Recueil du Symposium 2002 de Statistique Canada
Modélisation des données d'enquête pour la recherche économique et sociale

MODÈLES RÉGIONAUX FONDÉS SUR DES DONNÉES PROVENANT D'ENQUÊTES MULTIPLES

Sharon L. Lohr et Lynn M. R. Ybarra[1]

RÉSUMÉ

Typiquement, les méthodes d'estimation sur petits domaines consistent à combiner des estimations directes d'après des données d'enquête et des prévisions d'après un modèle de régression pour obtenir des estimations de paramètres de population dont l'erreur quadratique moyenne est réduite. Ici, nous considérons les effets des valeurs erronées ou manquantes des covariables utilisées dans le modèle de régression. Ces situations surviennent lorsque les données sur les covariables proviennent d'une autre enquête ou d'une source administrative incomplète. Nous présentons et élaborons les propriétés de modèles qui permettent d'utiliser des données d'enquête et des données administratives comme information auxiliaire pour l'estimation des variables d'intérêt d'une enquête principale. La méthode tient compte des plans de sondage et des données manquantes dans la structure de l'erreur.

MOTS CLÉS : Meilleure prédiction linéaire sans biais; erreur de mesure; modèle multivarié à effets mixtes.

1. INTRODUCTION

Les enquêtes nationales, comme la U.S. Current Population Survey (CPS) ou la U.S. National Crime Victimization Survey (NCVS), permettent de produire des estimations exactes de la pauvreté ou du taux de victimes de la criminalité au niveau national. Par contre, elles sont réalisées auprès d'échantillons dont la taille n'est pas suffisante pour produire des estimations fiables, en soi, pour les « petits domaines » comme les États, les comtés ou les groupes minoritaires, ou pour fournir des renseignements détaillés sur des événements comme la violence familiale qui ne touchent qu'une petite partie de la population. Les méthodes courantes d'estimation de la pauvreté au niveau du comté consistent à intégrer des données administratives auxiliaires provenant de sources telles que les dossiers fiscaux et les programmes de coupons alimentaires à titre de variables explicatives dans une équation de régression; la valeur prévue obtenue par régression est combinée à une estimation directe de la pauvreté d'après les données de la CPS pour estimer le taux de pauvreté au niveau du comté (Citro et Kalton, 1999). Si le modèle de régression donne des prévisions exactes, l'erreur quadratique moyenne de l'estimation sur petits domaines résultante est plus faible que celle de l'estimation directe de la pauvreté au niveau du comté calculée d'après les données de la CPS. Les propriétés des estimations sur petits domaines, comme le biais et l'erreur quadratique moyenne, sont déterminées conditionnellement aux données auxiliaires. Cette méthode se fonde sur l'hypothèse que les données auxiliaires sont disponibles pour tous les domaines et qu'elles ne sont pas entachées d'erreurs.

Cependant, dans nombre de situations, on dispose de données auxiliaires qui facilitent l'estimation, mais qui ne sont pas exactes. Ces données auxiliaires peuvent provenir d'une autre enquête ou de dossiers administratifs dans lesquels les valeurs manquantes ont été remplacées par des données imputées. Dans ces deux cas, les données auxiliaires comportent une erreur — erreurs d'échantillonnage et non due à l'échantillonnage pour les données d'enquête et erreur d'imputation pour les données administratives incomplètes. Nous donnons quatre exemples de situations où les données auxiliaires peuvent être entachées d'erreurs.

À l'heure actuelle, les estimations du revenu et de la pauvreté au niveau de l'État et du comté sont produites d'après les données de la CPS, qui est réalisée chaque année auprès d'un échantillon d'environ 60 000 ménages, afin

[1]Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona 85287-1804 USA (sharon.lohr@asu.edu; lynn.ybarra@asu.edu)

d'obtenir une estimation directe de la pauvreté pour ces petits domaines. Toutefois, si le Congrès approuve son financement, l'American Community Survey (ACS) sera réalisée auprès d'un échantillon de 3 millions de ménages chaque année. Pour la plupart des petits domaines, l'ACS donnera, en principe, des estimations plus précises des variables mesurées; elle a donc été proposée comme source de données auxiliaires en vue de rendre les estimations d'après la CPS plus exactes. Toutefois, pour nombre de petits domaines, l'ACS contient encore une erreur d'échantillonnage dont il faut tenir compte dans toute fourchette d'erreur accompagnant les estimations.

Le U.S. Bureau of the Census souhaite produire des estimations de la couverture par une assurance-maladie dans les États, en se concentrant au départ sur la couverture des enfants vivant dans une famille à faible revenu. Ces estimations servent à la répartition des fonds pour le State Children's Health Insurance Program. Le Bureau utilise les données de la CPS pour calculer l'estimation primaire de la couverture par une assurance-maladie. Une question de recherche (Campbell et coll., 2002) est celle de savoir s'il faut utiliser les données auxiliaires que pourraient fournir certains États, mais dont la qualité est variable.

Un autre exemple est celui de l'estimation du taux de victime d'actes de violence dans chaque État ou du montant total des dépenses au titre des services médicaux subies dans chaque État en raison des crimes de violence. La U.S. National Crime Victimization Survey (NCVS) fournit ces renseignements, mais les tailles d'échantillon sont trop faibles pour produire des estimations exactes selon l'État. Le programme des Uniform Crime Reports (UCR) du FBI, qui fournit des statistiques sur les crimes déclarés aux services de police, est une excellente source de données auxiliaires. Bien que le programme des UCR sous-estime le nombre de crimes et les coûts qu'ils font subir à la société, la corrélation entre les taux d'actes de violence calculés d'après ce programme et d'après la NCVS est positive. Néanmoins, la déclaration au programme des UCR est volontaire et l'ensemble de données comporte de nombreuses lacunes; en outre, les données déclarées par certains services de police pourraient être inexactes.

Aux États-Unis, on procède maintenant à l'intégration de nombreux plans de sondage afin de pouvoir combiner les estimations. Ainsi, la National Health Interview Survey (NHIS) et la National Health and Nutrition Examination Survey (NHANES) sont basées sur une même unité primaire d'échantillonnage (UPE). Les UPE sélectionnées pour la NHIS servent de base de sondage pour la NHANES. La NHIS est réalisée auprès d'un échantillon probabiliste stratifié à plusieurs degrés d'environ 100 000 personnes (40 000 ménages) par année. Le plan de sondage est décrit en détail dans Botman et coll. (2000). Par contre, la NHANES comporte un examen médical des participants et l'unité d'examen mobile ne peut rendre visite qu'à 15 UPE par année (environ 5 000 personnes), par opposition au 358 UPE de la NHIS. Étant donné la petite taille de l'échantillon, on cumule habituellement les données de la NHANES au fil du temps afin de produire des estimations. Les estimations calculées au niveau de l'État ou au niveau local d'après les données de la NHANES sont peu précises. Les données de la NHIS produisent de meilleures estimations des variables mesurées pour certaines localités, mais elles proviennent d'une interview au lieu d'un examen physique : par exemple, dans le cas de la NHANES, on estime la prévalence du diabète d'après les résultats d'examens médicaux, tandis que, dans le cas de la NHIS, on l'estime d'après les renseignements recueillis au moyen d'un questionnaire. Toutefois, nous supposons que les résultats basés sur le questionnaire sont fortement corrélés à ceux basés sur l'examen physique et, donc, que la NHIS pourrait fournir des données auxiliaires de haute qualité susceptibles d'être combinées aux données de la NHANES pour améliorer les estimations sur petits domaines.

Fay et Herriot (1979) ont d'abord étudié l'amélioration des estimations sur petits domaines en se servant de vecteurs connus de moyennes de covariables. Depuis, de nombreux autres modèles ont été étudiés. Prasad et Rao (1990) ont regroupé un grand nombre de ces estimateurs dans un cadre unifié et ont calculé des approximations du deuxième ordre des erreurs quadratiques moyennes des estimateurs. Schaible (1996) ont décrit des estimateurs sur petits domaines indirects utilisés par les organismes gouvernementaux américains. Rao (2003) a exposé en détail les travaux ayant trait à l'estimation sur petits domaines réalisés jusqu'à présent.

Supposons qu'il existe m domaines d'intérêt (par exemple, $m = 50$ si les États sont des petits domaines). Nous voulons évaluer la caractéristique Y_i du domaine i. Pour certains domaines (ou tous), nous possédons des données provenant de l'enquête principale. Représentons par y_i un estimateur sans biais de Y_i provenant de l'enquête dont la variance d'échantillonnage est $V(y_i) = \psi_i$. Supposons que les données administratives pour le domaine i, $\mathbf{A}_i$, sont sans erreur. Nous considérons que le vecteur de longueur p, $\mathbf{x}_i$, représente les mesures provenant des sources de données auxiliaires, qui peuvent être entachées d'une erreur d'échantillonnage et (ou) d'un biais. Chaque vecteur $\mathbf{x}_i$ estime la

caractéristique réelle du domaine en question, $\mathbf{X}_i$. Nous supposons que $E(\mathbf{x}_i) = \mathbf{X}_i + \mathbf{b}_i$ et que $V(\mathbf{x}_i) = \mathbf{\Sigma}_i$. Notons que, si $\mathbf{X}_i$ est mesuré exactement, le biais $\mathbf{b}_i$ et la matrice des covariances $\mathbf{\Sigma}_i$ sont tous deux nuls.

Souvent, la caractéristique étudiée est une moyenne ou une proportion. Pour l'exemple de la criminalité, Y_i représente la proportion de personnes victimes d'un crime dans l'État *i*. La NCVS donne une estimation directe y_i; les données auxiliaires $\mathbf{x}_i$ proviennent du programme des UCR.

L'objectif ici est d'utiliser les données auxiliaires $\mathbf{x}_i$ pour améliorer l'estimation de la caractéristique étudiée Y_i. Lohr et Prasad (2001) ont élaboré des modèles au niveau de l'unité d'échantillonnage lorsque les données auxiliaires proviennent d'une autre enquête; cependant, dans ces modèles, il faut que l'on puisse apparier les données individuelles provenant des enquêtes principale et auxiliaires. Les modèles au niveau du domaine examinés ici nécessitent uniquement que l'on puisse coupler les données au niveau du domaine. À la section 2, nous décrivons les conséquences de l'utilisation du modèle de Fay-Herriot (1979) lorsque les $\mathbf{x}_i$ sont mesurés avec erreur et nous discutons d'une méthode bayésienne empirique. À la section 3, nous étendons le modèle multivarié de Fay-Herriot afin d'y intégrer l'erreur d'estimation. À la section 4, nous présentons des modèles à erreurs de mesure qui intègrent dans l'estimateur l'incertitude concernant $\mathbf{x}_i$. À la section 5, nous présentons nos conclusions.

2. MODÈLE DE FAY-HERRIOT ET ESTIMATION BAYÉSIENNE EMPIRIQUE

Le modèle de Fay-Herriot (1979) mène au meilleur prédicteur linéaire sans biais (MPLSB) de Y_i. Si nous supposons que y_i et Y_i suivent la loi de distribution normale, l'estimateur de Fay-Herriot peut être motivé au moyen d'une méthode bayésienne (voir Rao, 2003, chapitre 9). Dans ce cas, $y_i \mid Y_i, \psi_i \sim N(Y_i, \psi_i)$; le modèle de régression pour le paramètre de population est donné par

$$Y_i \mid \mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \alpha, \beta \sim N(\mathbf{A}_i^T \alpha + \mathbf{X}_i^T \beta, \sigma_v^2). \tag{2.1}$$

Nous supposons que les paramètres (y_i, Y_i) sont indépendants sur les domaines, conditionnellement à $\mathbf{A}_i$, $\mathbf{X}_i$, α, β et σ_v^2. Selon ce modèle, la distribution a posteriori de Y_i est

$$Y_i \mid y_i, \mathbf{A}_i, \mathbf{X}_i, \sigma_v^2, \alpha, \beta, \psi_i \sim N[\gamma_i^* y_i + (1-\gamma_i^*)(\mathbf{A}_i^T \alpha + \mathbf{X}_i^T \beta),\ \psi_i \gamma_i^*] \tag{2.2}$$

où $\gamma_i^* = \sigma_v^2 / (\sigma_v^2 + \psi_i)$. Notons que la moyenne a posteriori est le MPLSB pour Y_i si l'on connaît la valeur de tous les paramètres sur lesquels le modèle est conditionné. L'erreur quadratique moyenne (EQM) du MPLSB est la variance a posteriori.

Considérons maintenant ce qui se passe si l'on substitue une estimation $\hat{\mathbf{X}}_i$ au paramètre de population $\mathbf{X}_i$. En pratique, on pourrait utiliser $\mathbf{x}_i$ à la place de $\hat{\mathbf{X}}_i$. Posons que $\mathbf{C}_i = \text{EQM}(\hat{\mathbf{X}}_i)$ et que $\tilde{Y}_i^* = \gamma_i^* y_i + (1-\gamma_i^*)(\mathbf{A}_i^T \alpha + \hat{\mathbf{X}}_i^T \beta)$ est l'estimateur par substitution de Y_i. Alors, $\text{EQM}(\tilde{Y}_i^*) = \gamma_i^* \psi_i + (1-\gamma_i^*)^2 \beta^T \mathbf{C}_i \beta$. Notons que, si la matrice $\mathbf{C}_i$ est grande, l'erreur quadratique moyenne de l'estimateur par substitution peut être plus grande que ψ_i, la variance de l'estimateur fondée uniquement sur les données d'enquête. Donc, si l'information auxiliaire est inexacte, son utilisation peut produire une estimation moins exacte que si l'on n'utilisait pas de données auxiliaires. Le poids relatif, γ_i^*, appliqué à l'estimateur direct, y_i, dans l'estimateur par substitution pourrait être top faible, puisqu'il ne tient pas compte de la totalité de l'erreur qui entache la valeur prévue d'après la régression. En outre, si l'on suppose que l'estimation de $\mathbf{X}_i$ n'est pas entachée d'erreur, dans (2.2), la variance a posteriori est plus faible que l'EQM réelle.

Nous pouvons corriger l'EQM en intégrant l'erreur d'estimation de $\mathbf{X}_i$ dans le modèle (2.1). Supposons que $\mathbf{x}_i$ obéit aussi à la loi de distribution normale : $\mathbf{x}_i \mid \mathbf{X}_i, \mathbf{\Sigma}_i \sim N(\mathbf{X}_i, \mathbf{\Sigma}_i)$ et que $\mathbf{X}_i \mid \mu_X, \mathbf{\Sigma} \sim N(\mu_X, \mathbf{\Sigma})$. Alors, la distribution a posteriori de $\mathbf{X}_i$ est

$$\mathbf{X}_i \mid \mathbf{x}_i, \mu_x, \Sigma_i, \Sigma \sim N[\mathbf{c}_i,\ \mathbf{D}_i], \tag{2.3}$$

où $\mathbf{c}_i = \boldsymbol{\Sigma}\ (\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma})^{-1}\ \mathbf{x}_i + \boldsymbol{\Sigma}_i\ (\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma})^{-1}\ \mu_X$ et $\mathbf{D}_i = \boldsymbol{\Sigma}\ (\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma})^{-1}\ \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i\ (\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma})^{-1}\ \boldsymbol{\Sigma}$. Si nous combinons (2.2) et (2.3), nous voyons que la moyenne de la distribution a posteriori de Y_i est $\gamma_i^* y_i + (1-\gamma_i^*)(\mathbf{A}_i^T\alpha + \mathbf{c}_i^T\beta)$ et que sa variance est $\gamma_i^*\psi_i + (1-\gamma_i^*)^2\,\beta^T\mathbf{D}_i\beta$. En émettant les hypothèses supplémentaires concernant la distribution des données auxiliaires, la variance a posteriori est correcte pour l'EQM. Cependant, le poids relatif γ_i^* ne rend toujours pas compte de l'erreur d'estimation de $\mathbf{X}_i$; il est possible que la variance a posteriori soit plus grande que ψ_i, de sorte que l'intégration des données auxiliaires $\mathbf{x}$ pourrait réduire la précision. Dans les modèles présentés aux sections 3 et 4, les poids relatifs sont rajustés de façon à éviter ce problème.

Dans la discussion qui précède, nous supposons que les paramètres de régression et les matrices des covariances sont connus. En général, on doit les estimer d'après les données; le cas échéant, sous des conditions appropriées de régularité, l'EQM de l'estimateur sur petits domaines résultant est égale à la variance a posteriori susmentionnée à laquelle s'ajoutent des termes d'ordre inférieur. Les termes supplémentaires qui figurent dans l'EQM en raison de l'estimation des paramètres sont donnés dans Ybarra (2003).

3. MODÈLE MULTIVARIÉ DE FAY-HERRIOT

Fay (1987) et Datta et coll. (1991) ont élaboré un modèle de type Fay-Herriot pour une réponse multivariée et montré qu'il produit souvent des estimateurs plus efficaces pour les paramètres des petits domaines étudiés que le modèle univarié de Fay-Herriot. Datta et coll. (1991) souhaitent estimer Y_i, le revenu médian des ménages de quatre personnes dans l'État i. L'estimation directe y_i est fondée sur les données de la CPS. L'information auxiliaire, $\mathbf{x}_i$ = (3/4) (revenu médian des ménages de cinq personnes) + (1/4) (revenu médian des ménages de trois personnes) provient aussi de la CPS. Selon ces auteurs, l'utilisation d'un modèle multivarié réduit l'EQM de l'estimateur de Y_i.

Nous étendons ce modèle de façon à permettre des observations manquantes et des erreurs de mesure, ainsi que des observations provenant de sources différentes. Représentons par $\mathbf{0}_k$ un vecteur de longueur k dont tous les éléments sont nuls et représentons par $\mathbf{I}_k$ la matrice d'identité $k \times k$. Supposons jusqu'à la fin de cette section que $\mathbf{x}_i$ est un estimateur sans biais de $\mathbf{X}_i$, de sorte que $\mathbf{b}_i = \mathbf{0}$.

Représentons par $\mathbf{U}_i = [\mathbf{X}_i^T \quad Y_i\ ]^T$ les valeurs de population pour chacun des i domaines, i = 1, …, m. Alors, un modèle établissant la relation entre les paramètres de population pour les domaines est donné par $\mathbf{U}_i = \mathbf{A}_i^T\alpha\ + \mathbf{v}_i$, où $\mathbf{v}_i\ \sim N(\mathbf{0}_{p+1}, \boldsymbol{\Sigma}_v)$. La matrice des covariances du modèle $\boldsymbol{\Sigma}_v$ est partitionnée comme suit

$$\boldsymbol{\Sigma}_v = \begin{bmatrix} \boldsymbol{\Sigma}_{vxx} & \boldsymbol{\Sigma}_{vxy} \\ \boldsymbol{\Sigma}_{vxy}^T & \boldsymbol{\Sigma}_{vyy} \end{bmatrix}.$$

Définissons le vecteur $\mathbf{u}_i$ et les matrices $\mathbf{Z}_i$ et $\boldsymbol{\Psi}_i$ pour trois cas :

1. $\mathbf{u}_i = [\mathbf{x}_i^T, y_i\ ]^T$, $\mathbf{Z}_i = \mathbf{I}_{p+1}$, $\boldsymbol{\Psi}_i$ = blockdiag($\boldsymbol{\Sigma}_i$, ψ_i) si $\mathbf{x}$ et y sont tous deux observés dans le domaine i;
2. $\mathbf{u}_i = \mathbf{x}_i$, $\mathbf{Z}_i^T = [\ \mathbf{I}_p, 0]^T$, $\boldsymbol{\Psi}_i = \boldsymbol{\Sigma}_i$ si $\mathbf{x}$ est observé dans le domaine i, mais non y;
3. $\mathbf{u}_i = y_i$, $\mathbf{Z}_i^T = [\ \mathbf{0}_p, 1]^T$, $\boldsymbol{\Psi}_i = \psi_i$ si y est observé dans le domaine i, mais non $\mathbf{x}$.

Alors, les observations $\mathbf{u}_i$ suivent le modèle $\mathbf{u}_i = \mathbf{Z}_i^T\mathbf{A}_i^T\alpha\ + \mathbf{Z}_i^T\mathbf{v}_i + \mathbf{e}_i$, où $\mathbf{e}_i\ \sim N(\mathbf{0}, \boldsymbol{\Psi}_i)$. La matrice des covariances de $\mathbf{u}_i$ est $\mathbf{V}_i = \mathrm{V}(\mathbf{u}_i) = \mathbf{Z}_i^T\boldsymbol{\Sigma}_v\mathbf{Z}_i + \boldsymbol{\Psi}_i$. Ce modèle concorde avec le modèle à structure diagonale par blocs des covariances décrit à la section 6.3 de Rao (2003). Le MPLSB de $\mathbf{U}_i$ est alors

$$\tilde{\mathbf{U}}_i = \mathbf{A}_i\tilde{\alpha} + \tilde{\mathbf{v}}_i, \tag{3.1}$$

où $\tilde{\mathbf{v}}_i = \boldsymbol{\Sigma}_v\mathbf{Z}_i\mathbf{V}_i^{-1}(\mathbf{u}_i - \mathbf{Z}_i^T\mathbf{A}_i\tilde{\alpha})$ et

$$\tilde{\alpha} = \left(\sum_i \mathbf{A}_i^T \mathbf{Z}_i \mathbf{V}_i^{-1} \mathbf{Z}_i^T \mathbf{A}_i \right)^{-1} \left(\sum_i \mathbf{A}_i^T \mathbf{Z}_i \mathbf{V}_i^{-1} \mathbf{u}_i \right). \tag{3.2}$$

Posons que $\mathbf{M}_i = (\mathbf{\Sigma}_{vxx} + \mathbf{\Sigma}_i)^{-1}$ et $\kappa_i = (\mathbf{\Sigma}_{vyy} - \mathbf{\Sigma}_{vxy}{}^T \mathbf{M}_i \mathbf{\Sigma}_{vxy}) / (\mathbf{\Sigma}_{vyy} - \mathbf{\Sigma}_{vxy}{}^T \mathbf{M}_i \mathbf{\Sigma}_{vxy} + \psi_i)$. En utilisant (3.1),

$$\tilde{Y}_{i,\mathrm{MFH}} = \kappa_i y_i + (1-\kappa_i)\{[\mathbf{0}_p^T, 1]\mathbf{A}_i\tilde{\alpha} + \mathbf{\Sigma}_{vxy}^T \mathbf{M}_i (\mathbf{x}_i - [\mathbf{I}_p, 1]\mathbf{A}_i\tilde{\alpha})\} \tag{3.3}$$

pour les cas 1 et 2 susmentionnés (notons que pour le cas 2, $\kappa_i = 0$). Pour le cas 3 (domaines *i* pour lesquels **x** n'est pas mesuré et pour lesquels les éléments de $\mathbf{M}_i$ sont donc nuls), $\tilde{Y}_{i,\mathrm{MFH}} = \kappa_i y_i + (1-\kappa_i)[\mathbf{0}_p^T, 1]\mathbf{A}_i\tilde{\alpha}$, et $\kappa_i = \mathbf{\Sigma}_{vyy} / (\mathbf{\Sigma}_{vyy} + \psi_i)$.

Dans l'estimateur sur petits domaines (3.3), le poids κ_i dépend donc de la variabilité de $\mathbf{x}_i$: la valeur de κ_i est plus faible et l'estimateur sur petits domaines dépend fortement de l'estimateur direct si la variabilité de $\mathbf{x}_i$ est importante. Si $\mathbf{X}_i$ est mesuré exactement (c.-à-d. si tous les éléments de $\mathbf{\Sigma}_i$ sont nuls), sous les hypothèses de normalité, l'estimateur multivarié de Fay-Herriot coïncide avec l'estimateur univarié de Fay-Herriot qui intègre les données sur **x** en tant que covariables.

L'EQM de l'estimateur (3.3) peut être calculée par les méthodes standards. Sous les hypothèses de régularité données dans Datta et coll. (1991) et dans Ybarra (2003), nous avons, pour les cas 1 et 3, que $\mathrm{EQM}(\tilde{Y}_{i,\mathrm{MFH}}) = \kappa_i\psi_i + O(m^{-1})$. Pour le cas 2, $\mathrm{EQM}(\tilde{Y}_{i,\mathrm{MFH}}) = \kappa_i^* + O(m^{-1})$, où κ_i^* est le numérateur de κ_i.

En pratique, on doit estimer $\mathbf{\Sigma}_v$ ainsi que α d'après les données. On peut, pour cela, utiliser la méthode des moments, du maximum de vraisemblance ou du maximum de vraisemblance limité. Voir Datta et coll. (2001) pour une comparaison des estimateurs de $\mathbf{\Sigma}_v$ dans le cas univarié.

4. MODÈLES À ERREURS DE MESURE

4.1 Estimation lorsqu'on connaît les paramètres de régression et les variances

Nous avons vu à la section 2 que si l'on ne tient pas compte de l'erreur dans $\mathbf{x}_i$, l'erreur quadratique moyenne est sous-estimée et le poids utilisé n'est pas optimal. La motivation qui pousse à l'utilisation d'un modèle à erreurs de mesure est que les covariables omises ou inexactes peuvent biaiser l'estimation. Supposons que le modèle (2.1) soit vérifié, mais que l'analyste l'ajuste sans le terme contenant β. Puisque le « mauvais » modèle est ajusté, les estimations des paramètres de régression α et les valeurs prévues peuvent être biaisées. Le biais dû à l'omission de **X** dans l'ensemble de covariables cause une augmentation de l'EQM des valeurs prévues. Par contre, si **X** est inclus dans les covariables, il faut tenir compte de l'erreur qui entache sa mesure dans le calcul de l'estimation et de l'erreur quadratique moyenne. Fuller (1987, 1990) a présenté un traitement complet de l'utilisation des modèles à erreurs de mesure pour l'estimation des paramètres de régression et pour la prévision.

Comme à la section 2, représentons par $\hat{\mathbf{X}}_i$ un estimateur du paramètre de population $\mathbf{X}_i$ pour lequel $\mathbf{C}_i = \mathrm{EQM}(\hat{\mathbf{X}}_i)$. Nous supposons qu'un tel estimateur existe pour chaque domaine : si **x** n'est pas mesuré dans le domaine *i*, alors nous pouvons utiliser un estimateur bayésien empirique ou une valeur imputée pour $\hat{\mathbf{X}}_i$. Considérons le modèle

$$y_i = \mathbf{A}_i^T \alpha + \hat{\mathbf{X}}_i^T \beta + r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) + e_i, \tag{4.1}$$

où $r_i(\hat{\mathbf{X}}_i, \mathbf{X}_i) = v_i + (\mathbf{X}_i - \hat{\mathbf{X}}_i)\beta$ et EQM(r_i) = $\sigma_v^2 + \beta^T \mathbf{C}_i \beta$. Comme précédemment, $V(e_i) = \psi_i$ représente la variance d'échantillonnage de y_i. Nous supposons ici que y_i, v_i et $\hat{\mathbf{X}}_i$ sont mutuellement indépendants. Maintenant, posons que

$$\tilde{Y}_i = \gamma_i y_i + (1-\gamma_i)\{\mathbf{A}_i^T \alpha + \hat{\mathbf{X}}_i^T \beta\}, \tag{4.2}$$

où $\gamma_i = (\sigma_v^2 + \beta^T \mathbf{C}_i \beta)/(\sigma_v^2 + \beta^T \mathbf{C}_i \beta + \psi_i)$. Si y_i est mesuré dans le domaine i, alors EQM $(\tilde{Y}_i) = \gamma_i \psi_i$, qui est, au plus, aussi grande que la variance ψ_i de l'estimateur direct, y_i. Si y_i n'est pas mesuré dans le domaine i, alors EQM $(\tilde{Y}_i) = \sigma_v^2 + \beta^T \mathbf{C}_i \beta$.

Notons que EQM $(\tilde{Y}_i) \leq$ EQM $(\tilde{Y}_i^*)$, où $\tilde{Y}_i^*$ est l'estimateur par substitution décrit à la section 2: Il y a égalité si $\beta^T \mathbf{C}_i \beta = 0$. L'EQM est également plus faible que celle que l'on obtiendrait en utilisant le modèle $y_i = \mathbf{A}_i^T \alpha + t_i + e_i$ au lieu de (4.1). Si nous utilisons l'estimateur bayésien empirique de la section 2 pour $\hat{\mathbf{X}}_i$, alors nous pouvons montrer que l'estimateur (4.2) est équivalent à l'estimateur multivarié de Fay-Herriot.

4.2 Estimation des paramètres de régression et des variances

En pratique, on ne connaît pas les paramètres σ_v^2, α et β et on doit les estimer d'après les données. Bien que la plupart des études portant sur les modèles régionaux s'appuient sur l'hypothèse que ψ_i est connue, il peut être nécessaire d'estimer $\mathbf{C}_i$ également.

Lindley (1947, p. 243) a proposé d'utiliser les moindres carrés pondérés pour estimer les paramètres de régression. Pour notre modèle, l'EQM des termes d'erreur est EQM ($r_i + e_i$) = $\sigma_v^2 + \psi_i + \beta^T \mathbf{C}_i \beta$. Donc, nous pouvons résoudre pour le paramètres inconnu en minimisant

$$Q_1(\alpha, \beta) = \sum_i \frac{(y_i - \mathbf{A}_i^T \alpha - \hat{\mathbf{X}}_i^T \beta)^2}{\psi_i + \sigma_v^2 + \beta^T \mathbf{C}_i \beta} \tag{4.3}$$

où la somme est faite sur les domaines i où y est mesuré. Gleser (1981) donne les propriétés sur grand échantillon des estimations résultantes des paramètres de régression.

Toutefois, l'utilisation de (4.3) nécessite que l'on connaisse σ_v^2 . Sinon, on peut utiliser les moindres carrés modifiés pour estimer les paramètres (Cheng et Van Ness, 1999, pp. 85 et 146). Dans ce cas, un estimateur sans biais de σ_v^2 est

$$Q_2(\alpha, \beta) = \frac{1}{m} \sum_i [(y_i - \mathbf{A}_i^T \alpha - \hat{\mathbf{X}}_i^T \beta)^2 - \psi_i - \beta^T \mathbf{C}_i \beta] \tag{4.4}$$

La minimisation de Q_2 par rapport à α et β donne les estimation des paramètres de régression. Notons toutefois que les termes de (4.4) peuvent être négatifs et que la minimisation pourrait avoir lieu sur les limites de l'espace des paramètres.

5. DISCUSSION

Nous décrivons et comparons dans le présent article les propriétés du premier ordre de trois méthodes, basées respectivement sur un modèle bayésien empirique, un modèle multivarié de Fay-Herriot et un modèle à erreurs de mesure, d'intégration de données auxiliaires dans l'estimation sur petits domaines, lorsque les données auxiliaires peuvent contenir une erreur. Les trois méthodes tiennent compte de la variabilité supplémentaire due au fait qu'on ne connaît pas exactement la valeur des données auxiliaires. Les modèles multivarié de Fay-Herriot et à erreurs de mesure comportent aussi un mécanisme de variation de la pondération relative de l'estimation directe et de la valeur

prévue d'après l'équation de régression, de façon à accorder plus d'importance à l'estimation directe si $\mathbf{X}_i$ est mesuré de façon inexacte. Le modèle à erreurs de mesure est le plus souple en ce qui concerne le choix de l'estimateur de $\mathbf{X}_i$; il donne les mêmes résultats que l'estimateur multivarié de Fay-Herriot dans le cas normal où on ne dispose d'aucune covariable de source administrative et qu'on utilise l'estimateur par rétrécissement de la section 2 pour estimer $\mathbf{X}_i$.

Les valeurs de l'EQM du premier ordre présentées ici sont calculées en supposant qu'on connaît les valeurs des paramètres de régression et des covariances. Si on ne les connaît pas et qu'on doit les estimer d'après les données, les estimateurs de l'EQM seront plus grands qu'il ne l'est indiqué ici. Les propriétés du deuxième ordre dépendent des grandeurs relatives de m, du nombre de domaines et de $\mathbf{C}_i$, c.-à-d. l'erreur d'estimation de $\mathbf{X}_i$. Ybarra (2003) compare les estimateurs des paramètres inconnus et calcule les termes du deuxième ordre de l'EQM des estimateurs.

Même si, dans certaines situations, le modèle à erreurs de mesure et le modèle multivarié de Fay-Herriot donnent les mêmes résultats, nous préférons le premier dans nombre de situations pratiques, car il est plus souple en ce qui concerne le choix de l'estimateur de $\mathbf{X}_i$. En outre, comme on peut utiliser des méthodes robustes pour estimer les paramètres de régression et les termes de variance, le modèle à erreurs de mesure s'adapte à des situations où les valeurs de $\mathbf{x}_i$ sont aberrantes à cause de la qualité variable des sources de données.

REMERCIEMENTS

La présente étude a été financée en partie par la subvention n^o 0105852 de la U.S. National Science Foundation.

RÉFÉRENCES

Botman, S.L., Moore, T.F., Moriarity, C.L. et Parsons, V.L. (2000), “Design and Estimation for the National Health Interview Survey, 1995-2004”, National Center for Health Statistics. *Vital Health Statistics* 2(130).

Campbell, J., Fisher, R. et Waddington, D. (2002), “Preliminary Research for Small Area Health Insurance Estimates”, paper presented at the October meeting of the Census Advisory Committee of Professional Associations, Arlington, Virginia.

Cheng, C.-L. et Van Ness, J.W. (1999), *Statistical Regression with Measurement Error*, London: Arnold.

Citro, C.F. et Kalton, G. (eds.) (1999), *Small-area Estimates of School-age Children in Poverty: Interim Report 3*, Panel on Estimates of Poverty for Small Geographic Areas, Committee on National Statistics. Washington, D.C.: National Academy Press.

Datta, G.S., Fay, R.E. et Ghosh, M. (1991), “Hierarchical and Empirical Multivariate Analysis in Small Area Estimation”, *Proceedings of the Bureau of the Census Annual Research Conference*, Washington: Bureau of the Census, pp. 63-79.

Datta, G.S., Rao, J.N.K. et Smith, D. D. (2001), “On Measures of Uncertainty of Small Area Estimators in the Fay-Herriot Model”, Technical Report, University of Georgia, Athens, Georgia.

Fay, R.E. (1987), “Application of Multivariate Regression to Small Domain Estimation”, in R. Platek et al. (eds.), *Small Area Statistics*, New York: Wiley, pp. 91-102.

Fay, R.E. et Herriot, R.A. (1979), “Estimates of Income for Small Places: An Empirical Bayes Application of James-Stein Procedures to Census Data”, *Journal of the American Statistical Association*, 78, pp. 269-277.

Fuller, W.A. (1987), *Measurement Error Models*, New York: Wiley.

Fuller, W.A. (1990), "Prediction of True Values for the Measurement Error Model", in P. J. Brown and W. A. Fuller (eds.), *Statistical Analysis of Measurement Error Models and Applications, Contemporary Mathematics Vol. 112*, Providence, RI: AMS, pp. 41-57.

Gleser, L.J. (1981), "Estimation in a Multivariate 'Errors-in-Variables' Regression Model: Large Sample Results", *Annals of Statistics*, 9, pp. 24-44.

Lindley, D.V. (1947), "Regression Lines and the Linear Functional Relationship". *Journal of the Royal Statistical Society Supp*., 9, pp. 218-244.

Lohr, S. et Prasad, N.G.N. (2001), "Small Area Estimation with Auxiliary Survey Data", Technical Report 01.08, Statistics Centre, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton.

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

Rao, J.N.K. (2003), *Small Area Estimation*, New York: Wiley, in press.

Schaible, W.L. (ed.) (1996), *Indirect Estimators in U.S. Federal Programs*. Lecture Notes in Statistics, 108, New York: Springer-Verlag.

Ybarra, L.M.R. (2003), *Small Area Estimation using Data from Multiple Surveys*, unpublished Ph.D. dissertation, Arizona State University.