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

Techniques d'enquête

Imputation par régression quantile semi-paramétrique pour une enquête complexe avec application au *Conservation Effects Assessment Project*

par Emily Berg et Cindy Yu

Date de diffusion : le 27 juin 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, juin 2019
Vol. 45, nº 2, pp. 265-288
Statistique Canada, nº 12-001-X au catalogue

Imputation par régression quantile semi-paramétrique pour une enquête complexe avec application au *Conservation Effects Assessment Project*

Emily Berg et Cindy Yu[1]

Résumé

L'élaboration de procédures d'imputation appropriées pour les données ayant des valeurs extrêmes ou des relations non linéaires avec des covariables constitue un défi important dans les enquêtes à grande échelle. Nous élaborons une procédure d'imputation pour les enquêtes complexes fondée sur la régression quantile semi-paramétrique. Nous appliquons cette méthode au *Conservation Effects Assessment Project* (CEAP), une enquête à grande échelle qui recueille des données utilisées pour quantifier la perte de sol provenant des champs de culture. Dans la procédure d'imputation, nous générons d'abord des valeurs imputées à partir d'un modèle semi-paramétrique pour les quantiles de la distribution conditionnelle de la réponse pour une covariable donnée. Ensuite, nous évaluons les paramètres d'intérêt à l'aide de la méthode généralisée des moments (MGM). Nous dérivons la distribution asymptotique des estimateurs MGM pour une classe générale de plans d'enquête complexes. Dans les simulations destinées à représenter les données du CEAP, nous évaluons les estimateurs de variance en fonction de la distribution asymptotique et comparons la méthode d'imputation par régression quantile (IRQ) semi-paramétrique à des solutions de rechange entièrement paramétriques et non paramétriques. La procédure de l'IRQ est plus efficace que les solutions de rechange non paramétriques et entièrement paramétriques, et les couvertures empiriques des intervalles de confiance se situent à moins de 1 % du niveau nominal de 95 %. Une application à l'estimation de l'érosion moyenne indique que l'IRQ pourrait être une option viable pour le CEAP.

Mots-clés : Plan d'échantillonnage informatif; B-spline; érosion.

1 Introduction

Les données manquantes ont d'importantes répercussions sur l'analyse des données d'enquête. Il peut arriver qu'il y ait des données manquantes si les unités échantillonnées refusent de participer à l'enquête, sont difficiles à localiser, ne répondent pas aux questions sensibles ou abandonnent les enquêtes longitudinales. Si les valeurs manquantes sont liées à la variable d'intérêt, une analyse des données complètes sans modification pour les valeurs manquantes est biaisée. La pondération et l'imputation sont deux grandes catégories d'ajustements de données manquantes.

Deux types d'ajustements de pondération sont la pondération par calage (D'Arrigo et Skinner, 2010; Kott, 2006) et la pondération fondée sur les scores de propension (Kim et Riddles, 2012). Dans la pondération par calage, les poids des répondants sont ajustés de façon à ce que la somme pondérée d'une variable auxiliaire pour les répondants soit égale à la moyenne correspondante pour l'échantillon complet ou à la moyenne de la population. Dans la pondération fondée sur les scores de propension, le poids d'échantillonnage est multiplié par l'inverse d'une probabilité estimée de réponse.

L'imputation complète l'ensemble de données en remplaçant les variables à réponse manquante par des valeurs imputées. Elle peut simplifier les analyses en cas de non-réponse partielle et améliorer l'uniformité

1. Emily Berg, Department of Statistics, Iowa State University. Courriel : emilyb@iastate.edu; Cindy Yu, Department of Statistics, Iowa State University.

des résultats entre les utilisateurs. Nous envisageons l'imputation d'une réponse y, qui pourrait être manquante, à l'aide d'une variable auxiliaire x observée pour l'échantillon complet. Afin de permettre une certaine souplesse dans les hypothèses du modèle, nous utilisons un modèle de régression quantile semi-paramétrique pour décrire la relation entre x et y.

Il existe un vaste éventail de procédures d'imputation (Kim and Shao, 2013). L'imputation fractionnaire paramétrique (Kim, 2011) et l'imputation multiple paramétrique (Rubin, 2004) génèrent des valeurs imputées à partir d'une estimation d'un modèle entièrement paramétrique pour la distribution conditionnelle de la réponse pour des covariables données. L'imputation hot deck (c'est-à-dire Andridge et Little, 2010), en revanche, inclue une catégorie de procédures non paramétriques dans laquelle les valeurs imputées sont sélectionnées parmi les répondants. Dans certaines procédures hot deck, les poids sont attribués selon une mesure de proximité, définie par des classes d'imputation (Brick et Kalton, 1996) ou une mesure (Rubin, 2004; Little, 1988), comme la distance entre les noyaux (Wang et Chen, 2009). L'imputation non paramétrique est plus robuste à l'erreur de spécification du modèle que les méthodes entièrement paramétriques. Toutefois, les estimateurs fondés sur les procédures non paramétriques peuvent être peu efficaces dans les petits échantillons. L'imputation par régression quantile (IRQ) semi-paramétrique est un compromis entre les procédures d'imputation non paramétriques et entièrement paramétriques. Dans l'IRQ, les valeurs imputées pour une seule valeur manquante sont les quantiles estimés de la distribution de l'observation manquante conditionnelle à une fonction des variables auxiliaires. Comme un modèle semi-paramétrique pour la fonction quantile est utilisé, l'IRQ est robuste à l'erreur de spécification du modèle, et comme les valeurs sont imputées à partir de quantiles estimés, l'IRQ résiste aux valeurs extrêmes. Chen et Yu (2016) élaborent l'IRQ pour un échantillonnage aléatoire simple d'une population infinie. Nous étendons le modèle de Chen et Yu (2016) pour permettre des probabilités de sélection inégales.

De nombreuses procédures d'imputation reposent sur une hypothèse de réponse manquant au hasard (RMH) (Rubin, 1976). Une hypothèse courante est que la variable de réponse (y, qui peut être manquante) est conditionnellement indépendante de l'indicateur de valeur manquante (« 1 » si une réponse est fournie et « 0 » autrement) pour les données observées. Une application directe de cette définition de la RMH à une enquête complexe précise l'indépendance de la variable de réponse et de la variable indicatrice manquante conditionnellement à la variable auxiliaire et aux indicateurs d'inclusion de l'échantillon (Little, 1982; Pfeffermann, 2011). Berg, Kim et Skinner (2016), nomme échantillon manquant au hasard la RMH définie conditionnellement aux indicateurs d'inclusion de l'échantillon. Selon une autre hypothèse, celle de la population manquant au hasard (Berg et coll., 2016), la variable de réponse est conditionnellement indépendante de l'indicateur de valeur manquante pour une variable auxiliaire donnée dans la superpopulation, de manière inconditionnelle aux indicateurs d'inclusion de l'échantillon. Berg et coll. (2016) montrent que ces deux hypothèses ne sont pas équivalentes. Nous discutons précisément de ces concepts de RMH à la section 2, et élaborons notre procédure pour qu'elle soit suffisamment souple pour s'adapter à l'une ou l'autre des conditions.

Notre intérêt pour la régression quantique semi-paramétrique pour une enquête complexe est motivé en partie par le *Conservation Effects Assessment Project* (CEAP), une enquête complexe visant à quantifier la perte de sol et d'éléments nutritifs provenant des champs de culture. Comme les distributions des variables de réponses sont fortement asymétriques et contiennent des valeurs extrêmes, la spécification d'un modèle d'imputation entièrement paramétrique adéquat est difficile, et les procédures d'imputation hot deck peuvent présenter de grandes variances. Nous étudions l'utilisation de l'IRQ pour régler ces problèmes relatifs à l'imputation pour le CEAP.

Nous démontrons la validité théorique et l'applicabilité de l'imputation par régression quantile semi-paramétrique dans le contexte d'une enquête complexe. La section 2 et la section 3, respectivement, présentent l'algorithme d'imputation et les propriétés asymptotiques. La section 4 et la section 5 démontrent les propriétés de l'IRQ par l'application au CEAP et les simulations, respectivement. La section 6 se termine par un résumé et une discussion des avenues de recherche futures.

2 Imputation par régression quantile pour les données d'enquête complexes

Imaginons un cadre conceptuel dans lequel les échantillons sont tirés d'une population finie générée à partir d'un modèle de superpopulation (Fuller, 2009b, chapitre 6). Supposons que x_i et y_i ont une distribution conjointe $f(x_i, y_i)$ dans la surpopulation. Nous définissons la distribution conditionnelle de y_i compte tenu de x_i par la fonction quantile conditionnelle. Supposons que $q_\tau(x_i)$ indique le τ^{e} quantile de la distribution conditionnelle de y_i, compte tenu de x_i, dans la superpopulation, où $q_\tau(x_i)$ est défini par

$$P(y_i \leq q_\tau(x_i) \,|\, x_i) = \tau. \tag{2.1}$$

Nous précisons un modèle pour les quantiles parce que les modèles de régression quantile peuvent décrire une grande variété de distributions, comme l'illustre la figure 2.1. Le panneau de gauche de la figure 2.1 illustre un modèle de régression quantile linéaire dans lequel chaque fonction quantile conditionnelle est représentée avec une ordonnée à l'origine différente et une pente différente. L'utilisation d'une pente différente permet de décrire des données avec des variances non constantes. Le panneau de droite de la figure 2.1 illustre une généralisation à la régression quantile semi-paramétrique, où le τ^{e} quantile de la distribution conditionnelle de y_i est représenté comme une fonction continue de x_i. Dans la procédure d'imputation, nous supposons que $q_\tau(\cdot)$ est une fonction avec $p+1$ dérivées continues. Nous obtenons la valeur approximative de $q_\tau(x_i)$ avec une fonction B-spline (de Boor, 2001; Chen et Yu, 2016; Yoshida, 2013; Hastie, Tibshirani et Friedman, 2009), comme nous l'expliquons plus en détail à la section 2.2. Pour permettre l'utilisation de la fonction B-spline, nous supposons que x_i a un support compact, mais n'avons pas besoin d'autres hypothèses de distribution pour x_i.

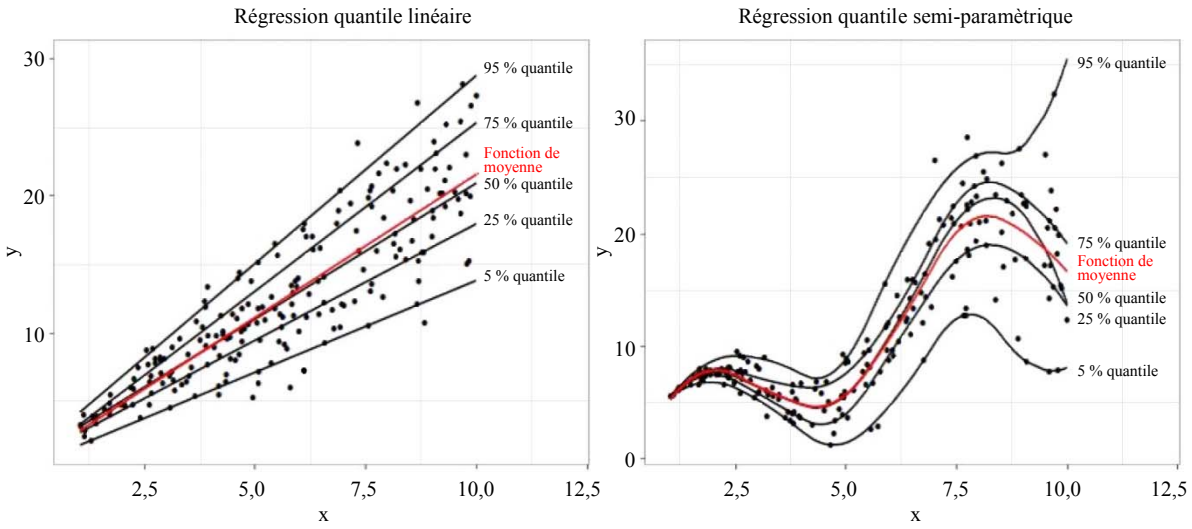

Figure 2.1 Illustration d'une régression quantile linéaire (à gauche) et d'une régression quantile semi-paramétrique (à droite).

Nous prenons en considération l'estimation des paramètres définis en fonction du modèle de superpopulation qui relie y_i à x_i, plutôt que des paramètres de population finie. Le véritable paramètre d'intérêt, $\boldsymbol{\theta}_o$, est un vecteur $d-$dimensionnel qui permet d'obtenir :

$$E\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right] = \mathbf{0}, \tag{2.2}$$

où $\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)$ est une fonction $r-$dimensionnelle à deux dérivées continues, et $r \geq d$. L'opérateur d'espérance $E[\cdot]$ indique des espérances par rapport au modèle de superpopulation. Veuillez prendre note que $E\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right] = E\left[E_{y|x}\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right]\right]$, où

$$\begin{aligned} E_{y|x}\left[\mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right)\right] &= \int_{-\infty}^{\infty} \mathbf{g}\left(y_i,\, x_i;\, \boldsymbol{\theta}_o\right) f_{y|x}\left(y_i \mid x_i\right) dy_i \\ &= \int_0^1 \mathbf{g}\left(F_{y|x}^{-1}(\tau),\, x_i;\, \boldsymbol{\theta}_o\right) \frac{f_{y|x}\left(F_{y|x}^{-1}(\tau) \mid x_i\right)}{f_{y|x}\left(F_{y|x}^{-1}(\tau) \mid x_i\right)} d\tau = \int_0^1 \mathbf{g}\left(q_\tau\left(x_i\right),\, x_i;\, \boldsymbol{\theta}_o\right) d\tau, \end{aligned} \tag{2.3}$$

et $F_{y|x}\left(y_i \mid x_i\right)$ et $f_{y|x}\left(y_i \mid x_i\right)$, respectivement, dénotent la fonction de distribution cumulative (fdc) et la fonction de densité de probabilité (fdp) de la distribution conditionnelle de y_i compte tenu de x_i. La deuxième égalité dans (2.3) découle de la transformation de l'intégrale de la probabilité et d'un changement de variables de y_i τ de distribution uniforme avec fdp $f(\tau) = I[\tau \in (0,1)]$, ou $I[\cdot]$ représente la variable indicatrice qui prend la valeur « 1 » si l'argument est vrai et « 0 » dans le cas contraire. La relation définie par la troisième égalité dans (2.3) joue un rôle important dans la procédure d'imputation. Pour chaque y_i, manquant nous générons J valeurs imputées définies par $\left\{\hat{q}_{\tau_1}\left(x_i\right), \ldots, \hat{q}_{\tau_J}\left(x_i\right)\right\}$, où $\hat{q}_{\tau_j}\left(x_i\right)$ estime la valeur de $q_{\tau_j}\left(x_i\right)$, et $\tau_1, \ldots, \tau_J$ forme une grille fine sur l'intervalle $[0, 1]$. Nous estimons ensuite la valeur

de $E_{y|x}[\mathbf{g}(y_i, x_i; \boldsymbol{\theta}_o)]$ en donnant une valeur approximative à l'intégrale dans la dernière expression de (2.3) avec une moyenne des valeurs J imputées.

La procédure d'imputation comporte deux étapes principales. Nous générons d'abord les valeurs imputées, en estimant $q_\tau(x_i)$ à l'aide d'une combinaison linéaire de fonctions de base B-spline. Nous estimons ensuite la valeur de $\boldsymbol{\theta}_o$ en utilisant la méthode généralisée des moments (MGM), en remplaçant la valeur y_i manquante par la valeur estimée de $E_{y|x}[\mathbf{g}(y_i, x_i; \boldsymbol{\theta}_o)]$ en fonction des valeurs estimées et de la relation (2.3). Pour officialiser la procédure, nous avons besoin d'hypothèses précises sur la conception et le mécanisme de réponse, que nous précisons à la section 2.1. La section 2.2 explique l'estimation de la fonction quantile et la section 2.3 décrit la méthode généralisée des moments (MGM). Les logiciels de mise en œuvre des procédures sont disponibles auprès des auteurs.

2.1 Hypothèses sur la conception et le mécanisme de réponse

Supposons que I_i est l'indicateur d'appartenance à l'échantillon, défini par $I_i = 1$ si l'unité i est sélectionnée. Supposons que π_i et π_{ij} représentent les probabilités d'inclusion de premier ordre et de deuxième ordre, respectivement, définies par

$$\left[\pi_i, \pi_{ij}\right] = \left[P(I_i = 1 \mid y_i, x_i), P\left(I_i = 1, I_j = 1 \mid y_i, x_i, y_j, x_j\right)\right]. \tag{2.4}$$

La dépendance de π_i sur y_i dans (2.4) représente une corrélation possible entre y_i et π_i qui peut rendre le plan d'échantillonnage informatif pour le modèle de régression quantile (2.1). Nous indiquons l'échantillon sélectionné par A, où $A = \{i : I_i = 1\}$.

Nous supposons que x_i est observé pour toutes les i dans A, tandis que y_i peut être manquant. Supposons que δ_i est l'indicateur de réponse, défini par $\delta_i = 1$ si y_i est observé et $\delta_i = 0$ si y_i est manquant. Supposons que $\delta_i \sim \text{Bernoulli}(p_i)$, où la probabilité de réponse p_i est définie ainsi :

$$p_i = P(\delta_i = 1 \mid y_i, x_i, I_i). \tag{2.5}$$

Pour définir une procédure d'imputation approximativement sans biais, nous avons besoin d'une hypothèse sur la relation entre δ_i et y_i. Une approche commune dans l'analyse des données manquantes consiste à présumer que la variable de réponse, y_i, est indépendante de l'indicateur de valeur manquante, δ_i, conditionnellement aux valeurs observées (Little, 1982; Pfeffermann, 2011). Cette hypothèse est une interprétation largement utilisée de la RMH définie dans Rubin (1976), et clarifiée dans Mealli et Rubin (2015). Pour une enquête complexe, la relation entre les probabilités d'inclusion, les probabilités de réponse et y peut être complexe si les indicateurs de réponse et les indicateurs d'inclusion de l'échantillon dépendent d'une variable qui n'est pas incluse dans le modèle d'imputation.

Nous suivons l'approche de Berg et coll. (2016) et examinons deux hypothèses relatives à la relation entre δ_i and y_i. Nous définissons un échantillon manquant au hasard (EMH) comme suit :

$$P(\delta_i = 1 \mid x_i, y_i, I_i) = P(\delta_i = 1 \mid x_i, I_i). \tag{2.6}$$

En revanche, nous définissons une population manquant au hasard (PMH) comme suit :

$$P(\delta_i = 1 | x_i, y_i) = P(\delta_i = 1 | x_i). \tag{2.7}$$

Berg et coll. (2016) discutent de situations où une hypothèse de PMH peut être considérée comme raisonnable, et fournissent des exemples où une hypothèse de PMH tient alors qu'une hypothèse d'EMH ne tient pas. Si les probabilités de réponse et les probabilités d'inclusion de l'échantillon dépendent d'une variable qui n'est pas incluse dans le modèle d'imputation, alors une hypothèse de PMH peut tenir, alors que ce n'est pas le cas pour une hypothèse d'EMH. Un exemple de variable qui peut être exclue du modèle d'imputation est une variable de plan. L'analyste peut omettre une variable de plan du modèle d'imputation si cette variable n'est pas disponible à l'étape de l'imputation ou si le modèle d'imputation est un modèle spécialisé qui établit un lien entre y_i et x_i. Nous élaborons la procédure de l'IRQ pour qu'elle soit suffisamment souple pour tenir compte soit de la PMH, soit de l'EMH. Dans la pratique, l'analyste peut décider laquelle est plus réaliste pour une application particulière. À la section 2.2, nous expliquons précisément comment la nature de l'hypothèse de RMH peut avoir une incidence sur l'utilisation des poids d'échantillonnage dans la procédure d'estimation. Dans la théorie de la section 3, nous nous concentrons sur la situation dans laquelle l'hypothèse (2.7) tient.

2.2 Régression quantile avec fonction B-Splines pénalisées

Nous évaluons approximativement la fonction quantile définissant la relation entre y_i et x_i dans la superpopulation avec une combinaison linéaire de fonctions de base B-splines. Une fonction B-Spline de base d'ordre p couvre l'espace linéaire des polynômes par morceaux de degré $p-1$ à dévirées continues jusqu'à l'ordre $p-2$. Les B-splines permettent d'améliorer l'efficacité de calcul par rapport à l'utilisation directe de splines polynomiaux (Hastie, Tibshirani et Friedman, 2009).

Pour définir une B-spline, nous empruntons la terminologie de Hastie, Tibshirani, et Friedman (2009) et de Chen et Yu (2016). Supposons que x_i a un support compact sur l'intervalle $[M_1, M_2]$. Nous définissons les nœuds internes $K_n - 1$, espacés à des emplacements équidistants de l'intervalle $[M_1, M_2]$ par $\kappa_i = M_1 + [M_2 - M_1][K_n]^{-1} i$, pour $i = 1, \ldots, K_n - 1$. Nous définissons les nœuds limites p pour M_1 par κ_k pour $k = -p+1, \ldots, 0$, et dénotons les nœuds limites à M_2 par κ_k pour $k = K_n, \ldots, K_n + p - 1$. Les fonctions B-spline de base du degré p pour la séquence de nœuds $\kappa_{-p+1}, \ldots, \kappa_{K_n+p-1}$ sont les éléments du vecteur de dimension $K_n + p$,

$$\mathbf{B}(x) = \left(B_{-p+1}^{[p]}(x), \ldots, B_{K_n}^{[p]}(x)\right)', \tag{2.8}$$

où $B_i^{[s]}(x)$ $(s = 1, \ldots, p)$ est défini de façon récursive par des différences divisées. De façon plus particulière,

$$B_i^{[1]}(x) = I[\kappa_i \le x \le \kappa_{i+1}], \quad \text{pour} \quad i = -p+1, \ldots, K_n + p - 2, \tag{2.9}$$

et

$$B_i^{[s]}(x) = \frac{x - \kappa_i}{\kappa_{i+s-1} - \kappa_i} B_i^{[s-1]}(x) + \frac{\kappa_{i+s} - x}{\kappa_{i+s} - \kappa_{i+1}} B_{i+1}^{[s-1]}(x), \tag{2.10}$$

pour $i = -p+1, \ldots, K_n + p - 1 - s$ et $s = 2, \ldots, p$.

L'estimateur de la fonction de régression quantile est défini par

$$\hat{q}_\tau(x) = \mathbf{B}(x)' \hat{\boldsymbol{\beta}}_\tau, \tag{2.11}$$

où l'estimateur $\hat{\boldsymbol{\beta}}_\tau$ est obtenu en minimisant la forme quadratique,

$$Q_\tau(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i w_i b_i \rho_\tau \left(y_i - \mathbf{B}(x_i)' \boldsymbol{\beta} \right) + \frac{\lambda_n}{2} \boldsymbol{\beta}' \mathbf{D}_m' \mathbf{D}_m \boldsymbol{\beta}, \tag{2.12}$$

où $w_i = \pi_i^{-1} \left(\sum_{i=1}^{n} \pi_i^{-1} \right)^{-1}$, λ_n est un paramètre de lissage spécifié, et $\rho_\tau(\cdot)$, b_i, et $\mathbf{D}_m$ sont définis comme suit. La fonction $\rho_\tau(u)$ au premier terme dans (2.12), est la fonction de vérification de Koenker et Bassett (1978) définie par

$$\rho_\tau(u) = u(\tau - I[u < 0]). \tag{2.13}$$

La fonction de vérification de Koenker (2.13) est un critère d'optimisation standard pour la régression quantile parce que $q_\tau(x)$ minimise la fonction $R(a) = E[\rho_\tau(y - a) | x]$ pour les a. Le deuxième terme dans (2.12) impose une pénalité de rugosité à la fonction de régression quantile estimée. La matrice $\mathbf{D}_m$ est la m^{e} matrice de la différence avec l'élément (i, j), $d_{ij} = (-1)^{j-i} C(m, j-i) I[0 \le j - i \le m] + (1 - I[0 \le j - i \le m])$, où $C(a, b)$ est la fonction de choix. Lorsque $m = 2$, $\mathbf{D}_m$ a une interprétation liée à l'intégrale du carré de la deuxième dérivée de la fonction définie par la fonction B-Spline. Comme la deuxième dérivée d'une ligne droite est zéro, l'utilisation de $\mathbf{D}_m$ pour $m = 2$ réduit la fonction de régression quantile estimée vers une ligne droite. Le choix approprié de b_i au premier terme dans (2.12) dépend des hypothèses formulées au sujet du mécanisme de non-réponse. Si (2.6) tient, on peut établir que $b_i = w_i^{-1}$, ce qui mène à l'équation d'estimation non pondérée de Chen et Yu (2016). Si (2.6) n'est pas satisfaite, l'estimateur non pondéré peut mener à un biais, et fixer $b_i = 1$ constitue une façon d'obtenir un estimateur approximativement sans biais (Berg et coll., 2016). Nous nous concentrons sur le choix conservateur de $b_i = 1$, qui permet d'obtenir des estimateurs cohérents en fonction de (2.7) sans qu'il soit nécessaire de satisfaire à (2.6).

Remarque 1. Par souci de simplicité, nous considérons un x_i univarié avec support dans un intervalle fermé. Chen et Yu (2016) montrent que la procédure s'étend directement à un vecteur $h-$dimensionnel $\mathbf{x}_i$, dont chaque élément a un support dans un intervalle fermé. Pour étendre la procédure à un vecteur $\mathbf{x}_i$, Chen et Yu (2016) définissent $\mathbf{B}(\mathbf{x}_i) = \left(\mathbf{B}(x_{1i})', \mathbf{B}(x_{2i})', \ldots, \mathbf{B}(x_{hi})' \right)$, où $x_{\tilde{h}i}$ est le $\tilde{h}^{\text{e}}$ élément de $\mathbf{x}_i$, pour $\tilde{h} = 1, \ldots, h$.

2.3 Estimation MGM fondée sur l'imputation par régression quantile

Rappelons-nous que le paramètre d'intérêt de la population est défini par l'équation d'estimation dans (2.2). Nous définissons un estimateur d'échantillon complet de $\boldsymbol{\theta}_o$ par

$$\hat{\boldsymbol{\theta}}_A = \text{argmin}_\theta \mathbf{G}_{n,A}(\boldsymbol{\theta})' \mathbf{G}_{n,A}(\boldsymbol{\theta}), \tag{2.14}$$

où

$$\mathbf{G}_{n,A}(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}), \tag{2.15}$$

w_i est défini selon (2.12), et $i = 1, \ldots, n$ indexe les éléments dans A. L'estimateur défini dans (2.15) est un estimateur MGM, selon lequel chaque élément de $\mathbf{G}_{n,A}$ définit un écart entre un moment d'échantillonnage et le paramètre de population correspondant. Par exemple, si $\boldsymbol{\theta}_o = E[y_i]$, alors $\mathbf{g}_i(y_i; \boldsymbol{\theta}_o) = (y_i - \boldsymbol{\theta}_o)$. D'autres exemples sont fournis dans l'étude en simulation de la section 5. Comme y_i n'est pas observée pour les non-répondants, il est impossible d'arriver à $\hat{\boldsymbol{\theta}}_A$.

Une version imputée de (2.15) est définie en remplaçant $\mathbf{g}(y_i, x_i, \boldsymbol{\theta})$ pour une unité non observée i par un estimateur de la valeur attendue. En partant de (2.3), un estimateur de $E_{y|x}[\mathbf{g}(y_i, x_i, \boldsymbol{\theta})]$ est $\int_0^1 \mathbf{g}(\hat{q}_{\tau i}, x_i, \boldsymbol{\theta})\, d\tau$, où $\hat{q}_{\tau i} = \hat{q}_\tau(x_i) = \mathbf{B}(x_i)' \hat{\boldsymbol{\beta}}_\tau$. Nous définissons ensuite l'estimateur $\hat{\boldsymbol{\theta}}$ par

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \left\{ \mathbf{G}_n(\boldsymbol{\theta})' \mathbf{G}_n(\boldsymbol{\theta}) \right\}, \tag{2.16}$$

où

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \left\{ \delta_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}) + (1 - \delta_i) \int_0^1 \mathbf{g}(\hat{q}_{\tau i}, x_i, \boldsymbol{\theta})\, d\tau \right\}. \tag{2.17}$$

Pour un $\mathbf{g}_i$ spécifique, le minimiseur de (2.16) a une expression en forme fermée. Dans le cas où $\boldsymbol{\theta}_o = E[y_i]$, et $\hat{\boldsymbol{\theta}}$ est l'estimateur de Hájek défini par

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{n} w_i \left\{ \delta_i y_i + (1 - \delta_i) \int_0^1 \hat{q}_\tau(x_i)\, d\tau \right\}.$$

Dans d'autres situations, il n'existe peut-être pas d'expression en forme fermée et il est possible d'utiliser des procédures numériques standard, comme Newton-Raphson, pour minimiser (2.17). En dérivant les résultats asymptotiques de la section 3, nous supposons que $\boldsymbol{\theta}_o$ est la valeur unique telle que $E[\mathbf{g}_i(y_i, \boldsymbol{\theta}_o)] = \mathbf{0}$, qui se rapporte à l'existence d'un minimum unique dans (2.16). Voir Fuller (1996, page 252) pour une condition semblable et une discussion de la théorie des estimateurs qui minimisent une forme quadratique.

Dans la pratique, une approximation de l'intégrale est requise. Nous utilisons une approximation du point milieu (c'est-à-dire Nusser, Carriquiry, Dodd et Fuller, 1996). Supposons que la séquence $0 < \tau_1 \leq \tau_2 \cdots \leq \tau_J < 1$ constitue les points milieux des sous-intervalles régulièrement espacés de [0,1] pour J. Pour le non-répondant i, des valeurs imputées J sont construites,

$$y_{ij}^{*} = \mathbf{B}(x_i)' \hat{\boldsymbol{\beta}}_{\tau_j}, \quad j = 1, \ldots, J, \tag{2.18}$$

où $\hat{\boldsymbol{\beta}}_{\tau_j}$ est obtenu en minimisant $Q_{\tau_j}(\boldsymbol{\beta})$ dans (2.12). Nous définissons l'estimateur $\hat{\boldsymbol{\theta}}_J$ pour satisfaire à

$$\hat{\boldsymbol{\theta}}_J = \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ \mathbf{G}_{n,J}(\boldsymbol{\theta})' \mathbf{G}_{n,J}(\boldsymbol{\theta}) \right\}, \tag{2.19}$$

où

$$\mathbf{G}_{n,J}(\boldsymbol{\theta}) := \mathbf{G}_n\left(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{n} w_i \left\{ \delta_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}) + (1-\delta_i) J^{-1} \sum_{j=1}^{J} \mathbf{g}\left(y_{ij}^{*}, x_i, \boldsymbol{\theta}\right) \right\}, \tag{2.20}$$

w_i est défini selon (2.12), et $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_{\tau_1}, \ldots, \hat{\boldsymbol{\beta}}_{\tau_J}\right)'$. La procédure d'imputation ci-dessus diffère de Chen et Yu (2016) en ce sens que l'approximation du point milieu pour l'intégrale est utilisée au lieu de l'intégration Monte Carlo. L'approximation du point milieu et l'intégration de Monte Carlo sont toutes deux justifiées par la transformation de l'intégrale de probabilité, qui relie l'espérance à la fonction quantile conditionnelle, comme expliqué en (2.3). Pour les fonctions avec dérivées secondes bornées, l'erreur dans l'approximation du point milieu est $O(J^{-2})$. Nous préférons également l'approximation du point milieu parce que dans les simulations, elle réduit la variance de l'estimateur et réduit l'instabilité de l'estimateur de variance en raison de quantiles extrêmes par rapport à la simulation de Monte Carlo. James et Wang (2015) discutent du problème potentiel des estimateurs instables pour les quantiles extrêmes des modèles de régression quantile non structurés.

3 Distributions asymptotiques et estimation de la variance

Nous dérivons une distribution normale asymptotique pour l'estimateur de l'IRQ $\hat{\boldsymbol{\theta}}$ défini dans (2.16), bien que, dans la pratique, l'estimateur $\hat{\boldsymbol{\theta}}_J$, défini dans (2.19), avec un nombre fini d'imputations (J), soit nécessaire. Cette approche d'élaboration d'une théorie selon une hypothèse d'un nombre infini de valeurs imputées a déjà été utilisée. Voir, par exemple, Clayton, Spiegelhalter, Dunn et Pickles (1998) et Robins et Wang (2000). Les simulations de la section 5 démontrent que la distribution normale asymptotique dérivée pour $J = \infty$ est une approximation raisonnable de la distribution pour l'estimateur généré avec un J fini. Nous exposons les principaux concepts qui sous-tendent les démonstrations du lemme 1, du lemme 2 et du théorème 1, en reportant les détails à la section B du supplément en ligne https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg et Yu, 2016).

La dérivation de la distribution asymptotique de $\hat{\boldsymbol{\theta}}$ se fait en trois étapes principales. Le lemme 1 donne la distribution asymptotique des estimateurs des coefficients de régression quantile. Le lemme 2 présente la distribution asymptotique de l'équation d'estimation (2.17). Ces deux lemmes sont analogues aux lemmes 1 et 2 de Chen et Yu (2016). Le théorème 1 fournit ensuite la distribution asymptotique de $\hat{\boldsymbol{\theta}}$.

3.1 Normalité asymptotique de $\hat{\theta}$

Nous imaginons une séquence d'échantillons et de populations finies indexée par N, où la taille d'échantillon $n \to \infty$ quand $N \to \infty$. Pour définir les conditions de régularité, nous introduisons la notation $\mathcal{F}_N$ afin de représenter un élément d'une séquence de populations finies de taille N et utilisons la notation « $|\mathcal{F}_N$ » pour indiquer que la distribution de référence est la distribution fondée sur un échantillonnage répété conditionnellement à la population finie de taille N. Par exemple, $E\left[\hat{Y} \mid \mathcal{F}_N\right]$ et $V\left\{\hat{Y} \mid \mathcal{F}_N\right\}$, respectivement, indiquent l'espérance conditionnelle et la variance du résultat $\hat{Y}$ par rapport à la distribution obtenue par randomisation générée par l'échantillonnage répété de $\mathcal{F}_N$. De façon similaire, $\hat{Y} \xrightarrow{d} Y \mid \mathcal{F}_N$ p.s. signifie que $\hat{Y}$ converge en distribution vers Y presque sûrement en ce qui concerne le processus d'échantillonnage répété de la séquence de populations finies quand $N \to \infty$. La convergence est de probabilité « 1 », parce que $\mathcal{F}_N$ est une réalisation aléatoire du modèle de superpopulation (2.1).

Les conditions de régularité sur le plan d'échantillonnage et les paramètres de réglage pour l'estimateur du modèle de fonction B-spline sont les suivantes :

1. Toute variable v_i de sorte que $E\left[\,|v_i|^{2+\delta}\right] < \infty$, où $\delta > 0$, satisfait

$$\sqrt{n}\,(\bar{v}_{\mathrm{HT}} - \bar{v}_N) \mid \mathcal{F}_N \xrightarrow{d} N(0, V_\infty) \quad \text{p.s.}, \tag{3.1}$$

   où $(\bar{v}_{\mathrm{HT}}, \bar{v}_N) = N^{-1}\sum_{i=1}^{N}(\pi_i^{-1} v_i I_i, v_i)$, $V_\infty = \lim_{N\to\infty} V_N$, et $V_N = nV\{\bar{v}_{\mathrm{HT}} \mid \mathcal{F}_N\}$ est la variance conditionnelle de la moyenne Horvitz-Thompson, $\bar{v}_{\mathrm{HT}}$, étant donné $\mathcal{F}_N$.
2. $nn_B^{-1} \to 1$ et $n_B N^{-1} \to f_\infty \in [0, 1]$, où n_B est la taille prévue de l'échantillon.
3. Il existe des constantes C_1, C_2, et C_3 de sorte que $0 < C_1 \le n_B N^{-1}\pi_i^{-1} \le C_2 < \infty$, et

$$\left| n_B\left(\pi_{ij} - \pi_i\pi_j\right)\pi_i^{-1}\pi_j^{-1} \right| \le C_3 < \infty \quad \text{p.s.} \tag{3.2}$$

4. La valeur servant à déterminer le nombre de nœuds intérieurs $K_n = O\left(n_B^{\frac{1}{2p+3}}\right)$.
5. $\lambda_n = O(n_B^{\nu})$ pour $\nu \le (2p+3)^{-1}(p+m+1)$.

La condition 3 est également utilisée dans Fuller (2009a). La condition 3 tient pour l'échantillonnage aléatoire simple, où $\left(\pi_{ij} - \pi_i\pi_j\right)\pi_i^{-1}\pi_j^{-1} = n^{-1}(n-1)(N-1)^{-1}N - 1$, et pour l'échantillonnage de Poisson, où $\left(\pi_{ij} - \pi_i\pi_j\right)\pi_i^{-1}\pi_j^{-1} = 0$. Fuller (2009a) explique que la condition 3 tient pour de nombreux plans stratifiés et que le concepteur a le contrôle pour assurer le respect de la condition 3.

Selon les hypothèses 4 et 5, Barrow et Smith (1978) montrent qu'il existe un $\boldsymbol{\beta}_\tau^*$ qui satisfait

$$\sup_{x\in[M_1,M_2]}\left| q_\tau(x) - b_\tau^a(x) - \mathbf{B}(x)'\,\boldsymbol{\beta}_\tau^* \right| = o\left(K_n^{-(p+1)}\right), \tag{3.3}$$

où $\mathbf{B}(x)'\,\boldsymbol{\beta}_\tau^*$ représente la meilleure approximation L_∞ pour $q_\tau(x)$, et $b_\tau^{(a)}(x)$ est un biais de l'approximation de la fonction B-spline pour la vraie fonction quantile, satisfaisant $b_\tau^a(x) = O\left(K_n^{-(p+1)}\right)$.

Pour obtenir plus de détails sur la forme du terme de biais, voir Chen et Yu (2016) et Yoshida (2013). La propriété (3.3) est largement utilisée dans la dérivation du lemme 1.

Les preuves des lemmes 1 et 2 utilisent un résultat obtenu dans le théorème 1.3.6 de Fuller (2009b). En raison de l'importance de ce théorème pour les résultats de cette section, nous intitulons ce théorème « Fait 1 » :

Fait 1. (Théorème 1.3.6 de Fuller, 2009b) : Supposons que :

$$\left(\hat{\theta} - \theta_N\right) | \mathcal{F}_N \xrightarrow{d} N\left(0, V_{11}\right) \quad \text{p.s.,} \quad \text{et} \quad \theta_N - \theta_o \xrightarrow{d} N\left(0, V_{22}\right). \tag{3.4}$$

Alors, $\left(\hat{\theta} - \theta_o\right) \xrightarrow{d} N\left(0, V_{11} + V_{22}\right)$.

Il est à noter que V_{11} dans le Fait 1 est une limite fixe et non une variance sous le plan, car la variance sous le plan est une fonction aléatoire de la population finie dans ce cadre. La condition $\left(\hat{\theta} - \theta_N\right) | \mathcal{F}_N \xrightarrow{d} N\left(0, V_{11}\right)$ p.s., s'applique à une vaste catégorie de plans, comme ceux dont il a été question dans Isaki et Fuller (1982).

Lemme 1. Selon les hypothèses 1 à 5 et pour les valeurs fixes $x_i \in [M_1, M_2]$ et $\tau \in (0, 1)$,

$$\sqrt{\frac{n}{K_n}}\left(\hat{q}_\tau\left(x_i\right) - \mathbf{B}\left(x_i\right)' \boldsymbol{\beta}_\tau^* + b_\tau^\lambda\left(x_i\right)\right) \xrightarrow{d} N\left(0, \mathbf{B}\left(x_i\right)' \boldsymbol{\Sigma}_\infty\left(\tau\right) \mathbf{B}\left(x_i\right)\right), \tag{3.5}$$

et

$$\sqrt{\frac{n}{K_n}}\left(\hat{q}_\tau\left(x_i\right) - q_\tau\left(x_i\right) + b_\tau^a\left(x_i\right) + b_\tau^\lambda\left(x_i\right)\right) \xrightarrow{d} N\left(0, \mathbf{B}\left(x_i\right)' \boldsymbol{\Sigma}_\infty\left(\tau\right) \mathbf{B}\left(x_i\right)\right), \tag{3.6}$$

où

$$\begin{aligned}
b_\tau^\lambda\left(x_i\right) &= \lim_{N\to\infty} \frac{\tilde{\lambda}_n}{n} \mathbf{B}\left(x_i\right)' \boldsymbol{\Omega}_n\left(\tau\right)^{-1} \mathbf{D}_m' \mathbf{D}_m \boldsymbol{\beta}_\tau^*, \\
\boldsymbol{\Omega}_n\left(\tau\right) &= \mathbf{H}\left(\tau\right) + \frac{\tilde{\lambda}_n}{n} \mathbf{D}_m' \mathbf{D}_m, \\
\boldsymbol{\Sigma}_\infty\left(\tau\right) &= \lim_{N\to\infty} \frac{1}{K_n} \boldsymbol{\Omega}_n\left(\tau\right)^{-1} \left(\mathbf{V}_{1,\infty}\left(\tau\right) + f_\infty \tau\left(1-\tau\right) \boldsymbol{\Phi}\right) \boldsymbol{\Omega}_n\left(\tau\right)^{-1}, \\
\mathbf{H}\left(\tau\right) &= E\left[p_i \mathbf{B}\left(x_i\right) f_{y|x,i}\left(q_{\tau i}\right) \mathbf{B}\left(x_i\right)'\right], \\
\boldsymbol{\Phi} &= E\left[p_i \mathbf{B}\left(x_i\right) \mathbf{B}\left(x_i\right)'\right], \\
\mathbf{V}_{1,\infty}\left(\tau\right) &= \lim_{N\to\infty} \frac{n}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \delta_i \delta_j \mathbf{B}\left(x_i\right) \psi_\tau\left(u_{i\tau}\right) \mathbf{B}\left(x_j\right)' \psi_\tau\left(u_{j\tau}\right),
\end{aligned} \tag{3.7}$$

$u_{i\tau} = y_i - \mathbf{B}\left(x_i\right)' \boldsymbol{\beta}_\tau^*$, $\psi_\tau(u) = \tau - I\left[u < 0\right]$, $\tilde{\lambda}_n = n\hat{N}N^{-1}\lambda_n$, $\hat{N} = \sum_{i=1}^{n} \pi_i^{-1}$, et $f_{y|x,i}\left(q\right)$ est la fdp de y_i étant donné x_i évaluée à q.

L'idée principale de la preuve du lemme 1 est de montrer que l'estimateur du coefficient de régression quantile présente une représentation des quantiles de Bahadur, qui est détaillée dans le corollaire 1 suivant :

Corollaire 1 : Selon la preuve du lemme 1, l'estimateur du coefficient de régression quantile a la représentation des quantiles de Bahadur suivante :

$$\sqrt{\frac{n}{K_n}}\left(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^*_\tau + \frac{\tilde{\lambda}_n}{n}\boldsymbol{\Omega}_n(\tau)^{-1}\mathbf{D}'_m\mathbf{D}_m\boldsymbol{\beta}^*_\tau\right) = \sqrt{\frac{n}{K_n}}\,\boldsymbol{\Omega}_n(\tau)^{-1}\frac{1}{N}\sum_{i=1}^{n}\pi_i^{-1}\delta_i\mathbf{B}(x_i)\psi_\tau(u_{i\tau}) + o_p(1). \quad (3.8)$$

La dérivation de la représentation des quantiles de Bahadur suit l'approche de base de Koenker (2005) et de Yoshida (2013). Pour tenir compte du plan d'échantillonnage complexe, la condition (3.2) est utilisée pour lier des sommes de covariances induites par des probabilités d'inclusion de deuxième ordre non négligeables. Pour les variables aléatoires indépendantes d'une population infinie (par exemple Chen et Yu, 2016; Yoshida,2013; Koenker, 2005), les covariances correspondantes sont de zéro. Compte tenu de la représentation des quantiles de Bahadur (3.8), le lemme 1 découle de l'application de la condition de régularité dans (3.1) et du Fait 1 aux éléments de la moyenne de Horvitz-Thompson dans (3.8). Le $V_{1,\infty}$ dans $\boldsymbol{\Sigma}_\infty(\tau)$ joue essentiellement le rôle de V_{11} dans le Fait 1 et constitue la limite de la variance sous le plan de la moyenne de Horvitz-Thompson. Le deuxième terme dans $\boldsymbol{\Sigma}_\infty(\tau)$ est la variance asymptotique de l'espérance sous le plan de la moyenne de Horvitz-Thompson et joue le rôle de V_{22} dans le Fait 1.

Le lemme 2 et le théorème 1 exigent des conditions de régularités supplémentaires au sujet de l'équation d'estimation. Les conditions de régularité de l'estimation sont semblables à celles de Chen et Yu (2016) et sont donc reportées à la section A du supplément en ligne https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg et Yu, 2016).

Lemme 2. Selon les hypothèses du lemme 1 et les conditions de régularité énoncées à la section A du supplément en ligne https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/Supplement ToQRI.pdf, (Berg et Yu, 2016),

$$\sqrt{n}\,\mathbf{G}_n(\boldsymbol{\theta}_o) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_G(\boldsymbol{\theta}_o)), \quad (3.9)$$

où

$$\begin{aligned}
\mathbf{V}_G(\boldsymbol{\theta}_o) &= f_\infty V\{\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)\} + \lim_{N\to\infty}\mathbf{V}_{\xi,N}(\boldsymbol{\theta}_o), \\
\mathbf{V}_{\xi,N} &= nN^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)\boldsymbol{\xi}_j(\boldsymbol{\theta}_o)', \\
\boldsymbol{\xi}_i(\boldsymbol{\theta}_o) &= \delta_i\mathbf{g}_i(y_i;\boldsymbol{\theta}_o) + (1-\delta_i)\int_0^1\mathbf{g}_i(q_\tau(x_i);\boldsymbol{\theta}_o)\,d\tau + \delta_i\mathbf{h}_{ni}(\boldsymbol{\theta}_o), \\
\mathbf{h}_{n_i}(\boldsymbol{\theta}_o) &= \int_0^1 E\left[(1-p_j)\dot{\mathbf{g}}_{j,y}(q_\tau(x_j);\boldsymbol{\theta}_o)\mathbf{B}(x_j)'\right]\boldsymbol{\Omega}_n(\tau)^{-1}\mathbf{B}(x_i)\psi_\tau(u_{i\tau})\,d\tau,
\end{aligned} \quad (3.10)$$

et $\dot{\mathbf{g}}_{i,y}(y_i;\boldsymbol{\theta}_o)$ est la dérivée partielle de $\mathbf{g}_i(a;\boldsymbol{\theta})$ par rapport à a et évaluée à y_i.

La preuve du lemme 2 est centrée sur un développement en série de Taylor donné par

$$\begin{aligned}\mathbf{g}_i\left(\hat{q}_\tau\left(x_i\right); \boldsymbol{\theta}_o\right) &= \mathbf{g}_i\left(q_\tau\left(x_i\right); \boldsymbol{\theta}_o\right) + \dot{\mathbf{g}}_{i,y}\left(q_\tau\left(x_i\right); \boldsymbol{\theta}_o\right)\left(\hat{q}_\tau\left(x_i\right) - q_\tau\left(x_i\right)\right) \\ &\quad + \ddot{\mathbf{g}}_{i,y}\left(q_\tau\left(x_i\right); \boldsymbol{\theta}_o\right)\left(\tilde{q}_\tau\left(x_i\right) - q_\tau\left(x_i\right)\right)^2, \end{aligned} \tag{3.11}$$

où $\tilde{q}_\tau\left(x_i\right)$ se situe entre $\hat{q}_\tau\left(x_i\right)$ et $q_\tau\left(x_i\right)$, et $\ddot{\mathbf{g}}_{i,y}\left(q_\tau\left(x_i\right); \boldsymbol{\theta}_o\right)$ indique le vecteur des dérivées partielles des éléments de $\dot{\mathbf{g}}_{i,y}\left(a, \boldsymbol{\theta}_o\right)$ par rapport à a et évalué à $q_\tau\left(x_i\right)$. Selon des arguments semblables à ceux de Chen et Yu (2016), $n\left\| \ddot{g}_{i,y}\left(q_\tau\left(x_i\right); \boldsymbol{\theta}_o\right)\left(\tilde{q}_\tau\left(x_i\right) - q_\tau\left(x_i\right)\right)^2 \right\| = O(1)$. Le lemme 2 découle ensuite de l'approximation linéaire pour $\hat{q}_\tau\left(x_i\right) - q_\tau\left(x_i\right)$ dans le lemme 1.

Théorème 1. Sous les hypothèses des lemmes 1 et 2, l'estimateur de l'IRQ $\hat{\boldsymbol{\theta}}$ défini dans (2.16), construit avec $J = \infty$, satisfait $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right)$, où

$$\boldsymbol{\Sigma}_\theta = \left[\boldsymbol{\Gamma}\left(\boldsymbol{\theta}_o\right)' \boldsymbol{\Gamma}\left(\boldsymbol{\theta}_o\right)\right]^{-1} \boldsymbol{\Gamma}\left(\boldsymbol{\theta}_o\right)' \mathbf{V}_G\left(\boldsymbol{\theta}_o\right) \boldsymbol{\Gamma}\left(\boldsymbol{\theta}_o\right)\left[\boldsymbol{\Gamma}\left(\boldsymbol{\theta}_o\right)' \boldsymbol{\Gamma}\left(\boldsymbol{\theta}_o\right)\right]^{-1}, \tag{3.12}$$

$\mathbf{G}(\boldsymbol{\theta}) = E\left[\mathbf{G}_N\left(\boldsymbol{\theta}, \mathbf{y}\right)\right]$, $\mathbf{G}_N\left(\boldsymbol{\theta}, \mathbf{y}\right) = N^{-1}\sum_{i=1}^{N} \delta_i g\left(y_i, x_i\right)$, et $\boldsymbol{\Gamma}\left(\boldsymbol{\theta}_o\right) = E\left[\partial / \partial \boldsymbol{\theta}\, \mathbf{G}_N\left(\boldsymbol{\theta}\right)\right]$.

Selon Pakes et Pollard (1989), le théorème 1 est satisfait si ce qui suit tient

1. $\sup_\theta \left|\mathbf{G}_n\left(\boldsymbol{\theta}\right) - \mathbf{G}\left(\boldsymbol{\theta}\right)\right| = o_p(1)$,
2. Si $\zeta_n \to 0$, $\sup_{\left|\boldsymbol{\theta}-\boldsymbol{\theta}_o\right| < \zeta_n} \left|\mathbf{G}_n\left(\boldsymbol{\theta}\right) - \mathbf{G}\left(\boldsymbol{\theta}\right) - \mathbf{G}_n\left(\boldsymbol{\theta}_o\right)\right| = o_p\left(n_B^{-0,5}\right)$,

où ζ_n est arbitrairement petit. En raison de la complexité du plan d'échantillonnage, la preuve que ces conditions tiennent se fait en deux étapes, en tenant compte d'abord de l'écart $\left|\mathbf{G}_n\left(\boldsymbol{\theta}\right) - \mathbf{G}_N\left(\boldsymbol{\theta}\right)\right|$ et ensuite de l'écart $\left|\mathbf{G}_N\left(\boldsymbol{\theta}\right) - \mathbf{G}\left(\boldsymbol{\theta}\right)\right|$. Le résultat découle ensuite de l'inégalité du triangle.

3.2 Estimation de la variance

Nous estimons la variance de $\hat{\boldsymbol{\theta}}_J$ en utilisant la méthode de linéarisation (Fuller, 2009b, page 64). Nous utilisons la matrice de covariance asymptotique dans (3.12) pour estimer la variance de $\hat{\boldsymbol{\theta}}_J$, l'estimateur de $\boldsymbol{\theta}_o$ défini dans (2.19), construit avec un nombre fini de valeurs imputées. Pour estimer $\mathbf{V}_G\left(\boldsymbol{\theta}_o\right)$, un estimateur de variance conforme au plan est appliqué à un estimateur de la moyenne d'un estimateur de $\boldsymbol{\xi}_i\left(\boldsymbol{\theta}_o\right)$ défini dans (3.10). L'estimateur de $\boldsymbol{\xi}_i\left(\boldsymbol{\theta}_o\right)$ est obtenu en remplaçant $\boldsymbol{\theta}_o$ et $\boldsymbol{\beta}_\tau^*$ avec les estimateurs $\hat{\boldsymbol{\theta}}_J$ et $\hat{\boldsymbol{\beta}}_\tau$, respectivement.

L'estimateur de variance est défini comme

$$\hat{\boldsymbol{\Sigma}}_\theta = \left[\hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)' \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)\right]^{-1} \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)' \left[\hat{\mathbf{V}}_{G,\infty}\left(\hat{\boldsymbol{\theta}}_J\right)\right] \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)\left[\hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)' \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)\right]^{-1}, \tag{3.13}$$

où $\hat{\mathbf{V}}_{G,\infty}\left(\hat{\boldsymbol{\theta}}_J\right) = \hat{f}_\infty \hat{V}\left\{\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right\} + \hat{\mathbf{V}}_{\xi,N}\left(\hat{\boldsymbol{\theta}}_J\right)$,

$$\hat{V}\left\{\hat{\xi}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right\} = \frac{1}{\hat{N}}\sum_{i=1}^{n}\pi_i^{-1}\hat{\xi}_i\left(\hat{\boldsymbol{\theta}}_J\right)\hat{\xi}_i\left(\hat{\boldsymbol{\theta}}_J\right)' - \frac{1}{\hat{N}\left(\hat{N}-1\right)}\left(\sum_{i=1}^{n}\pi_i^{-1}\hat{\xi}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right)\left(\sum_{i=1}^{n}\pi_i^{-1}\hat{\xi}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right)', \tag{3.14}$$

$$\hat{\mathbf{V}}_{\xi,N}\left(\hat{\boldsymbol{\theta}}_J\right) = \frac{n}{\hat{N}^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\hat{\xi}_i\left(\hat{\boldsymbol{\theta}}_J\right)\hat{\xi}_j\left(\hat{\boldsymbol{\theta}}_J\right)',$$

$$\hat{\xi}_i\left(\hat{\boldsymbol{\theta}}_J\right) = \delta_i\mathbf{g}_i\left(y_i;\,\hat{\boldsymbol{\theta}}_J\right) + (1-\delta_i)\,J^{-1}\sum_{j=1}^{J}\mathbf{g}_i\left(\mathbf{B}(x_i)'\,\hat{\boldsymbol{\beta}}_{\tau_j};\,\hat{\boldsymbol{\theta}}_J\right) + \delta_i\hat{\mathbf{h}}_{ni}\left(\hat{\boldsymbol{\theta}}_J\right),$$

$$\hat{\mathbf{h}}_{ni}\left(\hat{\boldsymbol{\theta}}_J\right) = J^{-1}\sum_{j=1}^{J}N^{-1}\sum_{k=1}^{n}\pi_k^{-1}(1-\delta_k)\,\dot{\mathbf{g}}_{k,y}\left(\mathbf{B}(x_k)'\,\hat{\boldsymbol{\beta}}_{\tau_j};\,\hat{\boldsymbol{\theta}}_J\right)\mathbf{B}(x_k)'\,\hat{\boldsymbol{\Omega}}_n\left(\tau_j\right)^{-1}\mathbf{B}(x_i)\,\psi_\tau\left(\hat{u}_{i\tau_j}\right),$$

$$\hat{\boldsymbol{\Omega}}_n\left(\tau_j\right) = \hat{\mathbf{H}}\left(\tau_j\right) + \frac{\hat{f}_\infty\tilde{\lambda}_n}{n}\mathbf{D}_m'\mathbf{D}_m,$$

$$\hat{\mathbf{H}}(\tau) = \frac{1}{\hat{N}}\sum_{i=1}^{n}\pi_i^{-1}\delta_i\mathbf{B}(x_i)\,\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right)\mathbf{B}(x_i)',$$

$\hat{f}_\infty = n\hat{N}^{-1}$, $\hat{N} = \sum_{i=1}^{n}\pi_i^{-1}$, et $\hat{u}_{i\tau_j} = y_i - \mathbf{B}(x_i)'\,\hat{\boldsymbol{\beta}}_{\tau_j}$. Un estimateur de $\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right)$ est l'inverse d'un estimateur de la dérivée de la fonction quantile et est défini par

$$\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right) = \max\left\{\frac{2a_{n,\tau}}{\mathbf{B}(x_i)'\left(\hat{\boldsymbol{\beta}}_{\tau+a_{n,\tau}} - \hat{\boldsymbol{\beta}}_{\tau-a_{n,\tau}}\right)},\,0\right\}, \tag{3.15}$$

où la largeur de bande $a_{n,\tau}$ est donnée par

$$a_{n,\tau} = n^{-0,2}\left[\frac{4{,}5\phi\left(\Phi^{-1}(\tau)\right)^4}{\left(2\Phi^{-1}(\tau)^2+1\right)^2}\right], \tag{3.16}$$

avec $\phi(\cdot)$ et $\Phi(\cdot)$, respectivement la fdp et la fdc d'une distribution normale centrée réduite. Voir Wei, Ma et Carroll (2012) et Koenker (2005) pour des analyses au sujet de (3.15) et de (3.16), respectivement.

4 Application au *Conservation effects assessment project*

La composante des terres en cultures du *Conservation effects assessment project* (CEAP) consiste en une série d'enquêtes visant à mesurer la perte de sol et d'éléments nutritifs provenant des champs de culture. La première évaluation des terres cultivées a été une enquête nationale menée de 2003 à 2006. La collecte des données d'une deuxième enquête nationale, prévue pour 2015 à 2016, était en cours durant l'écriture de ce document. Chacune des périodes 2003 à 2006 et 2015 à 2016 est considérée comme un point dans le temps pour l'estimation. Les données sont recueillies sur plusieurs années (c'est-à-dire de 2003 à 2006 ou de 2015 à 2016) pour des raisons opérationnelles, et aucune unité ne fait partie de l'échantillon pendant deux ans au cours de la même période. Les changements temporels d'intérêt sont les changements entre deux périodes, plutôt que les changements entre deux années dans la même période. La structure temporelle

entraîne un déséquilibre des données, parce que certaines unités répondent dans les deux périodes, certaines unités ne répondent jamais et certaines unités ne répondent que dans l'une des deux périodes. Le fait de fournir à l'utilisateur de données un ensemble complet de données imputées avec un seul ensemble de poids simplifie les analyses comportant plus d'un point dans le temps.

Nous étudions la faisabilité de l'imputation pour le CEAP à l'aide d'un sous-ensemble de données recueillies de 2003 à 2005. Nous omettons les données recueillies en 2006 parce que le plan d'échantillonnage a changé et nous n'avons pas l'information nécessaire pour calculer les poids d'échantillonnage pour l'enquête de 2006. Les données de l'enquête de 2015 à 2016 ne sont pas encore recueillies. Cette analyse est considérée comme une étude de la faisabilité de l'utilisation de l'IRQ pour imputer les données manquantes dans le CEAP en vue de régler le problème plus vaste de l'estimation du changement au fil du temps.

Pour comprendre le plan d'échantillonnage du CEAP, il faut comprendre la conception du *National Resources Inventory* (NRI). Le NRI surveille l'état et les tendances de l'utilisation des terres, de la couverture terrestre et de l'érosion, en mettant l'accent sur les caractéristiques liées aux ressources naturelles et à l'agriculture. Les unités d'échantillonnage primaires du NRI sont des zones appelées segments, qui ont une superficie d'environ 160 acres. Chaque segment contient environ trois unités d'échantillonnage secondaires, qui sont des emplacements choisis au hasard appelés « points ». De 1982 à 1997, le même échantillon d'environ 300 000 segments, appelé échantillon de fondation, a été réexaminé tous les cinq ans. L'échantillon de fondation est un échantillon stratifié de segments, avec un taux d'échantillonnage typique d'environ 4 %. Voir Nusser et Goebel (1997) pour des détails sur le plan de l'échantillon de fondation du NRI. En 2000, le NRI est passé au plan d'échantillonnage annuel. Comme il est impossible de revoir chaque segment échantillonné dans l'échantillon de fondation chaque année, un plan de sondage avec renouvellement de panel est utilisé. Un sous-échantillon des segments de fondation, appelé « panel principal », est revu chaque année. Le panel principal est complété par un panel rotatif, qui change chaque année. Essentiellement, les panels principaux et rotatifs sont des échantillons stratifiés de l'échantillon de fondation. Les strates, appelées classes d'échantillonnage, dépendent des caractéristiques du segment du NRI observé de 1982 à 1997, comme la présence de terres humides, de terres cultivées et de forêts. Voir Nusser (2006) et Breidt et Fuller (1999) pour obtenir plus de détails sur les échantillons annuels du NRI.

Pour le CEAP, les responsables de la collecte des données visitent un sous-ensemble de points du NRI qui se trouvent dans les champs de culture échantillonnés et recueillent des informations plus détaillées sur les choix de cultures et les pratiques de conservation. L'échantillon de l'enquête de 2003 à 2005 du CEAP comprend essentiellement de segments du panel principal du NRI, du panel rotatif de 2002 et du panel rotatif de 2003 qui contiennent au moins un point de culture. Pour les segments contenant plus d'un point de culture, un point de culture a été sélectionné au hasard. La sélection d'un point par segment vise à améliorer la répartition géographique et à réduire le nombre de cas où un exploitant agricole associé à de multiples points d'échantillonnage est sélectionné dans l'échantillon, ce qui réduit le fardeau du répondant.

Comme le taux d'échantillonnage de la première phase pour le NRI est faible $(\approx 4\,\%)$, nous estimons l'échantillon du CEAP en tant qu'échantillon avec probabilités proportionnelles à la taille avec remise. Les probabilités de sélection pour le CEAP reflètent en grande partie le plan d'échantillonnage pour le NRI. Les détails de la construction des probabilités de sélection de premier et deuxième ordre pour le CEAP sont fournis à la section C du supplément en ligne https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg et Yu, 2016).

La collecte des données pour les champs de culture échantillonnés pour l'enquête du CEAP comprend plusieurs éléments. Une enquête auprès des agriculteurs qui recueille des renseignements détaillés sur les pratiques de gestion et de conservation agricoles constitue l'un de ces éléments importants. Il est possible d'obtenir une non-réponse dans le CEAP si un producteur refuse de participer à l'entrevue.

Les variables de réponse dans le CEAP sont des mesures de différents types de perte de sol et d'éléments nutritifs, obtenues à partir d'un modèle de processus réel appelé « Modèle de politique agricole/environnementale (APEX) ». Le modèle APEX convertit les données des enquêtes auprès des agriculteurs ainsi que les renseignements provenant des sources administratives et du NRI en mesures numériques de l'érosion. Dans le cadre de cette étude, nous examinons une mesure de la perte de sol causée par l'érosion en rigoles et en surface appelée RUSLE2, dont il est question plus en détail à la section 4.1.

L'enquête du NRI constitue une source pratique d'information auxiliaire pour l'imputation des variables de réponse du CEAP. Puisque les données de l'enquête du NRI sont recueillies au moyen de photographies aériennes de segments échantillonnés, il n'y a pas de non-réponse en raison de refus dans le NRI. Par conséquent, les données du NRI sont disponibles pour tous les points échantillonnés dans le CEAP. De plus, le NRI recueille des données sur l'utilisation des terres, les pratiques de conservation et l'érosion − des caractéristiques qui devraient être corrélées aux extrants du modèle APEX. En tant que variable auxiliaire, nous utilisons USLE, une mesure de l'érosion en rigoles et en surface recueillie dans le NRI.

Les domaines d'intérêt du CEAP sont les dix « régions de production du CEAP ». Nous nous concentrons sur l'estimation de la valeur moyenne de RUSLE2 pour sept États (Iowa, Illinois, Indiana, Michigan, Minnesota, Ohio et Wisconsin) qui forment la majeure partie de la région de production du CEAP appelée la *Corn Belt*. Nous utilisons la régression quantile semi-paramétrique pour imputer les valeurs manquantes pour RUSLE2 en utilisant USLE comme variable auxiliaire pour chacun de ces sept États dans la région de la *Corn Belt*.

4.1 Modèle et procédures d'imputation

La variable d'intérêt, RUSLE2, est une mesure de l'érosion en rigoles et en surface obtenue à partir du modèle APEX. Comme l'intérêt est dans l'érosion moyenne par acre, le paramètre d'intérêt θ, la moyenne RUSLE2 de l'érosion dans l'État est défini comme un ratio par

$$\theta = \frac{E\left[\sum_{i=1}^{m_{ek}} R_{ik} D_k m_k^{-1}\right]}{E\left[m_{ek} D_k m_k^{-1}\right]}, \tag{4.1}$$

où R_{ik} est l'érosion RUSLE2 pour le point i dans le segment k échantillonné pendant la période de 2003 à 2005, D_k est la zone du segment k, m_k est le nombre total de points dans le segment k, et m_{ek} est le nombre de points dans le segment k qui sont admissibles à l'enquête du CEAP. Comme il a été mentionné précédemment, la période de 2003 à 2005 est considérée comme un point dans le temps, et aucun point n'est échantillonné plus d'une fois dans cet ensemble d'années. Par conséquent, chaque unité échantillonnée a une valeur R_{ik} pour cet ensemble d'années et R_{ik} n'a pas besoin d'un indice inférieur t pour l'année.

L'érosion RUSLE2 est une version améliorée d'une mesure plus simple de l'érosion en rigoles et en surface appelée USLE. La mesure USLE est le produit de cinq indices numériques associés à l'inclinaison et à la longueur des pentes, aux précipitations, à l'érosion du sol, aux pratiques de conservation et à la gestion des cultures. Bien que RUSLE2 ne soit observée que pour les répondants à l'enquête du CEAP, USLE est disponible à partir de l'échantillon principal du NRI pour tous les points de l'échantillon du CEAP. Nous utilisons la moyenne de la mesure USLE pour les années 2003 à 2005 comme covariable dans le modèle d'imputation. Plus précisément, pour le point i dans le segment k, nous définissions $U_{ik} = 3^{-1}\sum_{t=2003}^{2005} U_{tik}$, où U_{tik} est la perte de sol mesurée par USLE dans le NRI pour le point i dans le segmentn k pour l'année t.

Comme les mesures RUSLE2 et USLE sont fortement asymétriques, le modèle de régression quantile est appliqué après transformation de R_{ik} et de U_{ik} par une puissance de 0,2. Le modèle de régression quantile postulé pour la superpopulation peut être exprimé comme $P(y_{ik} \le \tau \mid x_{ik}) = q_\tau(x_{ik})$, où $y_{ik} = R_{ik}^{0,2}$, et $x_{ik} = U_{ik}^{0,2}$. La fonction inconnue $q_\tau(x_{ik})$ est estimée par une combinaison linéaire de fonctions de base B-spline générées à partir de x_{ik}. Pour définir la fonction B-spline pénalisée, nous avons établi que $p = 3$, $m = 2$, $K_n = 16$, et $\lambda = 0{,}004$.

Comme la quantité d'intérêt est l'érosion par acre, l'estimateur $\hat{\theta}$ de θ défini dans (4.1) est un ratio de deux estimateurs. C'est-à-dire, $\hat{\theta} = \hat{\theta}_2^{-1}\hat{\theta}_1$, où $\hat{\theta}_1$ est un estimateur de $\theta_1 = E[D_k U_{ik}]$, et $\theta_2 = E[D_k]$. L'estimateur de θ_2 est l'estimateur Hájek, $\hat{\theta}_2 = \left(\sum_{k=1}^{n} \pi_{ik}^{-1} D_k\right)\left(\sum_{k=1}^{n} \pi_{ik}^{-1}\right)^{-1}$, où π_{ik} est la probabilité de sélectionner le point i dans le segment k dans l'échantillon du CEAP. L'estimateur $\hat{\theta}_1$ de θ_1 est obtenu par MGM avec $g(y, \theta_1) = (D_k y^5 - \theta_1)$.

4.2 Estimations et estimations de la variance

Le tableau 4.1 contient des estimations de la perte de sol RUSLE2 moyenne selon l'IRQ, ainsi que des erreurs-types estimées pour sept États de la région de la *Corn Belt* du CEAP. À des fins de comparaison, l'estimateur de cas complet $(\bar{R}_{cc})$ et l'erreur-type estimée correspondante sont également fournis au tableau 4.1. L'estimateur de cas complet est le ratio des estimateurs de Hájek construits en utilisant seulement les unités qui fournissent une réponse utilisable pour RUSLE2.

Pour chacun des sept états, l'estimateur de cas complet est plus grand que l'estimateur fondé sur les données imputées. La procédure d'imputation réduit l'estimateur de θ, par rapport à l'estimateur de cas

complet, parce que la moyenne pondérée de U_{ik} parmi les unités échantillonnées est plus petite que la moyenne de U_{ik} parmi les répondants, comme le montrent les deux dernières rangées du tableau 4.1.

Comme prévu, l'erreur-type estimée pour $\hat{\theta}$ est plus petite que l'erreur-type estimée pour l'estimateur de cas complet. Les ratios des variances estimées pour l'estimateur de cas complet par rapport aux variances estimées de $\hat{\theta}$ vont de 1,103, pour le Minnesota (MN), à 1,252, pour l'Indiana (IN). Cette comparaison démontre le potentiel de gains d'efficience découlant de l'utilisation de l'imputation. La réduction de l'écart-type estimé se produit parce que la procédure d'imputation utilise U_{ik} pour l'échantillon complet, tandis que l'estimateur de cas complet est fondé seulement sur R_{ik} pour le sous-ensemble de répondants.

Tableau 4.1
Estimateur de cas complet $\left(\bar{R}_{cc}\right)$ et estimateur IRQ-MGM $\left(\hat{\theta}\right)$ de la perte de sol RUSLE2 moyenne (θ), erreurs-types correspondantes, taille d'échantillon (n), nombre de répondants (n_r), et moyennes pondérées de covariables pour les unités échantillonnées $\left(\hat{U}_s\right)$ et moyennes pondérées de covariables parmi les répondants $\left(\hat{U}_r\right)$ pour sept états dans la *Corn Belt*

	IL	IN	IA	MI	MN	OH	WI
$\bar{R}_{cc}$	0,3301	0,2994	0,3464	0,3214	0,1741	0,3700	0,5226
$\mathrm{SE}\left(\bar{R}_{cc}\right)$	0,0112	0,0179	0,0144	0,0209	0,0068	0,0213	0,0354
$\hat{\theta}$	0,3281	0,2901	0,3408	0,3145	0,1646	0,3636	0,4977
$\mathrm{SE}\left(\hat{\theta}\right)$	0,0106	0,0160	0,0134	0,0189	0,0063	0,0201	0,0337
n	1 823	1 151	1 492	935	1 649	1 053	662
n_r	1 275	751	1 011	585	1 008	698	414
$\hat{U}_r$	4,0775	3,7781	5,2046	1,6029	2,1063	2,1071	4,7586
$\hat{U}_s$	4,0909	3,6107	5,0385	1,5776	1,8973	2,0761	4,2232

5 Simulations

Nous construisons une étude de simulation pour représenter les propriétés des données et du plan du CEAP. Un ensemble étendu de simulations utilisant les modèles de simulation de Chen et Yu (2016) donne des résultats semblables et n'est pas présenté ici pour plus de concision. Les objectifs des simulations sont d'évaluer l'estimateur de variance et de comparer l'IRQ aux alternatives non paramétriques et entièrement paramétriques.

La procédure d'imputation entièrement paramétrique est l'imputation fractionnaire paramétrique (Kim, 2011). Le modèle d'imputation spécifié pour l'imputation fractionnaire paramétrique (IFP) est $y_i = \gamma_0 + \gamma_1 x_i + \epsilon_i$, où $\epsilon_i \sim N\left(0, \sigma_\epsilon^2\right)$. les valeurs imputées pour l'IFP sont générées comme $y_{ij}^* \sim N(\hat{\gamma}_0 + \hat{\gamma}_1 x_i, \hat{\sigma}_\epsilon^2)$, où $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}_\epsilon^2)'$ satisfait $\mathbf{S}_w(\hat{\boldsymbol{\gamma}}) = \mathbf{0}$,

$$\mathbf{S}_w(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \pi_i^{-1} \delta_i \mathbf{d}_i, \tag{5.1}$$

et $\mathbf{d}_i = \left(y_i - \gamma_0 - \gamma_1 x_i, (y_i - \gamma_0 - \gamma_1 x_i)\, x_i, (y_i - \gamma_0 - \gamma_1 x_i)^2 / \sigma_\epsilon^2 - 1\right)'$. En incorporant π_i^{-1} dans la fonction de score (5.1), l'estimateur est cohérent si le modèle de population est un modèle linéaire avec erreurs iid à distribution normale et si l'hypothèse de RMH dans (2.7) ou dans (2.6) tient.

La procédure d'imputation non métrique (IPN) est fondée sur Wang et Chen (2009). Pour l'IPN, la j^{e} valeur imputée pour le non-répondant i, y_{ij}^*, est générée à partir d'une distribution multinomiale avec espace d'échantillonnage $\{y_s : I_s = \delta_s = 1\}$. Plus précisément,

$$P\left(y_{ij}^* = y_s\right) = \frac{\pi_i^{-1} K\left\{(x_i - x_s)/h\right\}}{\sum_{j=1}^{N} I_j \delta_j \pi_j^{-1} K\left\{(x_i - x_j)/h\right\}}, \tag{5.2}$$

où $K(\cdot)$ est un noyau normal de largeur de bande h sélectionné en appliquant la méthode de Sheather et Jones (1991), telle qu'elle est mise en œuvre dans la fonction R $dpik$, à $\{x_i : I_s = \delta_s = 1\}$.

La procédure de l'IRQ est mise en œuvre conformément à la description fournies aux sections 2 et 3. Pour définir la fonction B-spline pénalisée, nous avons établi que $p = 3$, $m = 2$, $K_n = 16$, et $\lambda = 0{,}004$. La valeur de $\lambda = 0{,}004$ est la médiane des valeurs sélectionnées à l'aide de la fonction R « Cobbs » dans 1 000 échantillons d'une simulation préliminaire. Pour sélectionner λ à l'aide de « Cobbs », nous utilisons d'abord la fonction R « Cobbs » pour obtenir λ_{τ_j} pour $\tau_1, \ldots, \tau_J$. Le λ sélectionné est le minimum de $\left\{\lambda_{\tau_j} : j = 1, \ldots, J\right\}$, qui introduit le moins de lissage parmi les λ_{τ_j} sélectionnés.

Dans des simulations qui ne sont pas présentées ici, nous tenons également compte de l'imputation multiple. Des modifications aux procédures standard d'imputation multiple sont nécessaires pour produire des estimateurs sans biais pour une situation selon laquelle l'hypothèse de l'EMH (2.6) ne tient pas (Berg et coll., 2016; Reiter, Raghunathan et Kinney, 2006). Étant donné qu'une exploration des modifications à l'imputation multiple nécessaires pour assurer une estimation uniforme dépasse la portée de la présente étude, nous limitons notre attention à l'IFP, à l'IPN et à l'IRQ.

Pour les trois procédures d'imputation, la MGM fondée sur les valeurs imputées est utilisée pour estimer les paramètres. Il convient de souligner que cela diffère de Wang et Chen (2009), qui utilisent la probabilité empirique plutôt que la MGM. Le nombre d'imputations pour la simulation est $J = 50$. La taille de l'échantillon Monte Carlo (MC) est de 1 000.

Nous prenons en considération l'estimation de plusieurs paramètres : $\theta_1 = E[y_i]$, $\theta_2 = V\{y_i\}$, $\theta_3 = \text{Cor}\{y_i, x_i\}$, $\theta_4 = E[E[y_i \mid x_i \le 0{,}65]]$, et $\theta_5 = P(y_i \le 8)$. À l'exception de θ_5, les estimateurs de MGM de ces paramètres satisfont aux hypothèses requises pour la théorie de la section 3. En particulier, la fonction $\mathbf{g}_i(\cdot; \boldsymbol{\theta})$ qui définit l'estimateur de $(\theta_1, \theta_2, \theta_3, \theta_4)$ a deux dérivées continues. L'estimateur de θ_5 ne fait pas partie du cadre de la section 3, parce que $I[a \le 8]$ est une fonction non lisse de a; toutefois, nous évaluons les propriétés empiriques de $\hat{\theta}_5$ définies comme suit :

$$\hat{\theta}_5 = \left(\sum_{i=1}^{n} \pi_i^{-1}\right)^{-1} \sum_{i=1}^{n} \pi_i^{-1} \left\{\delta_i I[y_i \le 8] + (1 - \delta_i)\, J^{-1} \sum_{j=1}^{J} I\left[y_{ij}^* \le 8\right]\right\}. \tag{5.3}$$

Pour obtenir plus de détails sur la fonction $\mathbf{g}_i(\cdot; \boldsymbol{\theta})$ qui définit les estimateurs pour la simulation, voir la section D du supplément en ligne https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg et Yu, 2016).

5.1 Modèle de superpopulation et plan pour les simulations

Le modèle de superpopulation représente quatre aspects des données et de l'enquête du CEAP : (1) la forme de fonction d'espérance, (2) l'inclusion d'une relation moyenne-variance, (3) l'utilisation de la probabilité proportionnelle à la taille (PPT) avec échantillonnage avec remise et (4) la taille de l'échantillon et les taux de réponse. Le modèle spécifique pour la simulation est $y_i = m(x_i) + e_i$, où $e_i \sim N\left(0, \sigma_e^2 m(x_i)^2\right)$, $m(x_i) = 2 + 10\left(1 + 8\exp(-5x_i)\right)^{-\frac{5}{4}}$, et $x_i \sim$ Trunc. Norm. $(0{,}5;\ 0{,}3)$. Le plan d'échantillonnage est PPT avec remise, où la probabilité de sélectionner une unité i en un seul tirage est $\left(\sum_{i=1}^{N} \tilde{\psi}_i\right)^{-1} \tilde{\psi}_i$, $\text{logit}(\tilde{\psi}_i) = -3 - 0{,}33z_i + 0{,}1y_i$, $z_i \sim$ Trunc. Norm. $(0{,}5;\ 0{,}3)$, et $N = 50\ 000$. Le nombre de tirages est $n = 1\ 500$, ce qui mène à une taille d'échantillon médiane de 1 477, où la taille de l'échantillon est le nombre d'unités uniques dans l'échantillon. Les probabilités de sélection du premier et du deuxième ordre correspondant à $\tilde{\psi}_i$ sont $\pi_i = 1 - (1 - \tilde{\psi}_i)^n$, et $\pi_{ij} = 1 - (1 - \tilde{\psi}_i)^n - (1 - \tilde{\psi}_j)^n + (1 - \tilde{\psi}_j - \tilde{\psi}_i)^n$. L'indicateur de réponse $\delta_i \sim \text{Bernoulli}(p_i)$, où $\text{logit}(p_i) = 0{,}5x_i + 1{,}5z_i$, ce qui donne un taux de réponse médian de 0,631.

Selon le modèle pour y_i compte tenu de x_i, l'hypothèse de PMH (2.7) tient pour cette simulation. L'intégration de z_i dans les modèles pour p_i et π_i est l'approche utilisée dans Berg et coll. (2016) qui cause l'échec de l'hypothèse de l'EMH (2.6). La variable z_i peut être interprétée comme une variable de plan qui est omise du modèle d'imputation.

5.2 Résultats

Le tableau 5.1 contient trois mesures pour comparer l'estimateur de l'IRQ aux estimateurs de l'IFP et de l'IPN. L'erreur quadratique moyenne (EQM) de MC relative en pourcentage pour l'estimateur k $(k = \text{IFP, IPN})$ est définie par

$$\text{Pct. Rel. EQM}(k) = 100\frac{\text{EQM}_{\text{MC}}\left(\hat{\theta}(k)\right) - \text{EQM}_{\text{MC}}\left(\hat{\theta}(\text{IRQ})\right)}{\text{EQM}_{\text{MC}}\left(\hat{\theta}(\text{IRQ})\right)}, \tag{5.4}$$

où $\hat{\theta}(k)$ est l'estimateur fondé sur la procédure d'imputation k. La variance relative en pourcentage pour l'estimateur k est définie par

$$\text{Pct. Rel. Var}(k) = 100\frac{\text{Var}_{\text{MC}}\left(\hat{\theta}(k)\right) - \text{Var}_{\text{MC}}\left(\hat{\theta}(\text{IRQ})\right)}{\text{Var}_{\text{MC}}\left(\hat{\theta}(\text{IRQ})\right)}, \tag{5.5}$$

si $k = \text{IPN, IFP}$. Le pourcentage d'EQM attribuable au biais au carré est défini par

$$\text{Pct. Biais}\,(k) = 100\frac{\left(E_{\text{MC}}\left(\hat{\theta}\,(k)\right) - \theta\right)^2}{\text{EQM}_{\text{MC}}\left(\hat{\theta}\,(k)\right)}, \tag{5.6}$$

où k = IPN, IFP, IRQ. L'EQM de l'estimateur IRQ est plus petite que l'EQM des estimateurs de l'IPN et de l'IFP pour tous les paramètres. L'estimateur de l'IFP est biaisé parce que le modèle sous-jacent à la procédure de l'IFP ne tient pas compte de la non-linéarité dans les courbes quantiles ou des variances non constantes. La procédure de l'IPN présente une variance relativement grande pour les tailles d'échantillon comme celles obtenues dans l'enquête du CEAP. Le biais MC au carré de l'IRQ est inférieur à 0,5 % de l'EQM de MC pour tous les paramètres.

Les deux dernières colonnes du tableau 5.1 contiennent le biais relatif de l'estimateur de variance et la couverture empirique des intervalles de confiance de 95 % de la théorie normale. Le biais relatif de l'estimateur de variance est défini par

$$\text{Biais Rel.} = \frac{E_{\text{MC}}\left[\hat{V}\left(\hat{\theta}\right)\right] - V_{\text{MC}}\left(\hat{\theta}\right)}{V_{\text{MC}}, \left(\hat{\theta}\right)}, \tag{5.7}$$

où $E_{\text{MC}}\left[\hat{V}\left(\hat{\theta}\right)\right]$ est la moyenne MC des estimateurs de variance et $V_{\text{MC}}\left(\hat{\theta}\right)$ est la variance MC de l'estimateur de l'IRQ. Le biais relatif de MC de l'estimateur de variance pour l'estimateur de l'IRQ se situe entre -6 % et -1 %. Les couvertures empiriques des intervalles de confiance théoriques normaux se situent à moins de 1 % du niveau nominal de 95 %.

Tableau 5.1
Propriétés MC des estimateurs et des estimateurs de variance pour la simulation avec échantillonnage PPT avec remise. Pct. Rel. EQM (5.4) : Différence entre la variance MC de l'estimateur de l'IFP ou de l'IPN et l'EQM de MC de l'estimateur de l'IRQ, par rapport à l'EQM de MC de l'estimateur de l'IRQ. Pct. Rel. Var. (5.5) : Différence entre la variance MC de l'estimateur de l'IFP ou de l'IPN et l'EQM de MC de l'estimateur de l'IRQ, par rapport à l'EQM de MC de l'estimateur de l'IRQ. Pct. Biais (5.6) : Pourcentage de l'EQM de MC des estimateurs de l'IFP, de l'INP et de l'IRQ en raison du biais de MC au carré. Biais Rel = Biais relatif de MC de l'estimateur de variance défini en (5.7). Couverture = Couverture de MC des intervalles de confiance à 95 %

	Pct. Rel. EQM		Pct. Rel. Var.		Pct. Biais			Biais Rel.	Couverture
	IPN	IFP	IPN	IFP	IPN	IFP	IRQ	IRQ	IRQ
θ_1	0,509	1,624	0,211	1,589	0,304	0,041	0,006	-2,386	0,945
θ_2	3,308	1,882	1,011	-0,151	2,225	1,998	0,002	-1,113	0,951
θ_3	1,518	5,449	0,979	2,605	0,840	2,999	0,311	-5,772	0,943
θ_4	515,980	26,752	10,501	12,415	82,101	11,508	0,222	-3,182	0,952
θ_5	5,879	61,416	5,659	-2,345	0,223	39,510	0,015	–	–

6 Discussion

L'IRQ est élaborée pour un contexte d'enquête complexe. D'autres choix de poids sont examinés, et un estimateur de variance à forme fermée est fourni basé sur une approximation linéaire. La cohérence et la normalité asymptotique des estimateurs sont démontrées dans le cadre d'un nombre infini de valeurs imputées. Dans les simulations conçues pour représenter les données du CEAP, l'estimateur de variance fondé sur la distribution asymptotique a un biais relatif inférieur à 6 % en valeur absolue et conduit à des intervalles de confiance dont la couverture est proche du niveau nominal pour J fini. De plus, l'estimateur basé sur l'IRQ est plus efficace qu'un estimateur basé sur l'IFP ou l'IPN, parce que l'IRQ offre un compromis raisonnable entre le biais et la variance.

La procédure d'imputation par régression quantile est appliquée pour estimer l'érosion moyenne dans sept États du Midwest des États-Unis à l'aide des données du CEAP. L'analyse démontre que l'IRQ présente une solution de rechange viable aux ajustements de pondération actuellement utilisés pour tenir compte de la non-réponse dans le CEAP.

Les points à améliorer pour l'IRQ comprennent le choix de τ_j, le choix de b_i, le raffinement de l'estimation des courbes de quantiles et l'estimation de la variance pour les fonctions $\mathbf{g}(\cdot)$ non différenciables. L'élaboration de méthodes automatisées pour sélectionner les paramètres de nuisance, qui sont appropriés pour la sélection de plusieurs quantiles dans un contexte d'enquête complexe, est un domaine de recherche future. L'estimation des courbes quantiles, soumises à une restriction selon laquelle les courbes estimées ne se chevauchent pas, pourrait améliorer l'estimation des dérivées nécessaires pour l'estimateur de variance. La section E du supplément en ligne https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg et Yu, 2016) présente les points à améliorer.

Remerciements

Ces travaux ont été soutenus par l'accord de coopération nº 68-3A75-4-122 entre le *Natural Resources Conservation Service* de l'USDA et le *Center for Survey Statistics and Methodology* de l'*Iowa State University*.

Bibliographie

Andridge, R.R., et Little, R.J.A. (2010). A review of hot deck imputation for survey nonresponse. *Revue Internationale de Statistique*, 78, 40-64.

Barrow, D.L., et Smith, P.W. (1978). Asymptotic properties of the best $L_2[0, 1]$ approximation by Splines with variable knots. *Quaterly of Applied Mathematics*, 33, 293-304.

Berg, E.J., et Yu, C. (2016). Supplement to "Semiparametric quantile regression imputation for a complex survey with application to the conservation effects assessment project". Accessible à l'adresse : https://github.com/ emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf.

Berg, E.J., Kim, J.K. et Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4, 436-462.

Breidt, F.J., et Fuller, W.A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 391-403.

Brick, J.M., et Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Chen, S., et Yu, C. (2016). Parameter estimation through semiparametric quantile regression imputation. *Electronical Journal of Statistics*, 10, 3621-3647.

Clayton, D., Spiegelhalter, D., Dunn, G. et Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B*, 60, 71-87.

D'Arrigo, J., et Skinner, C. (2010). Estimation de la variance par linéarisation pour les estimateurs par calage généralisé en présence de non-réponse. *Techniques d'enquête*, 36, 2, 197-209. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2010002/article/11380-fra.pdf.

De Boor, C. (2001). *A Practical Guide to Splines* (édition révisée), New York: Springer-Verlag.

Fuller, W.A. (1996). *Introduction to Statistical Time Series: Second Edition*. New York: John Wiley & Sons, Inc.

Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Fuller, W.A. (2009b). *Sampling Statistics*. New York: John Wiley & Sons, Inc. Vol. 560.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer. Vol. 2, No. 1.

Isaki, T.C., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Jang, W., et Wang, J.H. (2015). A semiparametric Bayesian approach for joint-quantile regression with clustered data. *Computational Statistics and Data Analysis*, 84, 99-115.

Kim, J.K., et Park, M. (2010). Calibration estimation in survey sampling. *Revue Internationale de Statistique*, 78, 21-39.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1), 119-132.

Kim, J.K., et Riddles, M.K. (2012). Théorie concernant les estimateurs ajustés sur le score de propension dans les sondages. *Techniques d'enquête*, 38, 2, 171-180. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2012002/article/11754-fra.pdf.

Kim, J.K., et Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*, Chapman and Hall/CRC, Boca Raton.

Koenker, R., et Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33-50.

Koenker, R. (2005). *Quantile Regression*. Cambridge university press. No. 38.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9547-fra.pdf.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data missing values. *Applied Statistics*, 37, 23-38.

Mealli, F., et Rubin, D. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102, 995-1000.

Nusser, S.M., et Goebel, J.J. (1997). The National Resources Inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4(3), 181-204.

Nusser, S.M. (2006). National Resources Inventory (NRI), US. *Encyclopedia of Environmetrics Second Edition*, 1-3.

Nusser, S.M., Carriquiry, A.L., Dodd, K.W. et Fuller, W.A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91(436), 1440-1449.

Pakes, A., et Pollard, D. (1989). Simulation and the asymptotic of optimization estimators. *Econometrica,* 57(4), 1027-1057.

Pfeffermann, D. (2011). Modélisation des données d'enquêtes complexes : Pourquoi les modéliser ? Pourquoi est-ce un problème ? Comment le résoudre ? *Techniques d'enquête*, 37, 2, 123-146. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11602-fra.pdf.

Reiter, J.P., Raghunathan, T.E. et Kinney, S.K. (2006). L'importance de la modélisation du plan d'échantillonnage dans l'imputation multiple pour les données manquantes. *Techniques d'enquête*, 32, 2, 161-168. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9548-fra.pdf.

Robins, J.M., et Wang, N. (2000). Inference for imputation estimators. *Biometrika*. 87, 113-124.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc. Vol. 81.

Sheather, S.J., et Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

Wang, D., et Chen, S.X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 490-517.

Wei, Y., Ma, Y. et Carroll, R.J. (2012). Multiple imputation in quantile regression. *Biometrika*, 99, 423-438.

Yoshida, T. (2013). Asymptotics for penalized spline estimators in quantile regression. *Communications in Statistics - Theory and Methods*, DOI 10.1080/03610926.2013.765477.