao (2003) et Rao et Molina (2015). Voir également Jiang et Lahiri (2006a), Jiang et Lahiri (2006b) et Jiang (2010) pour les publications récentes.

1. Zhanshou Chen, School of Mathematics and Statistics, Université normale Qinghai, Xining 810008, P.R. Chine. Courriel : chenzhanshou@126.com; Jiahua Chen et Qiong Zhang, Département de la statistique, Université de la Colombie-Britannique, Vancouver (C.-B.) Canada.

Comparativement aux quantiles, il existe relativement plus d'activités de recherche sur l'estimation des moyennes de petits domaines. Les études sur l'estimation de quantiles sur petits domaines gagnent du terrain. L'approche M-quantile de Chambers et Tzavidis (2006) a remporté un franc succès. Cette approche utilise l'approche M-quantile pour caractériser les distributions conditionnelles de la variable de réponse y pour des covariables $\mathbf{x}$. Cette information sert ensuite à prédire les valeurs de réponses non observées à partir desquelles les distributions de la population des petits domaines sont estimées. L'estimation de quantiles sur petits domaines est un avantage secondaire naturel et bien accueilli. Voir Tzavidis et Chambers (2005), Pratesi, Ranalli et Salvati (2008), Tzavidis, Salvati et Pratesi (2008), et Salvati, Tzavidis et Pratesi (2012) pour en connaître l'évolution.

Une autre approche d'estimation de quantiles sur petits domaines est proposée par Molina (2010). Supposons que s et r sont les ensembles d'unités échantillonnées et non échantillonnées dans une enquête et y_s et y_r sont les vecteurs des valeurs de réponses correspondantes. À l'aide d'une hypothèse paramétrique sur la distribution conjointe de y_s et y_r (ou les réponses transformées), ils ont proposé de calculer la distribution conditionnelle de y_r sachant y_s (et d'autre information). Après avoir dûment estimé la distribution conjointe et, par conséquent, la distribution conditionnelle, ils ont suggéré l'échantillonnage à partir de la distribution conditionnelle estimée pour créer une population artificielle, mais complète, une fois que y_r non observé était rempli. La distribution de la population est estimée à partir de la population complète. Cette approche fonctionne bien pour estimer les quantiles et les moyennes de petits domaines. Les autres méthodes dont nous sommes au courant comprennent celles de Tzavidis, Marchetti et Chambers (2010), Chaudhuri et Ghosh (2011) et Chen et Liu (2018). Tzavidis et coll. (2010) ont proposé un cadre général pour une estimation sur petits domaines robuste, en représentant l'estimateur sur petits domaines comme une fonction d'une variable explicative de la fonction de distribution cumulative de ces petits domaines. Chaudhuri et Ghosh (2011) ont proposé une vraisemblance empirique qui repose sur la méthode bayésienne. Chen et Liu (2018) ont proposé une approche pour les populations en admettant un modèle de régression linéaire à erreurs emboîtées combiné à des distributions des erreurs qui satisfont un modèle du rapport de densité semi-paramétrique (MRD). Selon les simulations, la méthode fondée sur le MRD ressort du lot quand les distributions des erreurs sont asymétriques.

Dans le présent document, nous nous intéressons à la situation où la fonction de régression n'est pas linéaire, même si le modèle de régression à erreurs emboîtées demeure dûment semblable à celui d'Opsomer et coll. (2008). De toute évidence, les méthodes obtenues à l'aide de modèles linéaires peuvent donner un biais considérable si l'hypothèse de la linéarité est enfreinte. Pour réduire l'éventuel risque de biais majeur, Opsomer et coll. (2008) ont proposé une meilleure prédiction linéaire sans biais empirique (EBLUP) pour les moyennes de petits domaines selon un modèle de régression non paramétrique à l'aide de splines pénalisés (P-splines); Jiang, Ngueyen et Rao (2010) ont conçu une approche de barrière adaptative en employant une technique de sélection de modèle non paramétrique; Sperlich et José Lombardía (2010) ont eu recours à la méthode d'inférence locale polynomiale dans le contexte de l'estimation sur petits domaines;

Rao, Sinha et Dumitrescu (2014) ont proposé une EBLUP robuste à l'aide d'un modèle mixte approximé P-splines; Torabi et Shokoohi (2015) ont proposé une analyse unifiée des réponses discrètes et continues grâce à des modèles de régression P-spline.

Nous suivons leur exemple et élargissons leurs résultats pour permettre des distributions d'erreur non normales dans le modèle de régression non paramétrique à erreurs emboîtées. Plus précisément, nous établissons l'hypothèse du modèle de régression non paramétrique à erreurs emboîtées, mais nous assouplissons l'hypothèse de la distribution des erreurs sur petits domaines d'une normale à un MRD semi-paramétrique souple. Nous utilisons l'approche de régression P-splines d'Opsomer et coll. (2008) pour ajuster la régression non linéaire. Nous appliquons ensuite la vraisemblance empirique pour estimer les paramètres du MRD à l'aide des résidus. Cela donne une estimation naturelle de la distribution des erreurs pour des domaines spécifiques. Nous appliquons ensuite une méthode des noyaux pour obtenir des estimations lissées des distributions des erreurs et des quantiles sur petits domaines. Nous construisons des estimations des quantiles dans deux situations : lorsque nous connaissons uniquement les moyennes de puissances des covariables au niveau de la population et lorsque nous disposons des valeurs des covariables de toutes les unités d'échantillonnage dans la population. Notre approche devrait hériter des mérites du travail à partir d'un modèle de régression non paramétrique et profiter du fait que l'hypothèse d'une distribution des erreurs paramétrique est évitée. Les estimations des quantiles sur petits domaines ainsi obtenues sont donc plus robustes. Les simulations indiquent que, lorsque la fonction de régression est approximativement linéaire, le rendement de l'approche proposée est concurrentiel. L'approche proposée donne un meilleur résultat quand la relation de régression est quadratique ou exponentielle.

Le reste du document est organisé comme suit. Dans la section 2, nous introduisons le modèle et les hypothèses. Dans la section 3, nous présentons l'approche proposée. Dans la section 4, nous proposons une procédure bootstrap pour estimer les erreurs quadratiques moyennes. Dans la section 5, nous avons recours à des méthodes de Monte Carlo pour évaluer le rendement de la méthode proposée et la comparer à certaines méthodes existantes. Un exemple d'application est présenté dans la section 6. La section 7 renferme quelques observations finales.

2 Modèle et hypothèses

Prenons une population finie contenant $N = \sum_{i=0}^{m} N_i$ unités d'échantillonnage divisées en $m+1$ petits domaines $\{(x_{ij}, y_{ij}) : j = 1, 2, \ldots, N_i\}$, $i = 0, 1, \ldots, m$. Prenons un modèle de régression non paramétrique à erreurs emboîtées avec une covariable :

$$y_{ij} = m_0(x_{ij}) + v_i + \varepsilon_{ij}, \tag{2.1}$$

où x_{ij} est une variable auxiliaire, v_i désigne an effet aléatoire propre au domaine et ε_{ij} représente des erreurs aléatoires. La fonction de régression $m_0(\cdot)$ n'est pas précisée, mais nous pouvons calculer assez facilement une approximation à l'aide d'une fonction spline

$$m_0(x; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)_+^p. \tag{2.2}$$

Ici, p est le degré de l'estimation spline, $x_+^p = x^p$ lorsque $x > 0$ et 0 autrement, κ_k, $k = 1, \ldots, K$ constituent un ensemble de constantes fixes appelées nœuds, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ est un vecteur de coefficient de la portion paramétrique du modèle, et $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$ est le vecteur des coefficients splines, K est le nombre de nœuds splines. Si l'emplacement des noeuds couvre toute la plage de x et K est suffisamment important, la classe de P-spline (2.2) peut servir à estimer une fonction lisse $m_0(\cdot)$ avec un niveau d'exactitude élevé, même pour un petit p (Boor, 2001). Ruppert, Wand et Carroll (2003) ont recommandé d'utiliser le nombre de nœuds splines K comme le minimum de 40 et le nombre de x uniques divisé par 4.

Nous supposons que nous obtenons un échantillon aléatoire de la population à l'aide d'un plan de sondage non informatif, de telle sorte que (2.1) demeure valide pour les unités échantillonnées. Notre tâche immédiate consiste à ajuster ce modèle à partir des données échantillonnées et de suivre l'approche d'Opsomer et coll. (2008). Pour simplifier la présentation, nous introduisons d'abord une notation matricielle. Supposons que n_i est le nombre d'unités échantillonnées dans le petit domaine i. Les valeurs de réponses des i^{e} domaines sont désignées comme $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$. Nous les regroupons ensuite pour créer le vecteur de réponse dont la longueur est n : $\mathbf{Y}_n' = (\mathbf{y}_0', \mathbf{y}_1', \ldots, \mathbf{y}_m')$. De même, nous définissons $\boldsymbol{\epsilon}_i$ et $\boldsymbol{\epsilon}_n$ pour le terme d'erreur. Nous utilisons $\mathbf{v} = (v_0, \ldots, v_m)'$ pour les effets aléatoires propres à un domaine et créons une matrice $\mathbf{D}$ de sorte que

$$\mathbf{D}\mathbf{v} = (v_0 1_{n_0}', v_1 1_{n_1}', \ldots, v_m 1_{n_m}')$$

où 1_k est un vecteur de longueur k de 1. Nous construisons ensuite des matrices $\mathbf{X}_n$ et $\mathbf{Z}_n$ de telle sorte que leurs rangées soient composées de

$$\mathbf{x}_{ij}' = (1, x_{ij}, \ldots, x_{ij}^p), \qquad \mathbf{z}_{ij}' = \left((x_{ij} - \kappa_1)_+^p, \ldots, (x_{ij} - \kappa_K)_+^p\right)$$

dans l'ordre approprié. Avec ces matrices et ces vecteurs, les données de l'échantillon sous le modèle (2.1) sont reliées par

$$\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\beta} + \mathbf{Z}_n\boldsymbol{\gamma} + \mathbf{D}\mathbf{v} + \boldsymbol{\epsilon}_n. \tag{2.3}$$

Opsomer et coll. (2008) ont ajusté ce modèle selon l'hypothèse voulant que les composantes de $\boldsymbol{\gamma}$, de $\mathbf{v}$ et de $\boldsymbol{\epsilon}$ soient toutes indépendantes et normalement distribuées de manière identique avec les variances σ_γ^2, σ_v^2 et σ_ϵ^2 respectivement. Nous obtenons les solutions de l'ajustement comme suit

$$\begin{aligned}
\hat{\mathbf{V}} &= \mathbf{Z}_n\boldsymbol{\Sigma}_\gamma\mathbf{Z}_n' + \mathbf{D}\hat{\boldsymbol{\Sigma}}_v\mathbf{D}' + \hat{\boldsymbol{\Sigma}}_\varepsilon, \\
\hat{\mathbf{v}} &= \hat{\boldsymbol{\Sigma}}_v\mathbf{D}'\hat{\mathbf{V}}^{-1}\left(\mathbf{Y}_n - \mathbf{X}_n\hat{\boldsymbol{\beta}}\right), \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}_n'\hat{\mathbf{V}}^{-1}\mathbf{X}_n\right)^{-1}\left(\mathbf{X}_n'\hat{\mathbf{V}}^{-1}\mathbf{Y}_n\right), \\
\hat{\boldsymbol{\gamma}} &= \hat{\boldsymbol{\Sigma}}_\gamma\mathbf{Z}_n'\hat{\mathbf{V}}^{-1}\left(\mathbf{Y}_n - \mathbf{X}_n\hat{\boldsymbol{\beta}}\right)
\end{aligned}$$

où $\hat{\boldsymbol{\Sigma}}_{\gamma}$, $\hat{\boldsymbol{\Sigma}}_{v}$, $\hat{\boldsymbol{\Sigma}}_{\varepsilon}$ sont des estimations du maximum de vraisemblance restreinte (REML) pour les matrices de covariance de $\boldsymbol{\gamma}$, $\mathbf{v}$ et $\boldsymbol{\epsilon}$, et $\hat{\mathbf{V}}$ est l'estimation de $\mathbf{V} \equiv \text{var}(\mathbf{Y}_n)$.

Opsomer et coll. (2008) ont ensuite établi la meilleure prédiction linéaire sans biais empirique de la moyenne de petit domaine :

$$\hat{\bar{Y}}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_i + \ldots + \hat{\beta}_p \bar{X}_i^p + \bar{\mathbf{z}}_i \hat{\boldsymbol{\gamma}} + \hat{v}_i, \tag{2.4}$$

où $\bar{X}_i, \ldots, \bar{X}_i^p$ représentent les moyennes de puissances des unités de population x_{ij} dans le domaine i, c'est-à-dire, $\bar{X}_i^s = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}^s$ pour $s = 1, \ldots, p$, et $\bar{\mathbf{z}}_i \hat{\boldsymbol{\gamma}}$ désigne les vraies moyennes des fonctions de base de spline pour le petit domaine i. De toute évidence, il est facile d'élargir la discussion qui précède à des modèles additifs non paramétriques comportant deux covariables ou plus (Lin et Zhang (1999), Ruppert et coll. (2003) et Wood (2006)).

Dans le présent document, nous suivons les travaux d'Opsomer et coll. (2008) pour obtenir toutes les valeurs ajustées. Pour l'estimation de quantiles sur petits domaines, nous supprimons l'hypothèse de la normalité pour ϵ_{ij}. Nous supposons plutôt que leur distribution $G_i(u)$ satisfait un MRD, de telle sorte que pour $i = 1, \ldots, m$,

$$\log\{dG_i(u)/dG_0(u)\} = \theta_i' \mathbf{q}(u), \tag{2.5}$$

avec une fonction de base prédéterminée $\mathbf{q}(u)$ et un paramètre de basculement propre au domaine θ_i. Nous pouvons inclure $i = 0$ dans l'équation ci-haut en posant $\theta_0 = 0$. Il faut que le premier élément de $\mathbf{q}(u)$ soit un pour que le premier élément de θ_i soit un paramètre de normalisation. Le MRD comprend des familles de distributions normale, gamma et bien d'autres familles comme des cas spéciaux. Des discussions sur le MRD se trouvent dans Anderson (1979), Qin et Zhang (1997), Kezioua et Leoni-Aubina (2008) et Chen et Liu (2013).

Les équations (2.1), (2.2) et (2.5) forment ensemble la plateforme du présent document pour l'estimation de quantiles sur petits domaines. Nos travaux sont différents de ceux d'Opsomer et coll. (2008) en ce sens que nous nous concentrons sur l'estimation de quantiles sur petits domaines sans énoncer d'hypothèse de normalité pour $G_i(\cdot)$. Parallèlement, le présent document est différent de Chen et Liu (2018) parce qu'il formule des postulats pour une relation de régression non paramétrique entre y_{ij} et x_{ij} au lieu d'une relation linéaire.

3 Approche proposée

Pour tout $\alpha \in (0, 1)$, le α^{e} quantile d'une distribution F se définit comme

$$\xi_\alpha = \inf\{u : F(u) \geq \alpha\}.$$

Si $\hat{F}(u)$ est une estimation de $F(u)$, son α- quantile est naturellement estimé à l'aide de

$$\hat{\xi}_{\alpha} = \inf\left\{u : \hat{F}(u) \geq \alpha\right\}. \tag{3.1}$$

Selon l'hypothèse de distribution pour ϵ_{ij}, nous avons

$$\begin{aligned} P\left(y_{ij} \leq u\right) &= \mathbb{E}\left\{P\left(\varepsilon_{ij} \leq u - m_0\left(x_{ij}\right) - v_i \mid x_{ij}, v_i\right)\right\} \\ &= \mathbb{E}\left\{G_i\left(u - m_0\left(x_{ij}\right) - v_i\right)\right\}. \end{aligned}$$

Par conséquent, nous obtenons la distribution de la population du i^{e} petit domaine comme suit

$$F_i(u) = N_i^{-1}\sum_{j=1}^{N_i} G_i\left(u - m_0\left(x_{ij}\right) - v_i\right).$$

Une fois que G_i et $m_0(\cdot)$ seront correctement estimés, il en ira de même des quantiles sur petits domaines.

Nous suivons l'idée de la vraisemblance empirique de Chen et Liu (2018) pour l'estimation de $G_i(\cdot)$. Supposons que les valeurs de ε_{ij} dans l'échantillon sont connues. Prenons un candidat G_0 sous la forme suivante

$$G_0(u) = \sum_{i,j} p_{ij} I\left(\varepsilon_{ij} \leq u\right),$$

où $I(\cdot)$ est une fonction indicatrice et $\sum_{i,j} = \sum_{i=0}^{m}\sum_{j=1}^{n_i}$. Par conséquent, nous avons $p_{ij} = dG_0\left(\varepsilon_{ij}\right)$ et, selon le MRD, $dG_i(\varepsilon_{st}) = p_{st}\exp\left\{\boldsymbol{\theta}_i'\mathbf{q}(\varepsilon_{st})\right\}$ pour $i = 0, 1, \ldots, m$ ce qui implique ceci

$$G_i(u) = \sum_{s,t} p_{st}\exp\left\{\boldsymbol{\theta}_i'\mathbf{q}(\varepsilon_{st})\right\} I(\varepsilon_{st} \leq u). \tag{3.2}$$

Selon Owen (2001), nous obtenons la fonction de vraisemblance empirique

$$L_n(G_0, G_1, \ldots, G_m) = \prod_{i,j} dG_i\left(\varepsilon_{ij}\right) = \left\{\prod_{i,j} p_{ij}\right\}\exp\left[\sum_{i,j}\left\{\ \boldsymbol{\theta}_i'\mathbf{q}\left(\varepsilon_{ij}\right)\right\}\right],$$

où le paramètre $\boldsymbol{\theta}$ et p_{ij} satisfait $p_{ij} \geq 0$, et où $s = 0, 1, \ldots, m$,

$$\sum_{i,j} p_{ij}\exp\left\{\boldsymbol{\theta}_s'\mathbf{q}\left(\varepsilon_{ij}\right)\right\} = 1. \tag{3.3}$$

À noter que nous avons utilisé la convention $\boldsymbol{\theta}_0 = 0$ pour simplifier la présentation. Parce que $G_1, \ldots, G_m$ sont entièrement déterminés par $\boldsymbol{\theta}' = \left(\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_m'\right)$ et G_0, nous écrivons le log-vraisemblance empirique comme ceci

$$\ell_n(\boldsymbol{\theta}, G_0) = \sum_{i,j}\log\left(p_{ij}\right) + \sum_{ij}\boldsymbol{\theta}_i'\mathbf{q}\left(\varepsilon_{ij}\right).$$

En optimisant $\ell(\boldsymbol{\theta}, G_0)$ en ce qui concerne G_0 selon les résultats des contraintes (3.3) dans les probabilités ajustées

$$\hat{p}_{ij} = n^{-1}\left\{1 + \sum_{s=1}^{m}\lambda_s\left[\exp\left\{\boldsymbol{\theta}_s'\mathbf{q}(\varepsilon_{ij})\right\} - 1\right]\right\}^{-1} \tag{3.4}$$

et le log-VE du profil

$$\ell_n(\boldsymbol{\theta}) = -\sum_{i,j} \log\left\{1 + \sum_{s=1}^{m} \lambda_s \left[\exp\left\{\boldsymbol{\theta}_s' \mathbf{q}\left(\varepsilon_{ij}\right)\right\} - 1\right]\right\} + \sum_{i,j} \boldsymbol{\theta}_i' \mathbf{q}\left(\varepsilon_{ij}\right)$$

où $(\lambda_1, \ldots, \lambda_m)$ est la solution de

$$\sum_{s,t} \frac{\exp\left\{\boldsymbol{\theta}_i' \mathbf{q}\left(\varepsilon_{st}\right)\right\} - 1}{1 + \sum_{l=1}^{m} \lambda_l \left[\exp\left\{\boldsymbol{\theta}_l' \mathbf{q}\left(\varepsilon_{st}\right)\right\} - 1\right]} = 0.$$

Puisque les valeurs de ε_{ij} ne sont pas disponibles, nous les remplaçons par les résidus obtenus à l'aide du modèle d'ajustement (2.1) selon l'hypothèse (2.2) :

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{m}_0\left(x_{ij}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) - \hat{v}_i$$

où

$$\hat{m}_0(x; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \hat{\beta}_0 + \hat{\beta}_1 x + \ldots + \hat{\beta}_p x^p + \sum_{k=1}^{K} \hat{\gamma}_k \left(x - \kappa_k\right)_+^p . \tag{3.5}$$

Supposons que $\hat{\ell}_n(\boldsymbol{\theta})$ est la fonction du log-VE $\tilde{\ell}_n(\boldsymbol{\theta})$ après avoir remplacé ε_{ij} par $\hat{\varepsilon}_{ij}$. Nous définissons l'estimateur VE maximum de $\boldsymbol{\theta}$ à l'aide de $\hat{\boldsymbol{\theta}} = \operatorname{argmax} \hat{\ell}_n(\boldsymbol{\theta})$ et nous estimons $G_i(u)$ comme ceci

$$\tilde{G}_i(u) = \sum_{s,t} \hat{p}_{st} \exp\left\{\hat{\boldsymbol{\theta}}_i' \mathbf{q}\left(\hat{\varepsilon}_{st}\right)\right\} I\left(\hat{\varepsilon}_{st} \le u\right) \tag{3.6}$$

où

$$\hat{p}_{st} = n^{-1} \left\{1 + \sum_{l=1}^{m} (n_l / n)\left[\exp\left\{\boldsymbol{\theta}_l' \mathbf{q}\left(\hat{\varepsilon}_{st}\right)\right\} - 1\right]\right\}^{-1}$$

et $\hat{\boldsymbol{\theta}}_0 = 0$. La routine R **drmdel** peut servir à calculer $\hat{\boldsymbol{\theta}}$ et $\hat{p}_{ij}$ qui offre 11 possibilités de fonction de base $\mathbf{q}(u)$.

Parce que $\tilde{G}_i(u)$ est discret, la distribution lissée des noyaux $\hat{G}_i(u)$ suivante donne une meilleure estimation des quantiles :

$$\hat{G}_i(u) = \sum_{j=1}^{n_i} \hat{w}_{ij} \Phi\left(\frac{\hat{\varepsilon}_{ij} - u}{b}\right), \tag{3.7}$$

où les poids sont choisis comme $\hat{w}_{ij} = \tilde{G}_i\left(\hat{\varepsilon}_{ij}\right) - \tilde{G}_i\left(\hat{\varepsilon}_{ij} -\right)$, b est un paramètre de largeur de bande, et $\Phi(\cdot)$ est la fonction de distribution de la normale type. Comme l'ont proposé Chen et Liu (2013), nous choisissons $b = 1{,}06 n^{-1/5} \min\left\{\hat{\sigma}, \hat{Q}/1{,}34\right\}$ où $\hat{\sigma}$ est l'écart-type de la distribution $\hat{G}_i$ et $\hat{Q}$ est son intervalle interquartile.

Dans certaines applications, seules les moyennes de puissances des covariables de la population sont connues et peuvent être utilisées pour établir une inférence statistique. Dans d'autres applications, les covariables de tous les membres de la population sont connues. Cela donne deux estimations possibles des quantiles. Dans le premier cas, nous estimons F_i comme

$$\hat{F}_i^{(a)}(u) = n_i^{-1}\sum_{j=1}^{n_i}\hat{G}_i\left(u - \hat{\bar{Y}}_i - \left\{\hat{m}_0\left(x_{ij};\, \hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\gamma}}\right) - \hat{m}_0\left(\bar{x}_i;\, \hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\gamma}}\right)\right\}\right), \tag{3.8}$$

où nous utilisons $\hat{m}_0\left(\bar{x}_i;\, \hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\gamma}}\right)$ précisés sous (3.5).

Lorsque les données du recensement sur x sont disponibles, nous estimons F_i comme ceci

$$\hat{F}_i^{(b)}(u) = N_i^{-1}\left\{\sum_{j\in s_i} I\left(y_{ij} \le u\right) + \sum_{j\in r_i}\hat{G}_i\left(u - \hat{m}_0\left(x_{ij}\right) - \hat{v}_i\right)\right\}, \tag{3.9}$$

où s_i et r_i sont des ensembles d'unités observées et non observées dans le petit domaine i. Le reste des spécifications est identique à (3.8).

Les estimations proposées ressemblent à celles de Chen et Liu (2018), mais nous utilisons une régression non paramétrique. Parce qu'il est plus facile de recueillir les moyennes de puissances des covariables de la population que les valeurs des covariables de toutes les unités de la population, $\hat{F}_i^{(a)}(u)$ s'applique de façon plus large que $\hat{F}_i^{(b)}(u)$. Les calculs sont également plus efficaces. Parce que $\hat{F}_i^{(b)}(u)$ utilise les valeurs des covariables de toutes les unités de la population, il devrait donner de meilleurs résultats statistiques quand les deux s'appliquent.

4 Estimation bootstrap des erreurs quadratiques moyennes

Les estimateurs de quantiles sur petits domaines proposés sont assemblés en suivant de nombreuses étapes intermédiaires. Il est difficile d'évaluer de manière analytique les variances ou l'erreur quadratique moyenne (EQM) de ces estimateurs. Nous suivons d'autres chercheurs (Sinha et Rao (2009), Tzavidis et coll. (2010) et Chen et Liu (2018)) afin d'élaborer une procédure bootstrap comme suit :

Étape 1 Obtenir les estimations $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$, $\hat{\sigma}_v^2$ et $\hat{m}_0\left(x,\, \hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\gamma}}\right)$ à partir du modèle (2.1), et calculer $\hat{G}_i(u)$ comme dans (3.7).

Étape 2 Générer une population bootstrap finie $H^* = \left\{y_{ij}^*,\, x_{ij}\right\}$, $i = 0, \ldots, m,\ j = 1, \ldots, N_i$ avec

$$y_{ij}^* = \hat{m}_0\left(x_{ij},\, \hat{\boldsymbol{\beta}},\, \hat{\boldsymbol{\gamma}}\right) + v_i^* + \varepsilon_{ij}^*,$$

où les résidus bootstrap ε_{ij}^* sont échantillonnés à partir de la FDC $\hat{G}_i(u)$, et les v_i^* sont générés à partir de $N\left(0,\, \hat{\sigma}_v^2\right)$.

Étape 3 À partir de la population bootstrap H^*, nous sélectionnons $n_i^* = n_i$ unités d'échantillonnage dans le petit domaine i par échantillonnage aléatoire simple sans remise, et nous le répétons L fois pour obtenir h_l^*, $l = 1, \ldots, L$. Pour chaque échantillon h_l^*, il faut calculer les estimations $\hat{F}_i^{(a)*l}(u)$ et $\hat{F}_i^{(b)*l}(u)$ comme dans (3.8) et (3.9) respectivement.

Étape 4 Calculer l'estimateur de l'EQM empirique de $\hat{\tau}$ comme ceci

$$\text{eqm}\left(\tau^*\right) = L^{-1}\sum_{l=1}^{L}\left(\hat{\tau}^{*l} - \tau^*\right)^2,$$

où $\hat{\tau}^{*l} = \tau\left(\hat{F}^{*l}(u)\right)$ désigne une fonction de $\hat{F}^{(a)*l}(u)$ ou $\hat{F}^{(b)*l}$ et $\tau^* = \tau\left(F^*(u)\right)$ où $F^*(u)$ est la FDC connue des populations bootstrap.

Étape 5 Répéter les étapes 2 à 4, B fois, et définir l'estimation de l'EQM bootstrap comme ceci

$$B^{-1}\sum_{b=1}^{B}\text{eqm}\left(\tau^*\right)_b,$$

où $\text{eqm}\left(\tau^*\right)_b$ est $\text{eqm}\left(\tau^*\right)$ calculée à la b^{e} répétition.

Le rendement de l'estimateur de l'EQM bootstrap sera examiné et déclaré dans la section portant sur la simulation.

5 Simulations de Monte Carlo

Dans la présente section, nous utilisons une simulation pour évaluer le rendement des estimateurs de vraisemblance empirique fondés sur le modèle proposé de régression spline pénalisé (VEP) et leurs estimations de l'EQM. Lorsque seules les moyennes de la population de covariables sont connues, les estimateurs proposés sont comparés uniquement à l'estimateur de la vraisemblance empirique fondé sur le modèle de régression linéaire à erreurs emboîtées (VEL) de Chen et Liu (2018) et à l'estimateur direct (ED). Lorsque les valeurs des covariables sont connues pour toutes les unités d'échantillonnage, la comparaison est élargie afin d'inclure également six estimateurs de Tzavidis et coll. (2010), désignés comme EBLUP/naïve, EBLUP/CD, EBLUP/RKM, M-quantile/naïve, M-quantile/CD et M-quantile/RKM. Ici, EBLUP/CD et M-quantile/CD désignent la EBLUP et l'estimateur de M-quantile s'obtient à partir de la FDC proposée par Chambers et Dunstan (1986), et les estimateurs correspondants fondés sur la FDC proposée par Rao, Kovar et Mantel (1990), désignés comme RKM.

Tout comme Chen et Liu (2018), nous devons choisir $\mathbf{q}(u)$ dans le MRD. Il y a deux candidats, soit $\mathbf{q}_1(u) = (1, u)'$ et $\mathbf{q}_2(u) = \left(1, \text{sign}(u)\sqrt{|u|}\right)'$. Selon quelques résultats provisoires de la simulation, $\mathbf{q}_1(u) = (1, u)'$ fonctionne bien pour le modèle de régression non paramétrique ajusté aux P-splines, mais ce n'est pas le cas de $\mathbf{q}_2(u)$. Le choix de $\mathbf{q}_2^*(u) = (1, u, u^2)'$ donne plutôt un rendement concurrentiel. Nous utilisons donc $\mathbf{q}_1(u)$ et $\mathbf{q}_2^*(u)$ dans notre simulation.

Dans le sillage de Rao et coll. (2014) et de Torabi et Shokoohi (2015), nous avons généré des données à partir de trois modèles :

$$\begin{aligned}
\text{A}: \quad & y_{ij} = 1 + x_{ij} + v_i + \varepsilon_{ij}, \\
\text{B}: \quad & y_{ij} = 1 + x_{ij} + x_{ij}^2 + v_i + \varepsilon_{ij}, \\
\text{C}: \quad & y_{ij} = 1 - x_{ij} + 0{,}5\exp\left(x_{ij}\right) + v_i + \varepsilon_{ij}.
\end{aligned}$$

Ceux-ci donnent respectivement des fonctions de régressions linéaires, quadratiques et exponentielles. Nous avons établi le nombre de petits domaines à 30 et la taille de la population d'un domaine à $N_i = 500(i+1)$, $i = 0, 1, \ldots, 29$. Nous avons généré la covariable x_{ij} à partir de $N(0, 1)$. Après avoir généré x_{ij}, nous les avons traités comme étant fixes dans la simulation. L'effet aléatoire propre au domaine v_i a été obtenu à partir de $N(0, 1)$, et les erreurs ε_{ij} ont été obtenues à partir des quatre distributions suivantes.

$$
\begin{aligned}
&\text{(i)}: \quad N(0, 1),\\
&\text{(ii)}: \quad t(3),\\
&\text{(iii)}: \quad \text{combinaison normale } 0{,}5N(-1; 1) + 0{,}5N(1; 1),\\
&\text{(iv)}: \quad N(0, \sigma_i^2), \text{ où } \sigma_i \sim U(0{,}5; 2),\ i = 0, \ldots, 29.
\end{aligned}
$$

La distribution (ii) a une queue lourde, les distributions (ii) et (iii) sont symétriques, et la distribution (iv) est hétéroscédastique.

Nous avons utilisé $R = 1\,000$ répétitions dans la simulation et prélevé des échantillons aléatoires de taille $n = 500$ sans remplacement de la population à chaque répétition. Pour éviter la possibilité que certains petits domaines comportent trop peu d'unités d'échantillonnage, nous avons pris $n - 60$ unités au niveau de la population et affecté deux autres unités à chaque petit domaine. Nous avons utilisé la routine R **mgcv** pour la méthode REML avec des options par défaut pour les valeurs de p et K lors de l'ajustement de la fonction P-spline (2.4). Nous avons calculé les estimations des quantiles sur petits domaines de 5 %, 25 %, 50 %, 75 % et 95 % désignés comme ED, VEL1, VEL2, VEP1, VEP2, pour l'estimateur direct, les estimateurs de Chen et Liu (2018) et les estimateurs proposés à l'aide de $\mathbf{q}_1(\cdot)$ et $\mathbf{q}_2(\cdot)$. Nous présentons leur erreur quadratique moyenne relative (AMSE) et les biais absolus (ABIAS) qui sont définis comme ceci :

$$
\begin{aligned}
\text{AMSE} &= \{R(m+1)\}^{-1} \sum_{i=0}^{m} \sum_{r=1}^{R} \left(\hat{\xi}_i^{(r)} - \xi_i^{(r)}\right)^2,\\
\text{ABIAS} &= (m+1)^{-1} \sum_{i=0}^{m} \left| R^{-1} \sum_{r=1}^{R} \hat{\xi}_i^{(r)} - R^{-1} \sum_{r=1}^{R} \xi_i^{(r)} \right|,
\end{aligned}
$$

où $\hat{\xi}_i^{(r)}$ est soit l'une des estimations des quantiles pour le i^{e} petit domaine dans la r^{e} répétition. Les résultats sous les modèles A, B, et C figurent aux tableaux 5.1 à 5.3, respectivement. La VEP et la VEL sont fondées sur $\hat{F}_i^{(a)}$ et sa version mirroir dans Chen et Liu (2018).

Sous le modèle A, le modèle linéaire est valide. Par conséquent, nous nous attendons à ce que la VEL soit supérieure. Selon le tableau 5.1, deux méthodes sont similaires pour les quantiles de 25 %, 50 % et 75 %. Les VEL donnent de meilleurs résultats que les VEP pour le quantile de 5 %, tandis que la comparaison est inversée pour le quantile de 95 %. Les VEP et les VEL donnent un meilleur résultat que l'ED pour les quantiles de 25 %, 50 % et 75 % avec des marges importantes. Cela donne l'impression générale que les méthodes proposées fonctionnent toujours de façon satisfaisante.

Sous le modèle B, le modèle linéaire se détériore légèrement. Les résultats du tableau 5.2 montrent que les estimateurs des VEP ont une AMSE plus basse pour les quantiles inférieurs. Les VEL ont toujours une AMSE basse, malgré un ABIAS plus élevé. L'avantage des VEP proposées dans les modèles de régressions non paramétriques à erreurs emboîtées tient dans les quantiles des niveaux du centre. Comme il y a moins d'observations dans les quantiles qui s'approchent des extrémités, il est difficile d'ajuster le modèle non paramétrique.

La linéarité est gravement enfreinte sous le modèle C. La VEL devrait avoir un mauvais rendement et cela est évident dans le tableau 5.3. Parallèlement, les VEP fonctionnent bien pour les quantiles des 25 %,

50 % et 75 %. Le choix de $\mathbf{q}_2^*(u)$ est aussi utile en général. Pour les quantiles extrêmes, les VEP ne justifient pas les difficultés comparativement à l'ED.

Tableau 5.1
AMSE et ABIAS des estimateurs de quantiles sur petits domaines sous le modèle A

		AMSE					ABIAS				
	α	ED	VEL1	VEL2	VEP1	VEP2	ED	VEL1	VEL2	VEP1	VEP2
Distribution des erreurs (i)	5 %	0,470	0,120	0,142	0,121	0,162	0,346	0,022	0,028	0,024	0,032
	25 %	0,219	0,074	0,080	0,074	0,082	0,081	0,006	0,006	0,006	0,006
	50 %	0,187	0,067	0,067	0,067	0,068	0,011	0,005	0,005	0,006	0,006
	75 %	0,218	0,074	0,079	0,074	0,082	0,081	0,007	0,005	0,008	0,006
	95 %	0,470	0,121	0,142	0,123	0,165	0,340	0,024	0,031	0,023	0,033
Distribution des erreurs (ii)	5 %	1,287	0,249	0,786	0,276	1,726	0,352	0,011	0,023	0,011	0,089
	25 %	0,297	0,196	0,217	0,178	0,186	0,084	0,022	0,036	0,021	0,031
	50 %	0,238	0,187	0,182	0,167	0,154	0,011	0,010	0,010	0,010	0,009
	75 %	0,304	0,197	0,233	0,179	0,189	0,081	0,023	0,038	0,023	0,032
	95 %	1,344	0,249	1,919	0,319	2,297	0,349	0,013	0,034	0,015	0,100
Distribution des erreurs (iii)	5 %	0,636	0,165	0,199	0,163	0,234	0,408	0,008	0,013	0,008	0,019
	25 %	0,340	0,132	0,147	0,133	0,152	0,109	0,010	0,007	0,011	0,008
	50 %	0,306	0,128	0,128	0,130	0,132	0,014	0,007	0,007	0,007	0,007
	75 %	0,340	0,133	0,151	0,134	0,156	0,108	0,011	0,009	0,012	0,008
	95 %	0,651	0,168	0,205	0,166	0,243	0,410	0,010	0,016	0,010	0,022
Distribution des erreurs (iv)	5 %	1,225	2,589	0,787	2,679	0,651	0,504	0,220	0,028	0,222	0,071
	25 %	0,574	0,681	0,380	0,652	0,349	0,114	0,174	0,047	0,157	0,017
	50 %	0,488	0,273	0,277	0,241	0,291	0,017	0,010	0,010	0,009	0,010
	75 %	0,571	0,700	0,383	0,670	0,349	0,121	0,183	0,057	0,166	0,012
	95 %	1,251	2,611	0,795	2,709	0,655	0,519	0,207	0,037	0,210	0,082

Tableau 5.2
AMSE et ABIAS des estimateurs de quantiles sur petits domaines sous le modèle B

		AMSE					ABIAS				
	α	ED	VEL1	VEL2	VEP1	VEP2	ED	VEL1	VEL2	VEP1	VEP2
Distribution des erreurs (i)	5 %	0,524	2,998	2,991	0,404	0,439	0,382	1,520	1,502	0,017	0,019
	25 %	0,474	0,182	0,183	0,259	0,262	0,177	0,118	0,123	0,018	0,017
	50 %	0,865	0,907	0,951	0,215	0,219	0,092	0,785	0,791	0,031	0,031
	75 %	1,963	0,985	1,170	0,817	0,825	0,132	0,602	0,616	0,021	0,021
	95 %	7,850	3,083	3,783	9,163	9,193	1,200	1,159	1,185	0,251	0,251
Distribution des erreurs (ii)	5 %	1,227	2,768	3,065	0,492	1,691	0,352	1,430	1,423	0,067	0,143
	25 %	0,562	0,280	0,268	0,331	0,327	0,189	0,087	0,087	0,027	0,024
	50 %	0,976	0,924	0,957	0,287	0,281	0,098	0,728	0,733	0,046	0,046
	75 %	2,119	1,023	1,231	0,817	0,854	0,129	0,557	0,572	0,034	0,034
	95 %	8,392	2,989	4,864	8,405	9,180	1,250	1,140	1,147	0,112	0,119
Distribution des erreurs (iii)	5 %	0,842	2,171	2,207	0,425	0,491	0,500	1,252	1,238	0,013	0,014
	25 %	0,657	0,209	0,209	0,292	0,296	0,176	0,076	0,077	0,010	0,011
	50 %	0,935	0,791	0,805	0,244	0,249	0,082	0,679	0,682	0,026	0,027
	75 %	1,983	0,981	1,086	0,739	0,752	0,131	0,588	0,597	0,024	0,024
	95 %	8,020	2,782	3,251	8,344	8,385	1,219	1,059	1,078	0,144	0,145
Distribution des erreurs (iv)	5 %	1,458	3,913	3,066	2,414	0,814	0,557	1,195	1,172	0,226	0,053
	25 %	0,919	0,460	0,397	0,474	0,472	0,206	0,154	0,137	0,058	0,017
	50 %	1,183	0,913	0,920	0,398	0,416	0,071	0,629	0,640	0,048	0,023
	75 %	2,195	1,223	1,209	1,022	0,902	0,163	0,471	0,511	0,033	0,031
	95 %	8,043	2,954	3,420	7,476	7,639	1,268	0,975	1,042	0,104	0,115

Tableau 5.3
AMSE et ABIAS des estimateurs de quantiles sur petits domaines sous le modèle C

		AMSE					ABIAS				
	α	ED	VEL1	VEL2	VEP1	VEP2	ED	VEL1	VEL2	VEP1	VEP2
Distribution des erreurs (i)	5 %	0,279	1,340	1,258	0,092	0,151	0,267	0,997	0,978	0,051	0,031
	25 %	0,146	0,316	0,263	0,087	0,098	0,068	0,282	0,280	0,035	0,046
	50 %	0,152	0,326	0,403	0,094	0,096	0,011	0,215	0,227	0,019	0,015
	75 %	0,335	0,868	1,368	0,225	0,244	0,029	0,665	0,700	0,043	0,044
	95 %	7,011	0,890	6,818	27,97	27,81	0,291	0,206	0,301	1,398	1,384
Distribution des erreurs (ii)	5 %	1,180	1,181	1,355	0,278	1,776	0,286	0,849	0,836	0,090	0,174
	25 %	0,205	0,461	0,395	0,201	0,208	0,063	0,317	0,327	0,085	0,098
	50 %	0,201	0,450	0,502	0,201	0,191	0,024	0,226	0,235	0,013	0,012
	75 %	0,528	0,943	1,422	0,390	0,422	0,017	0,641	0,681	0,096	0,104
	95 %	7,478	0,890	6,306	23,33	25,01	0,479	0,089	0,107	1,055	1,084
Distribution des erreurs (iii)	5 %	0,438	1,063	1,004	0,157	0,240	0,349	0,826	0,803	0,065	0,034
	25 %	0,299	0,328	0,289	0,158	0,181	0,120	0,158	0,161	0,009	0,020
	50 %	0,305	0,364	0,409	0,174	0,179	0,013	0,151	0,157	0,035	0,029
	75 %	0,428	0,709	1,035	0,275	0,308	0,077	0,499	0,524	0,015	0,017
	95 %	6,718	0,974	4,704	24,79	25,04	0,232	0,321	0,378	1,336	1,325
Distribution des erreurs (iv)	5 %	1,078	4,146	2,303	3,378	0,685	0,444	0,918	0,803	0,409	0,035
	25 %	0,530	0,829	0,531	0,668	0,380	0,107	0,105	0,156	0,147	0,071
	50 %	0,490	0,526	0,565	0,297	0,344	0,021	0,177	0,188	0,054	0,017
	75 %	0,718	1,454	1,412	1,149	0,542	0,076	0,438	0,542	0,061	0,048
	95 %	6,430	2,492	4,002	22,54	21,92	0,462	0,364	0,242	1,258	1,042

Ensuite, nous étudions les estimateurs qui s'appliquent lorsque les valeurs des covariables sont connues pour toutes les unités d'échantillonnage. La simulation comprend EB0, EB1, EB2, MQ0, MQ1 et MQ2 qui désignent respectivement EBLUP/naïve, EBLUP/CD, EBLUP/RKM, M-quantile/naïve, M-quantile/CD et M-quantile/RKM. Nous établissons une taille de population relativement petite $N_i = 500$ pour épargner certains calculs. Le tableau 5.4 présente l'AMSE de ces estimateurs sous les modèles A, B et C avec une distribution des erreurs $N(0, 1)$. Pour économiser de l'espace, nous ne présentons pas les résultats des biais correspondants. Les résultats des simulations montrent que la méthode proposée donne en général des AMSE et des ABIAS (non présentés) inférieurs. Elle fonctionne bien, même pour les quantiles à des niveaux relativement extrêmes.

Pour économiser de l'espace, nous regroupons les résultats de l'AMSE des cinq niveaux de quantiles figurant au tableau 5.5. L'entrée correspondant à A_i est l'AMSE moyenne de l'estimation des quantiles à des niveaux de 5 %, 25 %, 50 %, 75 %, et 95 % lorsque les données sont générées à partir du modèle A avec la distribution des erreurs (i). Nous constatons que, lorsqu'il y a plus de détails sur les covariables, les estimateurs des VEL et des VEP sont bien plus exacts que les résultats des tableaux 5.1 à5.3. Du modèle A au modèle C, la ligne de régression devient moins linéaire. Parallèlement, les estimateurs de quantiles proposés offrent plus d'avantages que les autres estimateurs.

Maintenant, nous évaluons l'estimateur de l'EQM bootstrap proposé dans la section 4. Parce que cette méthode comporte beaucoup de calculs, nous avons restreint la simulation à l'estimateur fondé sur $\hat{F}_i^{(b)}(u)$

avec la fonction de base $\mathbf{q}_1(u) = (1, u)'$ et $B = 100$, $L = 100$. Nous signalons les ratios moyens des EQM estimés et les EQM simulés dans tous les petits domaines. Plus le ratio s'approche de un, plus l'estimation de l'EQM bootstrap est exacte. Dans le tableau 5.6, nous pouvons voir que les ratios moyens tournent autour de un dans la majorité des situations, sauf pour la distribution des erreurs (iv) aux niveaux extrêmes des quantiles. Nous en concluons que l'estimateur de l'EQM bootstrap est généralement satisfaisant.

Tableau 5.4
AMSE de 10 estimateurs de quantiles lorsque toutes les valeurs de covariance sont connues avec une distribution des erreurs *N*(0, 1)

	α	EB0	EB1	EB2	MQ0	MQ1	MQ2	VEL1	VEL2	VEP1	VEP2
Modèle A	5 %	0,477	0,123	0,501	0,536	0,127	0,499	0,128	0,146	0,078	0,110
	25 %	0,139	0,073	0,154	0,198	0,074	0,154	0,073	0,078	0,065	0,073
	50 %	0,061	0,066	0,124	0,119	0,066	0,124	0,066	0,066	0,064	0,064
	75 %	0,145	0,074	0,149	0,204	0,074	0,149	0,074	0,080	0,066	0,073
	95 %	0,491	0,125	0,394	0,552	0,129	0,395	0,126	0,146	0,079	0,113
Modèle B	5 %	1,270	2,500	0,928	1,682	2,575	0,946	2,965	2,949	0,079	0,110
	25 %	0,351	0,152	0,239	0,262	0,149	0,239	0,193	0,193	0,069	0,069
	50 %	0,834	0,723	0,285	0,631	0,722	0,284	0,899	0,944	0,071	0,073
	75 %	0,314	0,634	0,532	0,257	0,644	0,530	0,986	1,160	0,082	0,084
	95 %	3,710	2,095	3,690	4,209	2,059	3,685	3,235	3,900	0,154	0,156
Modèle C	5 %	0,346	0,830	0,415	0,708	0,307	0,351	1,087	1,028	0,075	0,130
	25 %	0,345	0,173	0,169	0,388	0,110	0,154	0,263	0,224	0,066	0,075
	50 %	0,340	0,170	0,142	0,207	0,150	0,136	0,291	0,349	0,065	0,067
	75 %	0,288	0,577	0,211	0,191	0,376	0,227	0,731	1,088	0,068	0,087
	95 %	2,578	11,47	8,087	5,194	14,64	11,96	0,868	4,215	0,148	0,156

Tableau 5.5
AMSE moyenne pour 5 quantiles lorsque toutes les valeurs des covariables sont connues

Modèle	EB0	EB1	EB2	MQ0	MQ1	MQ2	VEL1	VEL2	VEP1	VEP2
A_{i}	0,263	0,092	0,264	0,322	0,094	0,264	0,093	0,103	0,070	0,087
A_{ii}	0,810	1,379	1,822	0,810	1,381	1,796	0,217	0,370	0,203	0,744
A_{iii}	0,754	0,183	0,408	0,819	0,183	0,407	0,149	0,168	0,135	0,168
A_{iv}	0,687	0,186	0,399	0,746	0,188	0,399	0,281	0,196	0,256	0,164
B_{i}	1,296	1,221	1,135	1,408	1,230	1,138	1,832	1,829	0,091	0,098
B_{ii}	1,442	1,714	2,348	1,496	1,718	2,343	1,596	1,812	0,230	0,504
B_{iii}	1,270	1,081	1,357	1,348	1,088	1,351	1,399	1,521	0,163	0,179
B_{iv}	1,346	1,177	1,315	1,436	1,183	1,317	1,565	1,701	0,205	0,166
C_{i}	0,799	2,645	1,805	1,339	3,117	2,566	0,648	1,381	0,084	0,103
C_{ii}	1,441	3,439	3,368	2,232	3,967	3,898	0,725	1,168	0,241	0,377
C_{iii}	1,141	2,516	1,898	1,834	2,937	2,572	0,595	1,133	0,153	0,186
C_{iv}	1,149	2,499	1,909	1,821	2,933	2,639	0,767	1,176	0,280	0,179

Tableau 5.6
Ratios moyens des EQM bootstrap et des EQM simulés

α	A_{i}	A_{ii}	A_{iii}	A_{iv}	B_{i}	B_{ii}	B_{iii}	B_{iv}	C_{i}	C_{ii}	C_{iii}	C_{iv}
5 %	1,01	1,03	1,05	0,36	1,05	0,98	1,01	0,39	0,99	1,19	1,10	0,27
25 %	1,00	0,99	1,05	0,74	1,03	0,99	0,95	1,03	1,03	0,97	0,99	0,73
50 %	1,06	1,04	0,97	1,10	1,01	1,03	0,96	0,99	1,09	0,96	0,97	1,03
75 %	1,01	0,99	1,06	0,76	1,10	1,01	0,98	0,90	1,06	0,96	1,03	0,52
95 %	1,04	1,20	1,10	0,33	0,89	1,02	1,13	1,02	0,95	1,37	1,13	0,69

6 Application empirique

Nous illustrons maintenant les estimateurs proposés à partir de l'ensemble de données de l'*Enquête sur la dynamique du travail et du revenu* (EDTR) fourni par Statistique Canada (2014) et téléchargé du centre de données de la bibliothèque de l'Université de la Colombie-Britannique. Les données contiennent 147 variables et 47 705 unités d'échantillonnage. Nous remercions Statistique Canada de rendre l'ensemble de données disponible, mais nous ne nous attardons pas à l'objectif initial de l'enquête ici. Nous l'utilisons plutôt comme une superpopulation afin d'étudier l'efficacité de l'estimateur proposé des quantiles sur petits domaines.

Dans cette étude, nous avons isolé 9 des 147 variables. Il s'agit des variables `ttin`, `gender`, `spouse`, `edu`, `âge`, `yrx`, `tweek`, `jobdur` et `tpaid`, qui désignent respectivement : le revenu total, le sexe, si la personne vit avec son conjoint, le plus haut niveau de scolarité, l'âge, les années d'expérience, le nombre de semaines d'emploi, le niveau de scolarité, la durée de l'emploi actuel (en mois) et le nombre total d'heures rémunérées pour cet emploi. Après avoir retiré les unités qui renfermaient des valeurs manquantes dans ces 9 variables et celles où $\texttt{ttin} \leq 0$, nous avons obtenu un ensemble de données contenant 28 302 unités d'échantillonnage. Les moyennes de puissances des covariables au niveau de la population sont toujours calculées à partir de toutes les observations disponibles. Nous avons créé 28 sous-populations (soit les petits domaines) étiquetées comme $4(k-1)+i$, $k = 1, 2, \ldots, 7$, $i = 1, 2, 3, 4$ à partir des combinaisons de sexe-conjoint-scolarité. Ici, k désigne le niveau de scolarité et $i = 1, 2, 3, 4$ désignent respectivement un homme vivant avec un conjoint ou une conjointe, une femme vivant avec un conjoint ou une conjointe, un homme ne vivant pas avec un conjoint ou une conjointe et une femme ne vivant pas avec un conjoint ou une conjointe. Les niveaux de scolarité sont décrits comme suit.

k	Le plus haut niveau de scolarité
1	Pas plus de dix années d'études primaires et secondaires
2	De 11 à 13 années d'études primaires et secondaires (sans diplôme)
3	Diplôme d'études secondaires
4	Quelques études postsecondaires universitaires ou non universitaires, sans certificat
5	Certificat d'études postsecondaires universitaires ou non universitaires inférieur au baccalauréat
6	Baccalauréat
7	Certificat d'études universitaires supérieur au baccalauréat

Nous avons considéré le $\log(\texttt{ttin})$ comme la variable de réponse et ajusté les régressions non paramétriques linéaires et additives pour cinq autres variables. Selon toutes les données, le R-carré ajusté pour l'ajustement non paramétrique est 0,482, soit bien supérieur au 0,370 obtenu en ajustant la régression linéaire. Cela donne à penser qu'un modèle mixte non paramétrique est un bon choix. La figure 6.1 montre les courbes ajustées du log(`ttin`) pour ces deux covariables. Par ailleurs, le R-carré augmente pour s'établir à 0,483, même si le modèle comprend seulement les covariables `âge` et `tpaid`, ainsi qu'un effet aléatoire. Ces analyses exploratoires nous incitent à n'utiliser que ces deux covariables dans notre simulation. Nous avons effectué la simulation avec des tailles d'échantillons $n = 200; 500$ et 1 000. Pour que les proportions des échantillons dans les petits domaines s'approchent de leur taille, nous supposons que $n_i = a_i + 2$, $i = 1, \ldots, 28$, où a_i est généré à partir de la distribution multinomiale où $p_i = N_i / N$.

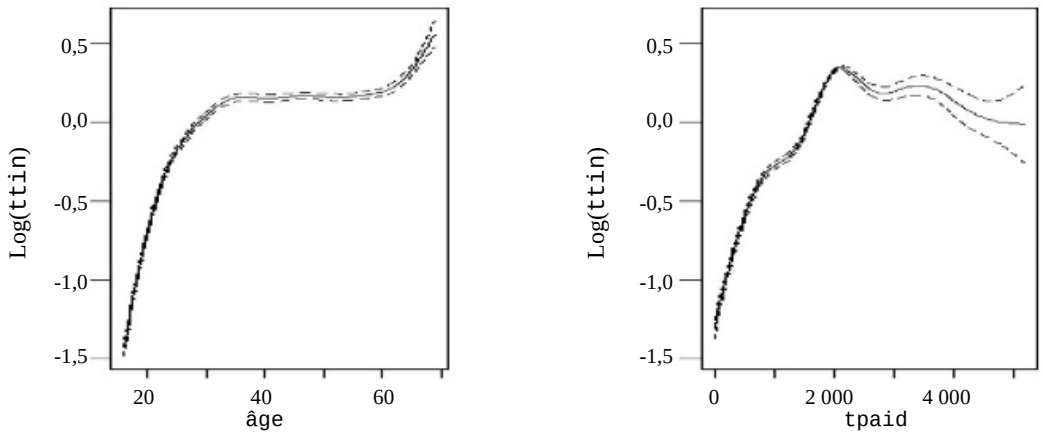

Figure 6.1 Courbes ajustées du log(`ttin`) pour `âge` et `tpaid`.

L'AMSE simulée de 10 estimateurs à partir de 1 000 répétitions est présentée au tableau 6.1. Nous remarquons d'abord que nos estimateurs des VEP donnent de meilleurs résultats que les autres estimateurs, en général, ce qui fait ressortir l'avantage qu'offre notre technique d'estimation sur petits domaines fondée sur un MRD non paramétrique. La VEP1, comparée à VEP2, obtient l'AMSE la plus basse pour les quantiles de 5 %, 25 %, et 50 %, mais une AMSE légèrement plus élevée pour les quantiles de 75 % et 95 %, ce qui démontre que l'hétéroscédasticité des données n'est pas grave. Malgré les estimateurs de VEP, nous constatons que les estimateurs de VEL obtiennent de meilleurs résultats que d'autres estimateurs pour le quantile de 5 % et qu'ils ont un rendement semblable pour les autres quantiles. Quand on augmente la taille de l'échantillon, l'AMSE de tous les estimateurs diminue. De toute évidence, il est difficile d'estimer le quantile de 5 % avec une grande précision parce que les données sont asymétriques vers la gauche; il y a donc peu d'observations pour estimer les quantiles inférieurs. Fait intéressant, VEL1 n'est pas autant touchée par l'asymétrie. Nous sommes d'avis que l'étape (3.7) du lissage par la méthode du noyau est utile

ici. Sans l'étape du lissage, VEL1 obtiendrait un bien moins bon résultat. Les simulations non présentées démontrent que l'ABIAS de tous les estimateurs diminue en général à mesure que la taille de l'échantillon augmente, et cela est surtout évident pour l'ED.

Pour vérifier le rendement du premier estimateur proposé, nous utilisons uniquement les données moyennes sur les covariables. Dans la figure 6.2, nous illustrons les quantiles de 2,5 %, 50 %, et 97,5 % de 1 000 estimations médianes sur petits domaines à l'aide de l'ED, de VEL1, VEL2, VEP1, VEP2, lorsque la taille de l'échantillon $n = 200$ et les véritables médianes sont indiquées par des points. L'axe des Y représente le revenu total et l'axe des X, le niveau de scolarité. On peut voir que les barres correspondant à VEP2 sont les plus courtes pour la plupart des petits domaines.

Le tableau 6.2 montre les estimations des EQM bootstrap ainsi que les ratios moyens des EQM bootstrap et simulées pour les estimateurs médians des petits domaines, où $\hat{F}_i^{(a)}(u)$ et $\hat{F}_i^{(b)}(u)$ et la taille de l'échantillon $n = 200$. Le nombre de répétitions de simulations est de 500 et la fonction de base $\mathbf{q}_1(u) = (1, u)'$ et $B = 100$, $L = 100$. Nous pouvons voir que l'estimateur $\hat{F}_i^{(a)}(u)$ a une EQM plus élevée que $\hat{F}_i^{(b)}(u)$, et que la plupart des ratios moyens s'approchent de un.

Tableau 6.1
AMSE des estimateurs de quantiles sur petits domaines à partir de données réelles

	α	EB0	EB1	EB2	MQ0	MQ1	MQ2	VEL1	VEL2	VEP1	VEP2
$n = 200$	5 %	0,784	0,769	0,901	0,714	0,763	0,885	0,245	0,421	0,242	0,336
	25 %	0,107	0,256	0,488	0,102	0,261	0,467	0,115	0,131	0,097	0,152
	50 %	0,080	0,119	0,236	0,064	0,116	0,223	0,076	0,095	0,056	0,102
	75 %	0,122	0,100	0,142	0,085	0,102	0,138	0,085	0,076	0,069	0,068
	95 %	0,233	0,190	0,280	0,141	0,138	0,266	0,217	0,179	0,117	0,096
$n = 500$	5 %	0,793	0,603	0,826	0,710	0,579	0,805	0,173	0,345	0,210	0,301
	25 %	0,072	0,110	0,207	0,076	0,119	0,197	0,069	0,127	0,063	0,091
	50 %	0,049	0,050	0,074	0,036	0,050	0,072	0,053	0,076	0,040	0,043
	75 %	0,108	0,044	0,060	0,055	0,046	0,058	0,054	0,047	0,046	0,043
	95 %	0,257	0,128	0,152	0,109	0,058	0,148	0,138	0,125	0,086	0,077
$n = 1\,000$	5 %	0,792	0,397	0,542	0,706	0,377	0,528	0,078	0,130	0,095	0,144
	25 %	0,054	0,056	0,098	0,066	0,067	0,095	0,041	0,043	0,038	0,056
	50 %	0,034	0,026	0,032	0,027	0,026	0,031	0,019	0,028	0,018	0,024
	75 %	0,102	0,024	0,030	0,043	0,026	0,030	0,037	0,033	0,019	0,023
	95 %	0,270	0,088	0,090	0,095	0,114	0,090	0,074	0,067	0,053	0,057

Tableau 6.2
Estimations des EQM bootstrap et des ratios moyens des EQM estimées et simulées

	$\hat{F}_i^{(a)}(u)$					$\hat{F}_i^{(b)}(u)$				
	5 %	25 %	50 %	75 %	95 %	5 %	25 %	50 %	75 %	95 %
EQM	0,542	0,196	0,117	0,098	0,165	0,204	0,093	0,068	0,062	0,102
Ratio	0,843	0,959	1,014	0,988	0,871	0,969	0,994	1,003	0,996	0,975

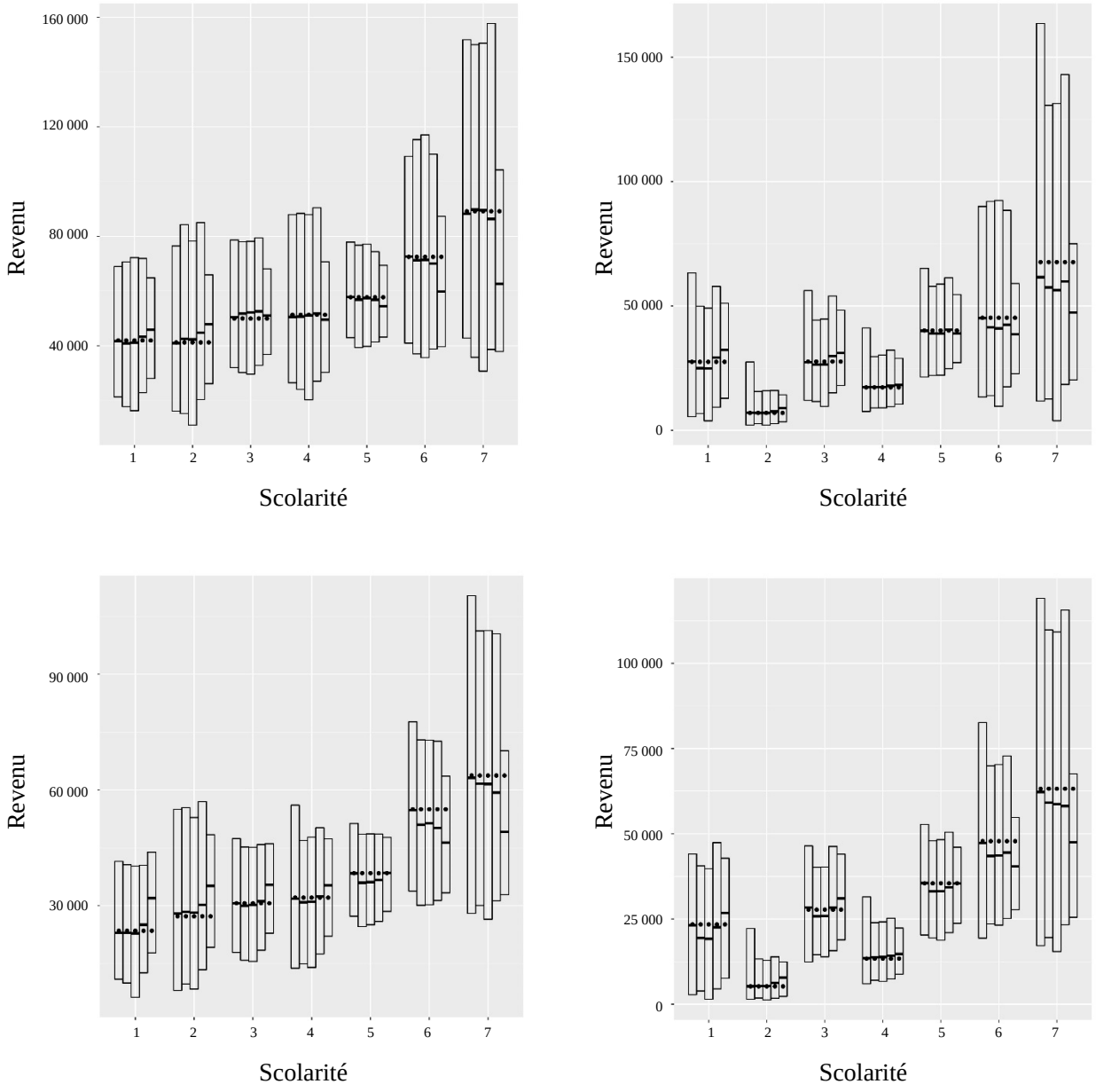

Figure 6.2 Les lignes du bas, du milieu et du haut de chaque barre désignent les quantiles de 2,5 %, 50 % et 97,5 % de 1 000 estimations sur petits domaines du revenu total. Le point dans chaque barre désigne la médiane véritable du petit domaine. Cinq barres dans chaque grappe sont formées par les estimations de l'ED, de VEL1, VEL2, VEP1 et VEP2. Dans les deux graphiques du haut : homme vivant (à gauche) et ne vivant pas (à droite) avec un conjoint ou une conjointe; dans les deux graphiques du bas : femme vivant (à gauche) et ne vivant pas (à droite) avec un conjoint ou une conjointe. Sept grappes dans chaque graphique correspondent à sept niveaux de scolarité.

7 Conclusion

Nous avons étudié l'estimation de quantiles sur petits domaines à l'aide du modèle de régression non paramétrique à erreurs emboîtées et d'une hypothèse de MRD semi-paramétrique pour les distributions des erreurs. Nous avons proposé deux estimateurs de quantiles à l'aide de l'approche de P-splines et de la vraisemblance empirique. Les résultats des simulations démontrent que les estimateurs proposés sont robustes et que leur efficacité est respectable tant pour des fonctions de régression linéaire que non paramétrique des quantiles intermédiaires. En principe, l'approche proposée peut être élargie aux modèles de régression non paramétriques à plusieurs covariables, même si elle donne bien plus de paramètres à estimer. Ce problème fera l'objet de travaux ultérieurs.

Remerciements

Nous remercions les professeurs Simon Wood, Matt Wand, Mahmoud Torabi et Song Cai pour leurs propositions utiles sur les routines R employées dans le présent document. Ces travaux ont obtenu le soutien de la Fondation nationale des sciences naturelles de la Chine (n° 11661067), de la Fondation des sciences naturelles de la province de Qinghai (n° 2015-ZJ-717,2019-ZJ-920), du programme de talents « Western Light » de l'Académie des sciences de Chine (2017) et du financement par l'intermédiaire de l'Institut canadien des sciences statistiques.

Bibliographie

Anderson, J.A. (1979). Multivariate logistic compounds. *Biometrika*, 66, 17-26.

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 80, 28-36.

Boor, C.D. (2001). *A Practical Guide to Splines.* New York: Springer.

Chambers, R., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chambers, R., et Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

Chaudhuri, S., et Ghosh, M. (2011). Empirical likelihood for small area estimation. *Biometrika*, 98, 473-480.

Chen, J., et Liu, Y. (2013). Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41, 1669-1692.

Chen, J., et Liu, Y. (2018). Small area quantile estimation. *Revue Internationale de Statistique*. Sur papier. arXiv:1705.10063.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Jiang, J. (2010). *Large Sample Techniques for Statistics.* New York: Springer.

Jiang, J., et Lahiri, P. (2006a). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Jiang, J., et Lahiri, P. (2006b). Mixed model prediction and small area estimation. *Test*, 15, 1-96.

Jiang, J., Ngueyen, T. et Rao, J.S. (2010). Méthode de l'enclos pour l'estimation non paramétrique sur petits domaines. *Techniques d'enquête*, 36, 1, 3-12. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2010001/article/11244-fra.pdf.

Kezioua, A., et Leoni-Aubina, S. (2008). On empirical likelihood for semiparametric two-sample density ratio models. *Journal of Statistical Planning and Inference*, 138, 915-928.

Lahiri, P.S., et Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.

Lin, X., et Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society, Series B*, 61, 381-400.

Molina, I., and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38, 369-385.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. et Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: B*, 70, 265-286.

Owen, A.B. (2001). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Pfeffermann, D. (2002). Small area estimation-New developments in small area estimation. *Revue Internationale de Statistique*, 70, 125-143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28, 40-68.

Pratesi, M., Ranalli, M.G. et Salvati, N. (2008). Semiparametric M-quantile regression for estimation for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. *Environmetrics*, 19, 687-701.

Qin, J., et Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609-618.

Rao, J.N.K. (2003). *Small Area Estimation.* New York: John Wiley & Sons, Inc.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation, 2nd Edition.* New York: John Wiley & Sons, Inc.

Rao, J.N.K., Sinha, S.K. et Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *The Canadian Journal of Statistics*, 42, 126-141.

Ruppert, D., Wand, M.P. et Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Salvati, N., Tzavidis, N. et Pratesi, M. (2012). Small area estimation via M-quantile geographically weighted regression. *Test*, 21, 1-28.

Sinha, S.K., et Rao, J.N.K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37(3), 381-399.

Sperlich, S., et José Lombardía, M. (2010). Local polynomial inference for small area statistics: Estimation, validation and prediction. *Journal of Non-parametric Statistics*, 22, 633-648.

Statistique Canada (2014). Survey of labour and income dynamics, 2011. Accessible à l'adresse : http://tinyurl.com/y2ys2zzs.

Torabi, M., et Shokoohi, F. (2015). Non-parametric generalized linear mixed models in small area estimation. *The Canadian Journal of Statistics*, 43, 82-96.

Tzavidis, N., et Chambers, R. (2005). Bias adjusted estimation for small areas with M-quantile models. *Statistics in Transition*, 7, 707-713.

Tzavidis, N., Salvati, N. et Pratesi, M. (2008). M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17, 393-411.

Tzavidis, N., Marchetti, S. et Chambers, R. (2010). Robust prediction of small area means and quantiles. *Australian and New Zealand Journal of Statistics*, 52, 167-186.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton, Floride: Chapman & Hall/CRC.