N° 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête 41-1

Estimation sur petits domaines en combinant des données provenant de plusieurs sources

par Jae-kwang Kim, Seunghwan Park et Seo-young Kim

Date de diffusion : le 29 juin 2015

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 infostats@statcan.gc.ca

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

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 « À propos de nous » > « Notre organisme » > « Offrir des services aux Canadiens »

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.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence ($p<0,05$)

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, juin 2015
Vol. 41, n° 1, pp. 21-37
Statistique Canada, n° 12-001-X au catalogue

Estimation sur petits domaines en combinant des données provenant de plusieurs sources

Jae-kwang Kim, Seunghwan Park et Seo-young Kim[1]

Résumé

Une approche basée sur un modèle au niveau du domaine pour combiner des données provenant de plusieurs sources est examinée dans le contexte de l'estimation sur petits domaines. Pour chaque petit domaine, plusieurs estimations sont calculées et reliées au moyen d'un système de modèles d'erreur structurels. Le meilleur prédicteur linéaire sans biais du paramètre de petit domaine peut être calculé par la méthode des moindres carrés généralisés. Les paramètres des modèles d'erreur structurels sont estimés en s'appuyant sur la théorie des modèles d'erreur de mesure. L'estimation des erreurs quadratiques moyennes est également discutée. La méthode proposée est appliquée au problème réel des enquêtes sur la population active en Corée.

Mots-clés : Modèle au niveau du domaine; information auxiliaire; modèles d'erreur de mesure; modèle d'erreur structurel; intégration des enquêtes.

1 Introduction

Combiner des données provenant de diverses sources est un problème important en statistique. Dans le contexte des sondages, combiner les données de plusieurs enquêtes peut améliorer la qualité des estimations sur petits domaines. Les données peuvent provenir d'un échantillon probabiliste sur lequel sont faites des mesures directes, d'un autre échantillon probabiliste sur lequel sont faites des mesures indirectes (comme l'état de santé autodéclaré), ou d'information auxiliaire au niveau du domaine. Bon nombre d'approches de combinaison de données, telles que les méthodes à bases de sondage multiples et les méthodes d'appariement statistique, requièrent l'accès à des données au niveau individuel, ce qui n'est pas toujours possible en pratique.

Nous considérons une approche de l'estimation sur petits domaines basée sur un modèle au niveau du domaine lorsqu'il existe plusieurs sources d'information auxiliaire. Pfeffermann (2002) et Rao (2003) ont procédé à une recension détaillée des méthodes utilisées en estimation sur petits domaines. Lohr et Prasad (2003) ont utilisé des modèles multivariés pour combiner l'information provenant de plusieurs enquêtes. Ybarra et Lohr (2008) ont considéré le problème de l'estimation sur petits domaines quand les données auxiliaires au niveau du domaine contiennent des erreurs de mesure. Merkouris (2010) a discuté de l'estimation sur petits domaines lorsque l'on combine des données provenant de plusieurs enquêtes. Raghunathan, Xie, Schenker, Parsons, Davis, Dodd et Feuer (2007), ainsi que Manzi, Spiegelhalter, Turner, Flowers et Thompson (2011) se sont servi de modèles hiérarchiques bayésiens pour combiner les données provenant de plusieurs enquêtes pour l'estimation sur petits domaines. Kim et Rao (2012) ont examiné une approche fondée sur le plan de sondage pour combiner les données provenant de deux enquêtes indépendantes.

Afin de décrire la situation, supposons que la population finie est constituée de H sous-populations, désignées par $U_1, \ldots, U_H$, et que nous souhaitons estimer les totaux de sous-population $X_h = \sum_{i \in U_h} x_i$

1. Jae-kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, É.-U.; Seunghwan Park, Department of Statistics, Seoul National University, Seoul, 151-747, Corée. Courriel : kkampsh@gmail.com; Seo-young Kim, Statistical Research Institute, Statistics Korea, Daejon, 302-847, Corée.

d'une variable x pour chaque domaine h. Nous supposons qu'il existe une enquête conçue pour mesurer x_i à partir de l'échantillon, mais que la taille de cet échantillon n'est pas suffisamment grande pour obtenir des estimations de X_h d'une précision raisonnable. Considérons l'une des enquêtes, appelée enquête A, comme étant l'enquête principale, et soit $\hat{X}_h$ un estimateur convergent sous le plan de X_h obtenu à partir de l'enquête A. Souvent, nous calculons $\hat{X}_h = \sum_{i \in A_h} w_{ia} x_i$, où A_h est le jeu d'unités de l'échantillon A pour la sous-population h et w_{ia} est le poids de l'unité i dans l'échantillon A.

En plus de l'enquête principale, supposons qu'il en existe une autre, appelée enquête B, donnant une mesure qui est une estimation grossière de x_i. Soit y_{1i} la mesure prise au moyen de l'enquête B. Nous pouvons supposer que y_{1i} est une mesure grossière de x_i présentant un certain niveau d'erreur de mesure. Donc, nous pouvons émettre l'hypothèse que

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \tag{1.1}$$

pour certains paramètres (β_0, β_1), où $e_{1i} \sim (0, \sigma_{e1}^2)$. Le modèle (1.1) étant propre à la variable, l'hypothèse de régression linéaire ou les hypothèses de variance égale peuvent être relâchées plus tard. Si $(\beta_0, \beta_1) = (0, 1)$, alors le modèle (1.1) signifie qu'il n'y a pas de biais de mesure. Notons que, dans (1.1), les paramètres du modèle (β_0, β_1) ne sont pas propres au domaine, mais peuvent différer pour des groupes de domaines, comme il est démontré dans l'application à l'enquête coréenne sur la population active présentée à la section 5. La spécification de modèles de régression distincts pour différents groupes peut donner lieu à de plus petites erreurs de modélisation et donc accroître l'efficacité statistique de la méthode proposée. Partant de l'enquête B, nous pouvons obtenir un autre estimateur $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ de X_h, où w_{ib} est le poids de l'unité i dans l'échantillon de l'enquête B, et B_h est l'échantillon B pour la sous-population h. Notons que l'on peut obtenir $\hat{Y}_{1h}$ pour chaque domaine, si les mêmes domaines sont définis dans les deux enquêtes A et B. Le modèle (1.1) peut être utilisé pour combiner l'information provenant des deux enquêtes.

Enfin, les données de recensement peuvent représenter une autre source d'information. Les données de recensement ne souffrent pas d'une erreur de couverture ni d'une erreur d'échantillonnage. Toutefois, elles peuvent présenter des erreurs de mesure et ne fournissent pas d'information mise à jour pour chaque mois ou chaque année. Soit y_{2i} la mesure de l'unité i d'après le recensement. Le total de sous-population $Y_{2h} = \sum_{i \in C_h} y_{2i}$ est disponible quand C_h est le jeu d'unités du recensement C pour la sous-population h.

Le tableau 1.1 résume les principales sources d'information que nous pouvons prendre en considération dans l'estimation sur petits domaines.

Tableau 1.1
Information disponible pour l'estimation sur petits domaines

Données	Observation	Estimation au niveau du domaine	Propriétés
Enquête A	Observation directe (x_i)	$\hat{X}_h, \hat{V}(\hat{X}_h)$	Erreur d'échantillonnage (grande)
Enquête B	Observation auxiliaire (y_{1i})	$\hat{Y}_{1h}, \hat{V}(\hat{Y}_{1h})$	Biais Erreur de mesure Erreur d'échantillonnage
Recensement	Observation auxiliaire (y_{2i})	Y_{2h}	Erreur de mesure Pas d'information mise à jour

Dans le présent article, nous considérons une approche d'estimation sur petits domaines au moyen d'un modèle au niveau du domaine combinant toute l'information disponible. L'approche proposée est basée sur les modèles d'erreur de mesure, dans lesquels les erreurs d'échantillonnage des estimateurs directs sont traitées comme des erreurs de mesure, et toutes les autres données auxiliaires sont combinées au moyen d'un ensemble de modèles de lien. L'approche proposée est appliquée au problème de l'estimation sur petits domaines dans le cas des enquêtes sur la population active en Corée, où trois estimations sont combinées pour produire des estimations sur petits domaines des taux de chômage.

La présentation de l'article est la suivante. À la section 2, nous exposons la théorie de base et nous envisageons le problème d'estimation sur petits domaines comme un problème de prédiction d'un modèle d'erreur de mesure. À la section 3, nous discutons de l'estimation des paramètres du modèle d'estimation sur petits domaines au niveau du domaine. À la section 4, nous décrivons brièvement l'estimation de l'erreur quadratique moyenne. À la section 5, nous appliquons la méthode proposée aux données de l'enquête sur la population active en Corée. Enfin, à la section 6, nous présentons nos conclusions.

2 Théorie de base

À la présente section, nous commençons par présenter la théorie de base qui sous-tend la combinaison de l'information pour l'estimation sur petits domaines. Nous examinons d'abord le cas simple de la combinaison de deux enquêtes. Supposons qu'il existe deux enquêtes, A et B, réalisées selon deux plans d'échantillonnage probabiliste distincts. Les deux enquêtes ne sont pas forcément indépendantes. À partir de l'enquête A, nous obtenons un estimateur sans biais sous le plan $\hat{X}_{h,a} = \sum_{i \in A_h} w_{ia} x_i$ et l'estimateur de sa variance $\hat{V}\left(\hat{X}_h\right)$. À partir de l'enquête B, nous obtenons un estimateur sans biais sous le plan $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ de $Y_{1h} = \sum_{i \in U_h} y_{1i}$. L'erreur d'échantillonnage de $\left(\hat{X}_h, \hat{Y}_{1h}\right)$ peut être exprimée par le *modèle d'erreur d'échantillonnage*

$$\begin{pmatrix} \hat{X}_h \\ \hat{Y}_{1h} \end{pmatrix} = \begin{pmatrix} X_h \\ Y_{1h} \end{pmatrix} + \begin{pmatrix} N_h a_h \\ N_h b_h \end{pmatrix} \tag{2.1}$$

et a_h et b_h représentent les erreurs d'échantillonnage associées à $\hat{X}_h / N_h$ et à $\hat{Y}_{1h} / N_h$ telles que

$$\begin{pmatrix} a_h \\ b_h \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V(a_h) & \mathrm{Cov}(a_h, b_h) \\ \mathrm{Cov}(a_h, b_h) & V(b_h) \end{pmatrix} \right].$$

Le paramètre d'intérêt est le total de population X_h de x dans le domaine h.

Partant de (1.1), nous obtenons le modèle au niveau du domaine qui suit :

$$Y_{1h} = N_h \beta_0 + \beta_1 X_h + \tilde{e}_{1h}, \tag{2.2}$$

où $(N_h, X_h, Y_{1h}, \tilde{e}_{1h}) = \sum_{i \in U_h} (1, x_i, y_{1i}, e_{1i})$. Nous pouvons exprimer (2.2) en fonction de la moyenne de population

$$\bar{Y}_{1h} = \beta_0 + \bar{X}_h \beta_1 + \bar{e}_{1h}, \tag{2.3}$$

où $(\bar{X}_h, \bar{Y}_{1h}, \bar{e}_{1h}) = N_h^{-1} \sum_{i \in U_h} (x_i, y_{1i}, e_{1i})$. Si nous utilisons un modèle d'erreurs emboîtées

$$e_{1hi} = \varepsilon_h + u_{hi} \tag{2.4}$$

où $\varepsilon_h \sim (0, \sigma_e^2)$ et $u_{hi} \sim (0, \sigma_u^2)$, alors $\bar{e}_{1h} \sim (0, \sigma_{e,h}^2), \sigma_{e,h}^2 = \sigma_e^2 + \sigma_u^2 / N_h$. Le modèle d'erreurs emboîtées, dont l'usage est assez fréquent en estimation sur petits domaines (par exemple, Battese, Harter et Fuller 1988), repose sur l'hypothèse que $\mathrm{Cov}(e_{1hi}, e_{1hj}) = \sigma_e^2$ pour $i \neq j$. Comme N_h est souvent assez grand, nous pouvons supposer sans risque que $\bar{e}_{1h} \sim (0, \sigma_{e,h}^2 = \sigma_e^2)$. Le modèle (2.2) est appelé *modèle d'erreur structurel* parce qu'il décrit la relation structurelle entre les deux variables latentes Y_{1h} et X_h. Les deux modèles, (2.1) et (2.2), sont souvent mentionnés dans la littérature traitant des modèles d'erreur de mesure (Fuller 1987). Donc, le modèle pour l'estimation sur petits domaines peut être considéré comme un modèle d'erreur de mesure, comme l'a suggéré Fuller (1991) qui a été le premier à utiliser l'approche du modèle d'erreur de mesure dans la modélisation au niveau de l'unité pour l'estimation sur petits domaines.

Maintenant, si nous définissons $(\bar{y}_{1h}, \bar{x}_h) = N_h^{-1} (\hat{Y}_{1h}, \hat{X}_h)$, en combinant (2.1) et (2.3), nous obtenons

$$\begin{pmatrix} \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \bar{X}_h \end{pmatrix} + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

qui peut également s'écrire sous la forme

$$\begin{pmatrix} \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}. \tag{2.5}$$

Donc, quand tous les paramètres du modèle (2.5) sont connus, le meilleur estimateur de $\bar{X}_h$ peut être calculé par

$$\hat{\bar{X}}_h = \left\{ (\beta_1, 1) V_h^{-1} (\beta_1, 1)' \right\}^{-1} (\beta_1, 1) V_h^{-1} (\bar{y}_{1h} - \beta_0, \bar{x}_h)' \tag{2.6}$$

où V_h est la matrice de variance-covariance de $(b_h + \bar{e}_{1h}, a_h)'$. La variance de $\hat{\bar{X}}_h$ est donnée par $\left\{ (\beta_1, 1) V_h^{-1} (\beta_1, 1)' \right\}^{-1}$. L'estimateur en (2.6) peut être appelé estimateur par les moindres carrés généralisés (MCG), parce qu'il s'appuie sur la méthode des moindres carrés généralisés de la théorie des modèles linéaires. La méthode MCG est utile parce qu'elle est optimale et qu'elle permet d'incorporer naturellement des sources d'information supplémentaires. Par exemple, si un autre estimateur $\bar{y}_{2h}$ de $\bar{Y}_{2h}$ est également disponible et satisfait

$$\bar{Y}_{2h} = \gamma_0 + \gamma_1 \bar{X}_h + \bar{e}_{2h}$$

et

$$\bar{y}_{2h} = \bar{Y}_{2h} + c_h,$$

alors le modèle MCG étendu s'écrit

$$\begin{pmatrix} \bar{y}_{2h} - \gamma_0 \\ \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} c_h + \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \tag{2.7}$$

et l'estimateur MCG peut être obtenu par

$$\hat{\bar{X}}_{h2} = \left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1} (\gamma_1, \beta_1, 1) V_{h2}^{-1} (\bar{y}_{2h} - \gamma_0, \bar{y}_{1h} - \beta_0, \bar{x}_h)'$$

où V_{h2} est la matrice de variance-covariance de $(c_h + \bar{e}_{2h}, b_h + \bar{e}_{1h}, a_h)'$. La variance de l'estimateur MCG est $\left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1}$. Si $\bar{y}_{2h}$ est indépendant de $(\bar{x}_h, \bar{y}_{1h})$, le gain d'efficacité, en termes de variance relative, qui découle de l'incorporation de $\bar{y}_{2h}$ dans l'estimateur MCG peut s'exprimer sous la forme

$$\frac{V\left(\hat{\bar{X}}_{h2}\right) - V\left(\hat{\bar{X}}_h\right)}{V\left(\hat{\bar{X}}_h\right)} = -\frac{\{V(\bar{y}_{2h}/\gamma_1)\}^{-1}}{\left\{V\left(\hat{\bar{X}}_h\right)\right\}^{-1} + \{V(\bar{y}_{2h}/\gamma_1)\}^{-1}},$$

où $V(\bar{y}_{2h}/\gamma_1) = V(c_h + \bar{e}_{2h})/\gamma_1^2$. Le gain est important si la variance d'échantillonnage de $\bar{y}_{2h}$ ainsi que la variance du modèle $V(\bar{e}_{2h})$ sont faibles. Si $\gamma_1 = 0$, alors le gain est nul.

Remarque 1 *Notons que le modèle (2.5) peut également s'écrire*

$$\begin{pmatrix} \beta_1^{-1}(\bar{y}_{1h} - \beta_0) \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} (b_h + \bar{e}_{1h})/\beta_1 \\ a_h \end{pmatrix}. \tag{2.8}$$

L'estimateur MCG obtenu à partir de (2.8), qui est le même que l'estimateur MCG obtenu à partir de (2.5), peut être exprimé sous la forme

$$\hat{\bar{X}}_h = \alpha_h \bar{x}_h + (1 - \alpha_h)\tilde{x}_h \tag{2.9}$$

où $\tilde{x}_h = \beta_1^{-1}(\bar{y}_{1h} - \beta_0)$ *et*

$$\begin{aligned} \alpha_h &= \frac{V(\tilde{x}_h) - \mathrm{Cov}(\bar{x}_h, \tilde{x}_h)}{V(\bar{x}_h) + V(\tilde{x}_h) - 2\mathrm{Cov}(\bar{x}_h, \tilde{x}_h)} \\ &= \frac{\sigma_{e,h}^2 + V(b_h) - \beta_1 \mathrm{Cov}(a_h, b_h)}{\sigma_{e,h}^2 + V(b_h) + \beta_1^2 V(a_h) - 2\beta_1 \mathrm{Cov}(a_h, b_h)}, \end{aligned}$$

L'estimateur $\tilde{x}_h$, *lorsqu'il est calculé en utilisant le paramètre estimé* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$, *est appelé estimateur synthétique, et l'estimateur optimal en (2.9) est souvent appelé estimateur composite. On peut montrer qu'en ignorant l'effet de l'estimation de* β, *la variance de l'estimateur composite est égale à*

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h) \tag{2.10}$$

et, comme $\alpha_h < 1$, *l'estimateur composite est plus efficace que l'estimateur direct.*

3 Estimation des paramètres

Maintenant, nous discutons de l'estimation des paramètres du modèle (2.3). L'estimateur MCG de $\beta = (\beta_0, \beta_1)$ peut être obtenu par minimisation de

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} \frac{(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2}{V(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)}. \tag{3.1}$$

Puisque

$$V(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1) = \sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)', \tag{3.2}$$

où $\sigma_{e,h}^2 = V(\bar{e}_{1h})$ et $\Sigma_h = V\{(a_h, b_h)'\}$, nous pouvons écrire

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} w_h(\beta_1)(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2, \tag{3.3}$$

où $w_h(\beta_1) = \left\{\sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)'\right\}^{-1}$. Maintenant, en résolvant $\partial Q^* / \partial \beta = 0$, nous obtenons

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w \tag{3.4}$$

et

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{(\bar{x}_h - \bar{x}_w)(\bar{y}_{1h} - \bar{y}_{1w}) - C(a_h, b_h)\}}{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{(\bar{x}_h - \bar{x}_w)^2 - V(a_h)\}}, \tag{3.5}$$

où

$$(\bar{x}_w, \bar{y}_w) = \left\{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\right\}^{-1} \sum_{h=1}^{H} w_h(\hat{\beta}_1)(\bar{x}_h, \bar{y}_h).$$

Notons que le poids $w_h(\beta_1)$ dépend de β_1. Donc, la solution (3.5) peut être obtenue à l'aide d'un algorithme itératif. Après avoir calculé $\hat{\beta}_1$ en utilisant (3.5), on obtient $\hat{\beta}_0$ en utilisant (3.4).

Passons maintenant à l'estimation de la variance du modèle $\sigma_{e,h}^2$. La méthode la plus simple est la méthode des moments (MOM). Autrement dit, nous pouvons utiliser

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_{e,h}^2 \tag{3.6}$$

pour obtenir un estimateur sans biais de $\sigma_{e,h}^2$. Sous le modèle des erreurs emboîtées donné par (2.4), nous avons $\sigma_{e,h}^2 = \sigma_e^2$ et

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_e^2. \tag{3.7}$$

Donc, comme dans Fuller (2009), l'estimateur MOM de σ_e^2 peut être exprimé par

$$\hat{\sigma}_e^2 = \sum_{h=1}^{H} \kappa_h \left\{ \left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h \hat{\beta}_1 \right)^2 - \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right) \right\} \tag{3.8}$$

où

$$\kappa_h \propto \left\{ \hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right) \right\}^{-1}$$

et $\sum_{h=1}^{H} \kappa_h = 1$. Comme κ_h dépend de $\hat{\sigma}_e^2$, la solution (3.8) peut être obtenue itérativement, en utilisant $\hat{\sigma}_e^2 = 0$ comme valeur initiale. Fay et Herriot (1979) ont utilisé une autre méthode qui est fondée sur la solution itérative de l'équation non linéaire :

$$\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h \right)^2}{\sigma_e^2 + \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right)'} = H - 2.$$

En écrivant l'équation susmentionnée sous la forme $g\left(\sigma_e^2\right) = H - 2$, une méthode de type Newton pour $g(\theta) = 0$ avec $\theta = \sigma_e^2$ peut être obtenue par

$$\theta^{(t+1)} = \theta^{(t)} + \frac{1}{g'\left(\theta^{(t)}\right)} \left(H - 2 - g\left(\theta^{(t)}\right) \right) \tag{3.9}$$

où

$$g'(\theta) = -\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h \right)^2}{\left\{ \theta + \left(-\hat{\beta}_1, 1 \right) \Sigma_h \left(-\hat{\beta}_1, 1 \right)' \right\}^2}.$$

En supposant que $\sigma_{e,h}^2 \equiv \sigma_e^2$, nous décrivons maintenant la procédure complète d'estimation des paramètres comme il suit :

Étape 1 Calculer l'estimateur initial de (β_0, β_1) en posant que $\hat{\sigma}_e^2 = 0$ dans (3.4) et (3.5).

Étape 2 En se basant sur la valeur courante de $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, calculer $\hat{\sigma}_e^2$ en utilisant l'algorithme itératif en (3.9).

Étape 3 Utiliser la valeur courante de $\hat{\sigma}_e^2$, calculer l'estimateur mis à jour de (β_0, β_1) au moyen de (3.4) et (3.5).

Étape 4 Répéter [Étape 2]-[Étape 3] jusqu'à la convergence.

La méthode d'estimation des paramètres proposée comprend l'estimation de $\beta = (\beta_0, \beta_1)$ par les MCG et l'estimation de σ_e^2 par les MOM itérativement. Notons que l'estimation de β est fondée sur des données provenant de tous les domaines. Si des modèles de régression distincts sont utilisés, la méthode d'estimation des paramètres proposée peut être appliquée à des groupes de domaines. Au lieu de cette

méthode d'estimation itérative distincte, nous pouvons également considérer une autre méthode fondée sur l'estimation du maximum de vraisemblance (EMV) sous des hypothèses distributionnelles paramétriques. Voir Carroll, Rupert et Stefanski (1995) et Schafer (2001) pour une discussion de l'EMV pour les paramètres des modèles d'erreur de mesure.

Remarque 2 *Si l'égalité* $\sigma_{e,h}^2 = \sigma_e^2$ *n'est pas vérifiée, nous pouvons considérer un modèle de rechange tel que*

$$\bar{e}_h \sim \left(0, \bar{X}_h \sigma_e^2\right). \tag{3.10}$$

Pour vérifier si le modèle (3.10) tient, on peut calculer

$$\nu_h = \left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h \hat{\beta}_1\right)^2 - \hat{\beta}_1^2 V\left(a_h\right) + 2\hat{\beta}_1 \hat{C}\left(a_h, b_h\right) - V\left(b_h\right) \tag{3.11}$$

et représenter graphiquement ν_h *en fonction de* $\bar{x}_h$. *Si le graphique montre une relation linéaire, alors (3.10) peut être traité comme un modèle raisonnable. Sous le modèle (3.10), nous pouvons obtenir* σ_e^2 *par une méthode du ratio :*

$$\hat{\sigma}_e^2 = \frac{\sum_{h=1}^{H} \kappa_h \nu_h}{\sum_{h=1}^{H} \kappa_h \hat{\bar{X}}_h} \tag{3.12}$$

où

$$\kappa_h \propto \left\{\hat{\bar{X}}_h \hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right)\right\}^{-1}$$

avec $\sum_{h=1}^{H} \kappa_h = 1$, $\hat{\bar{X}}_h$ *défini en (2.9), et* ν_h *défini en (3.11). Comme* κ_h *dépend aussi de* σ_e^2, *la solution (3.12) peut être obtenue par itération.*

Remarque 3 *Nous pouvons également considérer une transformation* $\bar{x}_h^* = T\left(\bar{x}_h\right)$ *et* $\bar{y}_{1h}^* = T\left(\bar{y}_{1h}\right)$ *afin d'améliorer l'approximation par une loi normale asymptotique. Pour vérifier l'écart par rapport à la normalité, nous représentons graphiquement* $n_{ha}\bar{V}\left(\bar{x}_h\right)$ *en fonction de* $\bar{x}_h$. *Si le graphique révèle une relation structurelle de* $\bar{x}_h$, *l'hypothèse de normalité peut être mise en doute. Maintenant, considérons la transformation suivante*

$$T\left(x\right) = \log\left(x\right). \tag{3.13}$$

Notons que la variance asymptotique de $\bar{x}_h^* = T\left(\bar{x}_h\right)$ *est égale à*

$$V\left(\bar{x}_h^*\right) \doteq \frac{1}{\left(\bar{x}_h\right)^2} V\left(\bar{x}_h\right).$$

Il s'agit d'une transformation stabilisant la variable qui est utile lorsque nous voulons améliorer l'approximation par la loi normale.

Après avoir obtenu l'estimateur MCG $\hat{\bar{X}}_h^*$ *de* $\bar{X}_h^*$, *nous devons appliquer la transformation inverse pour obtenir le meilleur estimateur de* $\bar{X}_h = T^{-1}\left(\bar{X}_h^*\right) := Q\left(\bar{X}_h^*\right)$. *La simple application de la transformation inverse donnera une estimation biaisée. Afin de corriger le biais, nous pouvons utiliser une linéarisation de Taylor d'ordre deux. En effectuant un développement en série de Taylor, nous obtenons*

$$Q\left(\hat{\bar{X}}_h^*\right) \doteq Q\left(\bar{X}_h^*\right) + Q'\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right) + \frac{1}{2}Q''\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right)^2$$

et donc, si nous utilisons $Q\left(\hat{\bar{X}}_h^*\right)$ *comme estimateur de* $\bar{X}_h = Q\left(\bar{X}_h^*\right)$, *nous obtenons, en laissant tomber les termes d'ordre plus faible,*

$$E\left\{Q\left(\hat{\bar{X}}_h^*\right)\right\} = \bar{X}_h + \frac{1}{2}Q''\left(\bar{X}_h^*\right)V\left(\hat{\bar{X}}_h^*\right).$$

Pour la transformation donnée par (3.13), nous avons $Q\left(\bar{X}_h^*\right) = \exp\left(\bar{X}_h^*\right)$ *et donc* $Q''\left(\bar{X}_h^*\right) = \bar{X}_h$. *Donc,* $\hat{\bar{X}}_h = Q\left(\hat{\bar{X}}_h^*\right)$, *et nous obtenons*

$$E\left(\hat{\bar{X}}_h\right) \cong \bar{X}_h + \frac{1}{2}\bar{X}_h V\left(\hat{\bar{X}}_h^*\right)$$

et l'estimateur de $\bar{X}_h$ *corrigé pour le biais est*

$$\hat{\bar{X}}_{h,bc} = \frac{\hat{\bar{X}}_h}{1 + 0,5V\left(\hat{\bar{X}}_h^*\right)}, \tag{3.14}$$

où $V\left(\hat{\bar{X}}_h^*\right)$ *est calculée par la méthode d'estimation de l'EQM dont nous discuterons à la section 4.*

4 Estimation de l'EQM

Passons maintenant à l'estimation de l'erreur quadratique moyenne (EQM) de l'estimateur MCG $\hat{\bar{X}}_h$ qui est donné par (2.9). Notons que l'estimateur MCG est une fonction de (β_0, β_1) et de σ_e^2. Si les paramètres du modèle sont connus, alors l'EQM de $\hat{\bar{X}}_h$ est égale à $M_{h1} = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\operatorname{Cov}(\bar{x}_h, \tilde{x}_h)$, comme il est discuté dans la remarque 1. Autrement dit, en écrivant $\theta = \left(\beta_0, \beta_1, \sigma_e^2\right)$ et $\hat{\bar{X}}_h = \hat{\bar{X}}_h(\theta)$, la prédiction réelle de $\bar{X}_h$ est calculée par $\hat{\bar{X}}_{eh} = \hat{\bar{X}}_h\left(\hat{\theta}\right)$. Afin de tenir compte de l'effet de l'estimation des paramètres du modèle, nous notons d'abord la décomposition qui suit de $\operatorname{EQM}\left(\hat{\bar{X}}_h^*\right)$:

$$\begin{aligned}\operatorname{EQM}\left(\hat{\bar{X}}_{eh}\right) &= \operatorname{EQM}\left(\hat{\bar{X}}_h\right) + E\left\{\left(\hat{\bar{X}}_{eh} - \hat{\bar{X}}_h\right)^2\right\} \\ &=: M_{h1} + M_{h2},\end{aligned}$$

qui a été prouvée pour la première fois par Kackar et Harville (1984) sous des hypothèses de normalité. Le premier terme, M_{h1}, est d'ordre $1/n_h$, où n_h est la taille de A_h, et le deuxième terme, M_{h2}, est d'ordre $1/n$ avec $n = \sum_{h=1}^{H} n_h$. Le deuxième terme est souvent beaucoup plus petit que le premier.

Nous considérons une approche jackknife pour estimer l'EQM. L'utilisation du jackknife pour obtenir une estimation corrigée pour le biais a été proposée au départ par Quenouille (1956). Jiang, Lahiri et Wan (2002) ont produit une justification rigoureuse de la méthode du jackknife pour l'estimation de l'EQM en estimation sur petits domaines. Les étapes qui suivent peuvent être utilisées pour le calcul du jackknife.

Étape 1 Calculer la k^{e} réplique $\hat{\theta}^{(-k)}$ de $\hat{\theta}$ en supprimant le k^{e} jeu de données de domaine $(\bar{x}_k, \bar{y}_{1k})$ du jeu de données complet $\{(\bar{x}_h, \bar{y}_{1h}); h = 1, 2, \ldots, H\}$. Ce calcul est effectué pour chaque k pour obtenir H répliques de θ : $\{\hat{\theta}^{(-k)}; k = 1, \ldots, H\}$ qui, à leur tour, fournissent H répliques de $\hat{\bar{X}}_h$: $\{\hat{\bar{X}}_h^{(-k)}; k = 1, 2, \ldots, H\}$, où $\hat{\bar{X}}_h^{(-k)} = \hat{\bar{X}}_h(\hat{\theta}^{(-k)})$.

Étape 2 Calculer l'estimateur de M_{h2} sous la forme

$$\hat{M}_{2h} = \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\bar{X}}_h^{(-k)} - \hat{\bar{X}}_h\right)^2. \tag{4.1}$$

Étape 3 Calculer l'estimateur de M_{h1} sous la forme

$$\hat{M}_{1h} = \hat{\alpha}_h^{(\text{JK})} V(\bar{x}_h) + \left(1 - \hat{\alpha}_h^{(\text{JK})}\right)\text{Cov}(\bar{x}_h, \tilde{x}_h) \tag{4.2}$$

où $\hat{\alpha}_h^{(\text{JK})}$ est un estimateur de α_h corrigé pour le biais donné par

$$\hat{\alpha}_h^{(\text{JK})} = \hat{\alpha}_h - \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\alpha}_h^{(-k)} - \hat{\alpha}_h\right),$$

$$\hat{\alpha}_h = \frac{\hat{\sigma}_e^2 + V(b_h) - \hat{\beta}_1\text{Cov}(a_h, b_h)}{\hat{\sigma}_e^2 + V(b_h) + \hat{\beta}_1^2 V(a_h) - 2\hat{\beta}_1\text{Cov}(a_h, b_h)},$$

et

$$\hat{\alpha}_h^{(-k)} = \frac{\hat{\sigma}_e^{(-k)2} + V(b_h) - \hat{\beta}_1^{(-k)}\text{Cov}(a_h, b_h)}{\hat{\sigma}_e^{(-k)2} + V(b_h) + \left(\hat{\beta}_1^{(-k)}\right)^2 V(a_h) - 2\hat{\beta}_1^{(-k)}\text{Cov}(a_h, b_h)}.$$

Remarque 4 *Pour la transformation donnée par (3.13), nous utilisons l'estimateur corrigé pour le biais (3.14) et la méthode d'estimation de son EQM doit être modifiée. En utilisant* $\hat{\bar{X}}_{eh,bc}$ *pour désigner l'estimateur corrigé pour le biais (3.14) évalué à* $\hat{\theta}$, *nous pouvons obtenir*

$$\begin{aligned}
\text{EQM}\left(\hat{\bar{X}}_{eh,bc}\right) &= \text{EQM}\left(\hat{\bar{X}}_{eh}\right) \\
&= \text{EQM}\left\{Q\left(\hat{\bar{X}}_{eh}^*\right)\right\} \\
&\cong \left\{Q'\left(\bar{X}_h^*\right)\right\}^2 \cdot \text{EQM}\left(\hat{\bar{X}}_{eh}^*\right) \\
&= \bar{X}_h^2 \cdot \text{EQM}\left(\hat{\bar{X}}_{eh}^*\right),
\end{aligned}$$

où la première égalité découle du fait que $\hat{\bar{X}}_{h,bc} - \hat{\bar{X}}_h$ *est d'ordre* $O_p\left(n_h^{-1}\right)$. *L'EQM de* $\hat{\bar{X}}_h^*$, *l'estimateur MCGE de* $\bar{X}_h^*$ *après transformation, est calculée au moyen de (4.1) et (4.2). Lorsque*

$\mathrm{EQM}\left(\hat{\bar{X}}^{*}_{eh}\right)$ *est estimée, nous devons la multiplier par* $\hat{\bar{X}}^{2}_{h}$ *pour obtenir l'estimateur de l'EQM de l'estimateur MCGE* $\hat{\bar{X}}_{eh,bc}$ *rétrotransformé.*

5 Application à l'Enquête sur la population active de la Corée

Nous examinons maintenant une application de la méthode proposée aux enquêtes sur la population active en Corée. Dans ce pays, deux enquêtes distinctes sur la population active sont utilisées pour obtenir des renseignements au sujet de l'emploi. L'une d'elles est l'Enquête sur la population active coréenne (PAC) et l'autre est l'Enquête sur la population active locale (PAL). L'enquête PAC est réalisée auprès d'un échantillon d'environ 7 000 ménages, tandis que l'enquête PAL est réalisée auprès d'un échantillon d'environ 200 000 ménages. Comme la PAL est une enquête à grande échelle faisant appel à un grand nombre d'intervieweurs à temps partiel, les données comportent un certain niveau d'erreurs de mesure. Nous supposons que l'enquête PAC est exempte d'erreur de mesure, quoiqu'elle présente d'importantes erreurs d'échantillonnage au niveau des petits domaines. L'échantillon de l'enquête PAC est un échantillon de deuxième phase tiré de l'échantillon de l'enquête PAL. Donc, les erreurs d'échantillonnage des estimations d'après les deux enquêtes sont corrélées. Soit $\bar{X}_h$ le taux de chômage (réel) dans le domaine h. Le niveau de petit domaine que nous considérons est appelé « Gu ». La Corée compte 229 « Gu ».

Nous observons $\bar{x}_h$ au moyen de l'enquête PAC et $\bar{y}_{1h}$ au moyen de l'enquête PAL. Pour construire des modèles de lien, nous commençons par diviser la population en deux régions, une région urbaine et une région rurale, en nous basant sur la proportion de ménages travaillant en agriculture. Nous spécifions des modèles distincts pour chaque région (même modèle mais en permettant des paramètres différents) et estimons les paramètres du modèle séparément. Le modèle structurel est

$$\bar{Y}_h = \beta_1 \bar{X}_h + e_h \tag{5.1}$$

avec $e_h \sim \left(0, \sigma_e^2\right)$. Ici, nous posons que $\beta_0 = 0$ pour garantir que l'estimateur MCG de $\bar{X}_h$ n'est pas négatif. Le modèle d'erreur d'échantillonnage reste le même. Dans ce cas, nous pouvons estimer β_1 comme il suit

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\left\{\bar{x}_h \bar{y}_{1h} - C\left(a_h, b_h\right)\right\}}{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\left\{\bar{x}_h^2 - V\left(a_h\right)\right\}}. \tag{5.2}$$

La variance d'échantillonnage de $\left(a_h, b_h\right)$ est calculée en utilisant la méthode d'échantillonnage à deux phases inverse décrite à l'annexe. La variance sous le modèle est estimée par la méthode des moments dans (3.8) avec $\hat{\beta}_0 = 0$. L'estimateur MCG peut être calculé en utilisant (2.9) avec $\tilde{x}_h = \hat{\beta}_1^{-1}\bar{y}_{1h}$.

En plus des deux enquêtes, nous pouvons aussi utiliser l'information provenant du recensement. Le modèle MCG intégrant les trois sources d'information peut être exprimé sous la forme

$$\begin{pmatrix} \bar{Y}_{2h} \\ \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

où $\bar{Y}_{2h}$ est le résultat du recensement pour le domaine h. Comme l'estimation d'après le recensement ne présente pas d'erreur d'échantillonnage, nous avons une seule erreur de modélisation e_{2h} qui représente l'erreur commise quand nous modélisons $E(\bar{Y}_{h2}) = \gamma_1 \bar{X}_h$. Les paramètres du modèle peuvent être obtenus en utilisant la méthode décrite à la section 3 avec $\Sigma_h = \text{diag}(0, V(a_h, b_h))$. L'estimateur MCG de $\bar{X}_h$ s'obtient facilement. L'EQM peut être calculée en utilisant le fait que

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix}' \left\{ V \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \right\}^{-1} \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} := M_{h1}$$

et en appliquant la méthode du jackknife pour corriger le biais.

La figure 5.1 donne le graphique du taux de chômage selon l'enquête PAC en fonction du taux de chômage selon l'enquête PAL pour les domaines urbains. La figure 5.1 montre qu'il existe une relation structurelle linéaire entre les estimations PAC et PAL. Au lieu du résidu habituel $\hat{\bar{e}}_h$ dans le modèle d'erreur structurel, nous utilisons $\hat{v}_h$ en tant que résidu dans le modèle de régression avec erreurs de mesure, où $\hat{v}_h = \bar{y}_{1h} - \hat{\beta}_1 \bar{x}_h$. La figure 5.2 donne le graphique de $\hat{v}_h$ en fonction de $\hat{\bar{X}}_h$ pour les domaines urbains. Le graphique montre que l'hypothèse de variance σ_e^2 égale est légèrement violée. Nous avons également considéré le modèle de variance hétéroscédastique décrit dans la remarque 2, mais les résultats n'ont pas varié de manière significative.

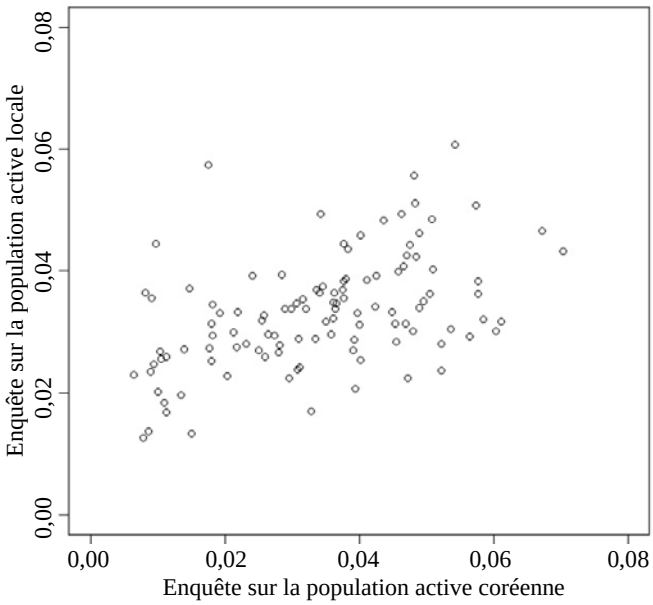

Figure 5.1 Graphique du taux de chômage selon les enquêtes PAC et PAL pour les domaines urbains.

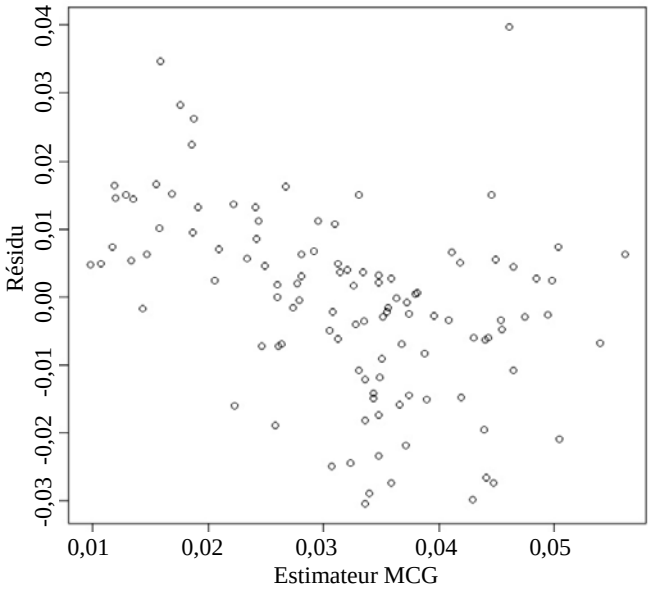

Figure 5.2 Graphique des résidus en fonction des valeurs estimées pour les domaines urbains.

Le tableau 5.1 donne les propriétés des estimations sur petits domaines en ce qui concerne l'EQM estimée. Nous avons examiné quatre estimateurs distincts de $\bar{X}_h$. PAC représente le résultat obtenu en utilisant les données de l'enquête sur la population active coréenne uniquement, PAL représente le résultat obtenu en utilisant les données de l'enquête sur la population active locale uniquement, MCG 1 représente le résultat obtenu en combinant les données des deux enquêtes PAC et PAL, et MCG 2 représente le résultat obtenu en combinant les données des enquêtes PAC et PAL et du recensement. Le tableau 5.1 montre que l'estimateur MCG 2 est celui qui donne les erreurs quadratiques moyennes les plus petites.

Tableau 5.1
Quartile de la performance des estimations sur petits domaines selon l'EQM pour les 229 domaines

EQM	1er Q	Médiane	3e Q	Moyenne
PAC	0,0000630	0,0001210	0,0002395	0,0002476
PAL	0,0001123	0,0001330	0,0001695	0,0001482
MCG 1	0,0000444	0,0000738	0,0001210	0,0000893
MCG 2	0,0000405	0,0000543	0,0000721	0,0000575

6 Conclusion

Le présent article décrit le traitement d'un problème d'estimation sur petits domaines comme un problème de prédiction d'un modèle d'erreur de mesure où les covariables, qui sont les estimations directes pour les petits domaines, sont sujettes à des erreurs d'échantillonnage. Dans notre approche du modèle d'erreur de mesure, les erreurs d'échantillonnage des estimateurs directs sont traitées comme des erreurs de mesure et le modèle d'erreur structurel peut être utilisé pour relier les autres estimations auxiliaires aux estimateurs directs. Le modèle proposé est en fait l'opposé du modèle d'Ybarra et Lohr

(2008), qui traitent l'estimateur direct comme une variable dépendante dans le modèle de régression et les estimations auxiliaires des erreurs non dues à l'échantillonnage comme des erreurs de mesure.

Dans notre approche, chaque estimation auxiliaire est traitée comme une variable dépendante dans le modèle de régression en utilisant l'estimation directe en tant que covariable et l'erreur d'échantillonnage de l'estimateur direct en tant qu'erreur de mesure. La variance de l'erreur de mesure est facile à estimer, parce qu'elle est essentiellement la variance d'échantillonnage de l'estimation directe. L'approche du modèle d'erreur de mesure est également très utile quand il existe plusieurs sources d'information auxiliaire au niveau des domaines. Contrairement à l'approche bayésienne, l'estimateur résultant ne s'appuie pas sur des hypothèses de modélisation paramétrique au sujet du modèle d'erreur structurel et reste optimal au sens de la minimisation des erreurs quadratiques moyennes parmi la classe d'estimateurs sans biais qui sont linéaires dans les données disponibles.

Dans l'exemple de l'application à l'enquête sur la population active de la Corée, deux estimations sur échantillon et l'information provenant du recensement sont utilisées pour calculer les estimations MCG des paramètres de petit domaine et les deux estimations sur échantillon sont corrélées en raison du plan d'échantillonnage à deux phases. Nous avons utilisé simplement des modèles de régression linéaire comme modèles de lien, principalement par souci de simplicité des calculs. Au lieu du modèle linéaire, on pourrait envisager un modèle linéaire généralisé afin d'améliorer le pouvoir de prédiction du modèle. Une telle extension ferait intervenir la théorie des modèles d'erreur de mesure non linéaires. Une étude plus approfondie de cette extension sera le sujet de futurs travaux de recherche.

Remerciements

Nous remercions un examinateur anonyme et le rédacteur associé de leurs commentaires constructifs. Les travaux de recherche du premier auteur ont été financés partiellement par l'entente de coopération NSF (MMS-121339).

Annexe

Échantillonnage à deux phases inverse

En échantillonnage à deux phases classique, l'échantillon de deuxième phase (A_2) est un sous-ensemble de l'échantillon de première phase (A_1). Nous considérons un autre type de plan d'échantillonnage possédant la structure inverse du plan d'échantillonnage à deux phases. Dans le plan d'échantillonnage à deux phases inverse, les étapes d'échantillonnage sont les suivantes :

Étape 1 À partir de la population finie, nous sélectionnons l'échantillon de première phase A_1 de taille n_1.

Étape 2 Dans l'échantillon de deuxième phase, nous sélectionnons A_2 à partir de $U - A_1$ de taille n_2. L'échantillon final A est constitué de A_1 et A_2. C'est-à-dire que $A = A_1 \cup A_2$ et $|A| = n = n_1 + n_2$.

L'échantillonnage à deux phases inverse est utilisé lorsqu'on augmente l'échantillon par une procédure d'échantillonnage additionnelle.

Pour discuter de l'estimation des paramètres sous échantillonnage à deux phases inverse, posons que $\pi_{1i} = \Pr(i \in A_1)$ est la probabilité d'inclusion d'ordre un pour A_1. Soit $\pi_{2i|1} = \Pr\left(i \in A_2 | A_1^c\right)$ la probabilité d'inclusion d'ordre un conditionnelle pour A_2 sachant $A_1^c = U - A_1$. Pour calculer la probabilité d'inclusion pour A, nous avons

$$\Pr(i \in A) = \Pr(i \in A_1) + \Pr\left(i \in A_2 | A_1^c\right) \Pr\left(i \in A_1^c\right).$$

Donc, nous pouvons utiliser $\pi_i = \pi_{1i} + (1 - \pi_{1i})\pi_{2i|1}$ pour calculer l'estimateur d'Horvitz-Thompson de la forme

$$\hat{Y}_{r,\mathrm{HT}} = \sum_{i \in A} \frac{1}{\pi_i} y_i. \tag{A.1}$$

Notons que, au lieu de (A.1), nous pouvons considérer la classe d'estimateurs suivante :

$$\hat{Y}_w = W \sum_{i \in A_1} \frac{1}{\pi_{1i}} y_i + (1 - W) \sum_{i \in A_2} \frac{1}{\pi_{2i|1}(1 - \pi_{1i})} y_i := W\hat{Y}_1 + (1 - W)\hat{Y}_2. \tag{A.2}$$

Puisque $\hat{Y}_1$ et $\hat{Y}_2$ sont tous deux sans biais pour Y, $\hat{Y}_w$ est également sans biais quel que soit le choix de W. Un choix raisonnable de W est $W = n_1/n$.

Sous échantillonnage aléatoire simple dans les deux plans, les deux estimateurs sont égaux à $\hat{Y} = N\bar{y}_n$, où $\bar{y}_n$ est la moyenne d'échantillon de y dans A. En écrivant $\bar{y}_1 = n_1^{-1} \sum_{i \in A_1} y_i$ et $\bar{y}_2 = \sum_{i \in A_2} y_i / n_2$, nous obtenons

$$\bar{y}_n = W\bar{y}_1 + (1 - W)\bar{y}_2 \tag{A.3}$$

où $W = n_1/n$. En utilisant

$$V(\bar{y}_1) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 \tag{A.4}$$

$$V(\bar{y}_2) = \left(\frac{1}{n_2} - \frac{1}{N}\right) S_y^2$$

$$\mathrm{Cov}(\bar{y}_1, \bar{y}_2) = \mathrm{Cov}\left(\bar{y}_1, \bar{y}_1^c\right) = -\frac{n_1}{N - n_1}\left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 = -\frac{1}{N} S_y^2,$$

où $\bar{y}_1^c = \sum_{i \in A_1^c} y_i / (N - n_1)$, nous obtenons, pour $W = n_1/n$,

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.5}$$

En outre,

$$\mathrm{Cov}(\bar{y}_1, \bar{y}_n) = \mathrm{Cov}\left[\bar{y}_1, W\bar{y}_1 + (1 - W)\bar{y}_2\right] = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.6}$$

Si l'égalité $W = n_1/n$ n'est pas vérifiée, alors (A.5) et (A.6) ne sont pas vérifiées.

Dans l'application à l'enquête sur la population active de la Corée à la section 5, puisque x et y mesurent le même item, nous pouvons supposer que $S_x^2 = S_y^2 = S_{xy}$ et la matrice de variance-covariance des erreurs d'échantillonnage peut être lissée sous la forme

$$V(a_h, b_h) = \begin{pmatrix} n_1^{-1} & n^{-1} \\ n^{-1} & n^{-1} \end{pmatrix} S_y^2.$$

Bibliographie

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,* 83, 28-36.

Carroll, R.J., Rupert, D. et Stefanski, L.A. (1995). *Measurement error in nonlinear models*. New York : Chapman & Hall.

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

Fuller, W.A. (1987). *Measurement error models*. New York : John Wiley & Sons, Inc.

Fuller, W.A. (1991). Small area estimation as a measurement error problem. Dans *Economic Models, Estimation, and Socioeconomic Systems: Essays in Honor of Karl A. Fox*, (Éds., Tij K. Kaul et Jati K. Sengupta), Elsevier Science Publishers, 333-352.

Fuller, W.A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc., Hoboken, NJ.

Jiang, J., Lahiri, P. et Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics,* 30, 1782-1810.

Kackar, R.N., et Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association,* 79, 853-862.

Kim, J.K., et Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika,* 99, 85-100.

Lohr, S.L., et Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data. *The Canadian Journal of Statistics,* 31, 383-396.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. et Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society A,* 174, 31-50.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society B,* 68, 509-521.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *Revue Internationale de Statistique*, 70, 125-144.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.I., Davis, W.W., Dodd, K.W. et Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, 102, 474-486.

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, NJ.

Schafer, D.W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics*, 57, 53-61.

Ybarra, L.M.R., et Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95, 919-931.