Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Comparaison de méthodes de régression sur données d'enquête dans le contexte de l'estimation de la pauvreté sur des petits domaines

par Stephen J. Haslett, Marissa C. Isidro et Geoffrey Jones

Décembre 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2010
Vol. 36, N° 2, pp. 169-184
Statistique Canada, N° 12-001-X au catalogue

Comparaison de méthodes de régression sur données d'enquête dans le contexte de l'estimation de la pauvreté sur des petits domaines

Stephen J. Haslett, Marissa C. Isidro et Geoffrey Jones [1]

Résumé

L'une des clés de la réduction ou de l'éradication de la pauvreté dans le tiers monde est l'obtention d'information fiable sur les pauvres et sur leur emplacement, afin que les interventions et l'aide soient dirigées vers les personnes les plus nécessiteuses. L'estimation sur petits domaines est une méthode statistique utilisée pour surveiller la pauvreté et décider de la répartition de l'aide de façon à réaliser les Objectifs du millénaire pour le développement. Elbers, Lanjouw et Lanjouw (ELL) (2003) ont proposé, pour produire des mesures de la pauvreté fondées sur le revenu ou sur les dépenses, une méthode d'estimation sur petits domaines qui est mise en œuvre par la Banque mondiale dans ses projets de cartographie de la pauvreté grâce à la participation des organismes statistiques centraux de nombreux pays du tiers monde, dont le Cambodge, le Laos, les Philippines, la Thaïlande et le Vietnam, et qui est intégrée dans le logiciel PovMap de la Banque mondiale. Dans le présent article, nous présentons la méthode ELL, qui consiste à modéliser d'abord les données d'enquête, puis à appliquer le modèle obtenu à des données de recensement, en nous penchant surtout sur la première phase, c'est-à-dire l'ajustement des modèles de régression, ainsi que sur les erreurs-types estimées à la deuxième phase. Nous présentons d'autres méthodes d'ajustement de modèles de régression, telles que la régression généralisée sur données d'enquête (RGE) (décrite dans Lohr (1999), chapitre 11) et celles utilisées dans les méthodes existantes d'estimations sur petits domaines, à savoir la méthode du meilleur prédicteur linéaire sans biais pseudo-empirique (pseudo-MPLSB) (You et Rao 2002) et la méthode itérative à équations d'estimation pondérées (IEEP) (You, Rao et Kovačević 2003), et nous les comparons à la stratégie de modélisation de ELL. La différence la plus importante entre la méthode ELL et les autres techniques tient au fondement théorique de la méthode d'ajustement du modèle proposée par ELL. Nous nous servons d'un exemple fondé sur la Family Income and Expenses Survey des Philippines pour illustrer les différences entre les estimations des paramètres et leurs erreurs-types correspondantes, ainsi qu'entre les composantes de la variance générées par les diverses méthodes et nous étendons la discussion à l'effet de ces différences sur l'exactitude estimée des estimations sur petits domaines finales. Nous mettons l'accent sur la nécessité de produire de bonnes estimations des composantes de la variance, ainsi que des coefficients de régression et de leurs erreurs-types aux fins de l'estimation sur petits domaines de la pauvreté.

Mots clés : Modèles pour petits domaines ; modèle de régression à erreurs emboîtées ; cartographie de la pauvreté.

1. Introduction

La pauvreté est un problème multidimensionnel très complexe pour lequel il n'existe pas de définition unique ni de méthode de mesure unique. Dans le présent article, nous adoptons le sens donné au terme pauvreté par la plupart des économistes. Les ménages considérés comme étant en état de pauvreté sont ceux dont le revenu est inférieur à un certain seuil de revenu appelé seuil de pauvreté. Chambers (2006) donne à cette approche le nom de pauvreté fondée sur le revenu et cette définition est celle adoptée par la Banque mondiale dans la mise en œuvre de ses projets de cartographie de la pauvreté par petit domaine exécutés en collaboration avec les organismes statistiques nationaux et utilisés, par exemple, pour surveiller les progrès réalisés en regard des Objectifs du millénaire pour le développement (site Web de l'ONU). Parfois, des mesures de la pauvreté fondées sur les dépenses sont utilisées à la place de celles fondées sur le revenu pour évaluer la pauvreté économique. Dans les contextes de santé publique, diverses mesures, telles que le poids et la taille normalisés selon l'âge, ainsi que le poids en fonction de la taille chez les enfants (insuffisance pondérale, retard de croissance et émaciation, respectivement) sont utilisés, par exemple, au Bangladesh (Haslett et Jones 2004) et au Népal (Haslett et Jones 2006).

Les enquêtes réalisées dans la plupart des pays du tiers monde permettent habituellement d'obtenir un niveau acceptable de précision pour la publication de statistiques sur la pauvreté aux premier et deuxième niveaux administratifs ou au niveau de la région géographique (par exemple pour les Philippines, national et régional, respectivement). Cependant, pour que les responsables de l'élaboration des politiques puissent diriger correctement l'aide et les interventions vers les communautés et les ménages qui en ont le plus besoin, ils doivent disposer de statistiques sur la pauvreté produites à un niveau plus fin de détail. Toutefois, les statistiques sur la pauvreté calculées au moyen de données d'enquête pour de plus petites régions géographiques ou un plus faible niveau administratif sont habituellement moins fiables (possèdent des erreurs-types plus élevées) à cause des tailles d'échantillon plus petites et c'est

1. Stephen J. Haslett, Marissa C. Isidro et Geoffrey Jones, Institute of Fundamental Sciences: Statistics, College of Sciences, Massey University, Private Bag 11-222, Palmerston North, Nouvelle-Zélande. Courriel : S.J.Haslett@massey.ac.nz.

dans de telles situations que l'estimation sur petits domaines entre en jeu.

La méthode d'estimation sur petits domaines utilisée le plus fréquemment pour mesurer la pauvreté dans les pays du tiers monde, qui a été proposée par Elbers, Lanjouw et Lanjouw (ELL) (2002, 2003), permet de produire des estimations plus précises pour les régions géographiques plus petites en combinant des données d'enquête avec des renseignements tirés d'un recensement récent. La méthode ELL comporte deux phases, à savoir l'ajustement d'un modèle (ou de modèles) de régression à des données d'enquête complexes, puis l'utilisation de ce modèle pour prédire le revenu et les dépenses par tête au niveau du ménage (valeurs qui sont transformées et agrégées pour estimer les statistiques sur la pauvreté au niveau du petit domaine).

Dans le présent article, nous nous penchons spécifiquement sur les divers algorithmes utilisés pour ajuster les modèles de régression à la première phase et pour estimer les erreurs-types des paramètres de régression et les composantes de la variance d'après des données d'enquête. Nous mettons l'accent sur les conséquences des décisions concernant la modélisation de régressions sur données d'enquête plutôt que sur le système entier et assez complet qu'ELL utilisent pour produire des estimations sur petits domaines.

L'exigence préliminaire, lorsque l'on applique la méthode ELL à des mesures économiques, est d'élaborer un modèle exact du revenu ou des dépenses par tête dans les ménages, bien que ce modèle soit souvent utilisé pour générer des fonctions non linéaires du revenu ou des dépenses (par exemple prévalence de la pauvreté – pourcentage de ménages sous le seuil de pauvreté, ou écart de pauvreté – somme des écarts relatifs entre le seuil de pauvreté et le revenu ou les dépenses des ménages ou des individus qui se trouvent sous le seuil de pauvreté). Le modèle de régression sur données d'enquête élaboré pour le revenu ou les dépenses joue un rôle essentiel dans la production de statistiques exactes sur la pauvreté, mais, comme nous le montrons plus loin, le modèle de régression proprement dit n'est pas toujours l'élément le plus important et d'autres questions, telles que l'estimation des composantes de la variance, méritent qu'on s'y attarde.

D'autres méthodes de régression sur données d'enquête pour l'estimation sur petits domaines – la méthode du meilleur prédicteur linéaire sans biais pseudo-empirique (pseudo-MPLSB) (You et Rao 2002), la méthode itérative à équations d'estimations pondérées (IEEP) (You et coll. 2003) et la méthode de régression généralisée sur données d'enquête (RGE) (Skinner, Holt et Smith 1989) sont examinées en tant que techniques d'ajustement de modèles à des données d'enquête et comparées à deux variantes de la méthode ELL d'ajustement de modèles de régression à des données d'enquête. Notre étude est fondée sur des données réelles tirées de la Family Income and Expenses Survey (FIES) des Philippines, plutôt que sur des données simulées.

Le plan de l'article est le suivant. À la section 2, nous donnons des renseignements généraux sur les modèles d'estimation sur petits domaines ; à la section 3, nous présentons le modèle pour le revenu (ou les dépenses) décrit par Elbers, Lanjouw et Lanjouw ; à la section 4, nous présentons un sommaire de la méthode ELL ; à la section 5, nous décrivons en détail les diverses méthodes d'ajustement, dont la méthode du meilleur prédicteur linéaire sans biais pseudo-empirique (5.1), la méthode IEEP (5.2) et la méthode de régression généralisée sur données d'enquête (5.3). À la section 6, nous discutons des différences entre les méthodes, tandis qu'à la section 7, nous présentons l'application des méthodes aux données de la FIES de 2000 des Philippines. Enfin, à la section 8, nous présentons nos conclusions et nos recommandations.

2. Modèles pour petits domaines

Ghosh et Rao (1994) classent les modèles pour petits domaines en deux grandes catégories, c'est-à-dire les modèles au niveau du domaine et les modèles au niveau de l'unité. Les modèles au niveau du domaine correspondent à l'ensemble de modèles qui peuvent être pris en considération quand on ne dispose que de variables auxiliaires propres au domaine. Les modèles au niveau de l'unité, par ailleurs, englobent ceux qui peuvent être considérés lorsqu'il existe des variables auxiliaires propres à l'unité et que des valeurs de la variable étudiée au niveau de l'unité peuvent être utilisées. Tous ces modèles sont des cas particuliers d'un modèle linéaire généralisé ou d'un modèle mixte linéaire généralisé et contiennent habituellement des effets fixes ainsi que des effets aléatoires.

Pour les modèles au niveau du domaine, on suppose que la moyenne de population $(\bar{Y}_a)$ du a^{e} petit domaine ou une fonction appropriée $\theta_a = g(\bar{Y}_a)$ est reliée aux variables auxiliaires propres au domaine $\mathbf{x}_a = (x_{a1}, ..., x_{ap})'$ au moyen d'un modèle linéaire

$$\theta_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a \qquad (1)$$

où $a = 1, ..., k$, $v_a \sim \text{iid}(0, \sigma_v^2)$, $\boldsymbol{\beta}$ est un vecteur de paramètres de régression, c_a est une constante positive connue ou estimée pour tenir compte de l'hétéroscédasticité, k est le nombre total de petits domaines d'intérêt et p est le nombre de variables auxiliaires. On suppose qu'un estimateur direct fondé sur le plan de sondage, $\hat{\bar{Y}}_a$, de la moyenne de population $\bar{Y}_a$ est disponible quand la taille de l'échantillon de domaine $n_a \geq 1$, et que

$$\hat{\theta}_a = \theta_a + e_a \qquad (2)$$

où $\hat{\theta}_a = g(\hat{\bar{Y}}_a)$ et les erreurs dues à l'échantillonnage e_a sont indépendantes et suivent une loi $N(0, V_a)$ de variance connue V_a. En combinant les équations (1) et (2), on obtient le modèle mixte linéaire au niveau du domaine :

$$\hat{\theta}_a = \mathbf{x}'_a\boldsymbol{\beta} + c_a v_a + e_a. \qquad (3)$$

Soulignons que (3) fait intervenir à la fois des variables aléatoires fondées sur le plan de sondage e_a et des variables aléatoires fondées sur le modèle v_a (Rao 1999), où les variables fondées sur le plan de sondage dépendent du mécanisme de sélection d'échantillons et celles fondées sur le modèle, de la structure de superpopulation dans laquelle le modèle est intégré.

Les modèles au niveau du domaine possèdent diverses extensions qui permettent, par exemple, de traiter des erreurs dues à l'échantillonnage corrélées, la dépendance spatiale des effets aléatoires de petit domaine, les séries chronologiques et les données transversales (voir Rao 2003, 1999, ainsi que Ghosh et Rao 1994).

Le modèle au niveau de l'unité repose sur l'hypothèse que la variable d'intérêt Y_{ah} pour la h^{e} unité dans le a^{e} petit domaine est reliée aux données auxiliaires propres à l'élément $\mathbf{x}_{ah} = (x_{ah1}, ..., x_{ahp})'$ par la voie d'un modèle de régression à erreurs emboîtées :

$$Y_{ah} = \mathbf{x}'_{ah}\boldsymbol{\beta} + v_a + e_{ah} \qquad (4)$$

où $a = 1, ..., k$, $h = 1, ..., N_a$, $\boldsymbol{\beta} = (\beta_0, ... \beta_{p-1})$ est le vecteur de dimensions $p \times 1$ des paramètres de régression et N_a est le nombre d'unités ou de ménages de la population dans le a^{e} petit domaine. Il est également supposé que les effets aléatoires, v_a, sont iid $N(0, \sigma_v^2)$ et sont indépendants des erreurs au niveau de l'unité, e_{ah}, qui sont supposées être iid $N(0, \sigma_e^2)$. Des extensions permettant que les erreurs soient hétéroscédastiques, avec constante(s) d'échelle connue(s) sont également possibles.

La méthode ELL s'appuie sur un modèle au niveau de l'unité, où les unités sont les ménages dans le cas des données sur les revenus ou les dépenses, et où la variation est modélisée au niveau de l'unité primaire d'échantillonnage, c'est-à-dire au niveau de la grappe et au niveau du ménage. Notons que ELL n'intègrent pas la variation du modèle au niveau du petit domaine, et ne le font que pour la grappe dans le petit domaine, et pour le ménage dans la grappe. Cette forme du modèle de base est celle utilisée pour les comparaisons dans le présent article, puisque la méthode ELL est la méthode standard d'estimation sur petits domaines de la pauvreté dans les pays du tiers monde. Dans les ensembles de données réelles que nous avons étudiés, cette variation supplémentaire au niveau du petit domaine était très faible. Toutefois, malgré ces preuves empiriques, d'importantes questions persistent quant à la meilleure façon d'estimer la composante de la variance au niveau du petit domaine en présence de variations au niveau de la grappe, en cas de pondération par les poids de sondage, surtout quand de nombreux petits domaines ne contiennent qu'une seule grappe échantillonnée.

Le modèle ELL possède un certain nombre d'autres caractéristiques qui ne sont pas toutes standard au sens statistique (voir Haslett et Jones 2005, par exemple). Le but du présent article n'est pas de discuter des différences entre les méthodes existantes en général, mais plutôt de nous concentrer directement sur les différences entre les méthodes d'ajustement des modèles de régression aux données d'enquête quand est utilisée la « structure de base » de la première phase de la méthode ELL pour ajuster les modèles de régression sur données d'enquête. Par conséquent, le présent article est axé sur la comparaison des méthodes existantes d'ajustement de modèles de régression aux données d'enquête sur le revenu ou les dépenses en utilisant un ensemble spécifié de variables indépendantes, même si la méthode ELL peut également être (et est) utilisée relativement couramment pour obtenir des estimations sur petits domaines pour des fonctions non linéaires (par exemple prévalence ou gravité de la pauvreté, ou écart de pauvreté) grâce à l'application des modèles de régression ajustés aux données d'enquête à des données de recensement.

La réponse à la question de savoir quel est le « meilleur ajustement du modèle de régression » aux données d'enquête analysées dans le présent article (comme celles à d'autres questions liées à la méthode ELL) est particulièrement importante, parce que des milliards de dollars destinés au financement de l'aide sont (ou pourraient être) répartis en se fondant sur les modèles de régression utilisés dans le cadre de l'estimation sur petits domaines de la pauvreté.

3. Modèle de revenu/consommation

La modélisation du revenu ou des dépenses par tête dans les ménages au lieu de mesures de la pauvreté proprement dites (telles que la prévalence de la pauvreté et l'écart de la pauvreté) est l'une des caractéristiques distinctives de la méthode ELL. Comme nous l'avons mentionné à la section précédente, cette méthode comprend l'ajustement du modèle de revenu ou de dépenses aux données d'enquête et l'application de ce modèle à des données de recensement avant de produire les estimations sur petits domaines des mesures de la pauvreté. Le modèle de revenu/dépenses est de la forme :

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + u_{bh} \qquad (5)$$

où $b = 1, ..., M$, $h = 1, ..., N_b$; Y_{bh} représente le revenu ou les dépenses par tête après transformation logarithmique de la h^{e} unité ou ménage dans la b^{e} grappe, M est le nombre total de grappes dans la population et N_b est le nombre total

de ménages dans la b^{e} grappe dans population. $\mathbf{x}_{bh}$ est un ensemble de variables auxiliaires disponibles dans l'enquête ainsi que dans le recensement, qui doivent généralement être contemporains ; u_{bh} est le terme d'erreur aléatoire représentant la part de Y_{bh} qui ne peut pas être expliquée par $\mathbf{x}_{bh}$. Les données sur le revenu et les dépenses possédant presque invariablement une distribution asymétrique, une transformation (habituellement logarithmique) est appliquée pour les rendre plus symétriques.

Les ménages pour lesquels des données sur le revenu ou les dépenses par tête sont recueillies sont rarement indépendants, et forment des grappes naturelles, souvent définies administrativement. Les ménages qui sont proches les uns des autres ou appartiennent à la même grappe ont tendance à être similaires à de nombreux égards. Dans les données d'enquête, les grappes sont habituellement aussi les unités primaires d'échantillonnage (UPE) du plan de sondage. Afin de tenir compte de la mise en grappe des ménages, on suppose habituellement que le terme d'erreur aléatoire u_{bh} du modèle de régression a la spécification suivante :

$$u_{bh} = v_b + e_{bh} \tag{6}$$

où v et e sont indépendants l'un de l'autre et non corrélés à $\mathbf{x}_{bh}$, v_b est le terme d'erreur appartenant en commun au b^{e} groupe ou grappe (par exemple barangay pour les Philippines) et e_{bh} est l'erreur au niveau du ménage à l'intérieur de la grappe. L'importance de ces termes est mesurée par leur variance ou les composantes de leur variance, σ_v^2 et σ_e^2, respectivement. Diverses méthodes existent pour estimer ces variances. Nous abordons ce sujet important aux sections qui suivent.

Le modèle (5) peut s'écrire

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + v_b + e_{bh} \tag{7}$$

dont la forme est similaire à celle du modèle au niveau de l'unité ou du modèle de régression à erreurs emboîtées mentionné à la section précédente. Cependant, si la forme du modèle est similaire, le groupe auquel il est fait référence est différent, par exemple Y_{ah} renvoie au h^{e} ménage dans le a^{e} petit domaine, tandis que Y_{bh} renvoie au h^{e} ménage dans la b^{e} grappe. Les grappes, fondées sur le plan de sondage, sont habituellement nettement plus petites que les domaines pour lesquels des estimations sur petits domaines sont recherchées et, généralement, elles ne sont pas toutes échantillonnées (contrairement aux petits domaines qui le sont presque tous). Par exemple, aux Philippines, des estimations sont demandées au niveau municipal, qui comprend les barangays ou grappes.

4. La méthode ELL

Pour la méthode ELL, l'estimation du paramètre de régression $\boldsymbol{\beta}$ est donnée dans Elbers et coll. (2002, page 11, note en bas de page 8) et dans le logiciel POVMAP développé pour la méthode ELL par Zhao (2006), sous la forme

$$\hat{\boldsymbol{\beta}}_{\mathrm{ELL}} = \left(\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{X}_b \right)^{-1} \left(\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{y}_b \right) \tag{8}$$

et la matrice de variance-covariance correspondante sous la forme

$$\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathrm{ELL}}) = \mathbf{D} \left[\left(\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{W}_b \mathbf{X}_b \right)^{-1} \right] \mathbf{D} \tag{9}$$

où $\mathbf{V}_b = (\sigma_e^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{1}_{n_b} \mathbf{1}'_{n_b})$, (σ_υ^2) est la variance au niveau de la grappe, tandis que (σ_e^2) est la variance au niveau du ménage, $\mathbf{I}_{n_b}$ est la matrice identité, $\mathbf{1}'_{n_b} = (1 \ldots 1)$ est un vecteur constant, $\mathbf{D} = (\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{X}_b)^{-1}$, $\mathbf{X}_b = (\mathbf{x}_{b1}, \ldots, \mathbf{x}_{bn_b})'$; $\mathbf{y}_b = (y_{b1}, \ldots, y_{bn_b})'$; $\mathbf{W}_b$ est une matrice diagonale des poids de sondage ; m est le nombre de grappes dans l'échantillon et n_b est le nombre de ménages dans chaque grappe échantillonnée. L'équation (8) repose sur l'hypothèse que $\mathbf{V}_b$ est connue. En pratique, nous devons estimer σ_e^2 et σ_v^2 pour obtenir l'estimateur $\hat{\mathbf{V}}_b$. Nous notons que l'expression de la variance dans (9) est dérivée sous un modèle hypothétique vaguement spécifié pour l'échantillon (voir Elbers et coll. 2002). Sous la méthode ELL, l'ajustement du modèle de revenu/dépenses (7) comprend l'obtention de l'estimation initiale de $\boldsymbol{\beta}$ par la méthode des moindres carrés pondérés (MCP) et l'utilisation des résidus du modèle initial pour estimer la matrice de covariance $\mathbf{V}_b$ nécessaire pour obtenir $\hat{\boldsymbol{\beta}}_{\mathrm{ELL}}$. Les estimations des variances au niveau de la grappe (σ_v^2) et au niveau du ménage (σ_e^2) sont calculées par Elbers et coll. (2002) de la façon suivante :

$$\hat{\sigma}_v^2 = \max\left(\frac{\sum_b w_b (u_{b.} - u_{..})^2}{\sum_b w_b (1 - w_b)} - \frac{\sum_b w_b (1 - w_b)\, \tau_b^2}{\sum_b w_b (1 - w_b)} \; ; \; 0 \right) \tag{10}$$

où $\tau_b^2 = \sum_h (e_{bh} - e_{b.})^2 / (n_b (n_b - 1))$; $w_b = \sum_h w_{bh} / \sum_b \sum_h w_{bh}$ sont les poids de sondage transformés par grappe, dont la somme sur les grappes est égale à un, et w_{bh} représente les poids de sondage changés d'échelle, dont la somme est égale à la taille totale de l'échantillon. Ici, $u_{b.} = \sum_h u_{bh}$ et $u_{..} = \sum_b \sum_h u_{bh}$ (qui est égal à zéro), où u_{bh} est défini comme dans l'équation (6).

Elbers et coll. (2002) ont proposé deux moyens de générer l'estimation de la composante de la variance au niveau du ménage, soit le calcul « direct » qui est désigné par $(\hat{\sigma}_e^2)$ ou le calcul fondé sur un modèle hétéroscédastique $(\hat{\sigma}_{e,bh}^2)$. Le calcul direct s'appuie sur la différence entre l'erreur quadratique moyenne estimée d'après la régression MCP initiale et l'estimation calculée de σ_v^2, tandis que le calcul fondé sur un modèle hétéroscédastique s'appuie sur une fonction de lien de type logistique pour borner la variance comme il suit :

$$\sigma_{e,bh}^2(\mathbf{z}_{bh}, \boldsymbol{\alpha}, A, B) = \left[\frac{A \exp(\mathbf{z}'_{bh}\boldsymbol{\alpha}) + B}{1 + \exp(\mathbf{z}'_{bh}\boldsymbol{\alpha})} \right] \tag{11}$$

où A et B sont les bornes supérieure et inférieure, respectivement, estimées par le vecteur de paramètres $\boldsymbol{\alpha}$ en utilisant la méthode classique du pseudo-maximum de vraisemblance (Elbers et coll. 2003) et où $\boldsymbol{z}_{bh}$ représente les variables auxiliaires. Elbers et coll. soutiennent qu'imposer une borne minimale nulle et une borne maximale égale à $A^* = (1,05)\ \max\{e_{bh}^2\}$ produit en général des estimations semblables des paramètres $\boldsymbol{\alpha}$. Cette contrainte permet d'estimer la forme plus simple

$$\ln\left[\frac{e_{bh}^2}{A^* - e_{bh}^2}\right] = \boldsymbol{z}_{bh}'\boldsymbol{\alpha} + r_{bh} \qquad (12)$$

où r_{bh} est un terme d'erreur et les autres variables sont les mêmes que celles définies précédemment. Dans la plupart des projets de cartographie de la pauvreté de la Banque mondiale, de légères modifications sont généralement apportées, par exemple l'ajout d'une constante δ à e_{bh}^2 dans le modèle (11).

En utilisant le modèle (12), et en appliquant la méthode delta, $\hat{\sigma}_{e,bh}^2$ est calculé sous la forme :

$$\hat{\sigma}_{e,bh}^2 = \left[\frac{A^* C_{bh}}{1 + C_{bh}}\right] + \frac{1}{2}\hat{\sigma}_r^2\left[\frac{A^* C_{bh}(1 - C_{bh})}{(1 + C_{bh})^3}\right] \qquad (13)$$

où $C_{bh} = \exp\{\boldsymbol{z}_{bh}'\hat{\boldsymbol{\alpha}}\}$, et $\hat{\sigma}_r^2$ est la variance estimée des résidus sous le modèle (12). Si la composante de la variance au niveau du ménage est fondée sur un modèle hétéroscédastique, alors $\mathbf{V}_b = (\sigma_{e,bh}^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{1}_{n_b}\mathbf{1}_{n_b}')$. La modélisation de l'hétéroscédasticité est effectuée en supposant que la variation au niveau du ménage dépend de certaines covariables.

Comme nous l'exposons plus en détail à l'annexe, la façon dont la matrice de pondération $\mathbf{W}_b$ entre dans le calcul de l'équation (9) susmentionnée aboutit à une matrice de covariance estimée asymétrique. Une approche un peu meilleure basée sur le « pseudo-maximum de vraisemblance », qui est décrite par Pfeffermann, Skinner, Holmes, Goldstein et Rasbash (1998), consiste à décomposer $\mathbf{X}_b'\mathbf{V}_b^{-1}\mathbf{X}_b$ en sommes distinctes de carrés et de produits croisés, et à pondérer chacune de manière appropriée – si nous écrivons $\mathbf{V}_b^{-1} = c\,\mathbf{I}_{n_b} + d\,\mathbf{1}_{n_b}\mathbf{1}_{n_b}'$, alors la pondération appropriée est $c\mathbf{X}_b'\mathbf{W}_b\mathbf{X}_b + d\mathbf{X}_b'\mathbf{W}_b\mathbf{1}_{n_b}\mathbf{1}_{n_b}'\mathbf{W}_b\mathbf{X}_b$.

Puisque la version ELL, $\mathbf{W}_b\mathbf{V}_b^{-1}$, n'est en général pas symétrique, dans l'équation (9), $\mathbf{D}$ ne l'est pas non plus. Donc, la matrice de covariance supposée de $\hat{\boldsymbol{\beta}}_{\mathrm{ELL}}$, $\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathrm{ELL}})$, n'est pas symétrique non plus. Le logiciel POVMAP essaye de résoudre ce problème en prenant la moyenne de $\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathrm{ELL}})$ et de sa transposée, ce qui force la matrice à être symétrique.

Mentionnons de nouveau que, sous la méthode ELL, l'ajustement de la régression aux données d'enquête et l'estimation des composantes de la variance ne constituent que la première phase. La phase suivante comporte la prédiction au niveau du ménage en se basant sur les données de recensement complètes et l'agrégation au niveau du petit domaine.

Les méthodes d'ajustement aux données d'enquête (calcul de l'estimation de $\boldsymbol{\beta}$ et de sa matrice de variance-covariance correspondante) pour les trois méthodes de régression proposées comme alternative de la méthode ELL sont présentées aux sections qui suivent.

5. Autres méthodes d'ajustement

5.1 La méthode du meilleur prédicteur linéaire sans biais pseudo-empirique

You et Rao (2002) ont proposé un estimateur de la moyenne de petit domaine en établissant un estimateur de $\boldsymbol{\beta}$ basé sur le modèle au niveau de l'unité (4). L'établissement de l'estimateur de $\boldsymbol{\beta}$ débute par le calcul du meilleur prédicteur linéaire sans biais (MPLSB) de v_a sachant les paramètres $\boldsymbol{\beta}$, σ_e^2 et σ_v^2 tirés du modèle au niveau du domaine agrégé (pondéré par les poids de sondage) :

$$\bar{Y}_{aw} = \bar{\mathbf{x}}_{aw}'\boldsymbol{\beta} + v_a + \bar{e}_{aw} \qquad (14)$$

qui procède comme il suit :

$$\hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2) = \gamma_{aw}(\bar{y}_{aw} - \bar{\mathbf{x}}_{aw}'\boldsymbol{\beta}) \qquad (15)$$

où $\bar{\mathbf{x}}_{aw} = \sum_{h=1}^{n_a} w_{ah}\mathbf{x}_{ah}$, $\bar{y}_{aw} = \sum_{h=1}^{n_a} w_{ah} y_{ah}$, $\gamma_{aw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\,\delta_a^2)$, $w_{ah} = \tilde{w}_{ah}/\sum_{h=1}^{n_a}\tilde{w}_{ah}$, $\delta_a^2 = \sum_{h=1}^{n_a} w_{ah}^2$ et $\tilde{w}_{ah}$ sont les poids de sondage au niveau de l'unité ; vient ensuite la résolution de l'équation d'estimation pondérée par les poids de sondage pour trouver $\boldsymbol{\beta}$:

$$\sum_{a=1}^{k}\sum_{h=1}^{n_a}\tilde{w}_{ah}\mathbf{x}_{ah}[y_{ah} - \mathbf{x}_{ah}'\boldsymbol{\beta} - \hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2)] = 0 \qquad (16)$$

d'après laquelle l'estimateur de $\boldsymbol{\beta}$ est obtenu sous la forme

$$\hat{\boldsymbol{\beta}}_w = \left\{\sum_{a=1}^{k}\sum_{h=1}^{n_a}\mathbf{x}_{ah}\mathbf{z}_{ah}'\right\}^{-1}\left\{\sum_{a=1}^{k}\sum_{h=1}^{n_a}\mathbf{z}_{ah}y_{ah}\right\} \qquad (17)$$

où $\mathbf{z}_{ah} = \tilde{w}_{ah}(\mathbf{x}_{ah} - \gamma_{aw}\bar{\mathbf{x}}_{ah})$. La matrice de covariance correspondante s'écrit alors :

$$\begin{aligned}\boldsymbol{\Phi}_w = {} & \sigma_e^2\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a}\mathbf{x}_{ah}\mathbf{z}_{ah}'\right)^{-1} \\ & \left(\sum_{a=1}^{k}\sum_{h=1}^{n_a}\mathbf{z}_{ah}\mathbf{z}_{ah}'\right)\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a}\mathbf{x}_{ah}\mathbf{z}_{ah}'\right)^{-1} \\ & + \sigma_v^2\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a}\mathbf{x}_{ah}\mathbf{z}_{ah}'\right)^{-1} \\ & \left\{\sum_{a=1}^{k}\left(\sum_{h=1}^{n_a}\mathbf{z}_{ah}\right)\left(\sum_{a=1}^{n_a}\mathbf{z}_{ah}\right)'\right\}\left\{\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a}\mathbf{x}_{ah}\mathbf{z}_{ah}'\right)^{-1}\right\}'. \qquad (18)\end{aligned}$$

Les composantes de la variance sont estimées en utilisant la méthode 3 de Henderson (Henderson 1953) pour produire des estimations sans biais, même en présence d'éléments corrélés dans le modèle. Les estimateurs des composantes de la variance sont les suivants :

$$\hat{\sigma}^2_{eH} = (n-k-p+1)^{-1}\sum_{a=1}^{k}\sum_{h=1}^{n_a}\hat{\varepsilon}^2_{ah} \qquad (19)$$

où $\{\hat{\varepsilon}^2_{ah}\}$ représente les résidus de la régression par les moindres carrés ordinaires (MCO) de $(y_{ah} - \bar{y}_a)$ sur $\{x_{ah1} - \bar{x}_{a.1}, ..., x_{ahp} - \bar{x}_{a.p}\}$ et $(\bar{y}_a, \bar{x}_{a.1}, ..., \bar{x}_{a.p})$ sont les moyennes d'échantillon dans le a^{e} groupe.

$$\hat{\sigma}^2_{vH} = n_*^{-1}\left[\sum_{a=1}^{k}\sum_{h=1}^{n_a}\hat{u}^2_{ah} - (n-p)\hat{\sigma}^2_{eH}\right] \qquad (20)$$

où $n_* = n - \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\Sigma_{a=1}^{k} n_a^2\,\bar{\mathbf{x}}_a\,\bar{\mathbf{x}}'_a]$ avec $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_k)$, et les $\{\hat{u}_{ah}\}$ sont les résidus de la régression par les MCO de y_{ah} sur $\{x_{ah1}, ..., x_{ahp}\}$. Pour le modèle (7), l'indice inférieur a est remplacé par b.

Cependant, les estimateurs de Henderson susmentionnés ne tiennent pas compte des poids de sondage. Pour contourner ce problème, You et coll. (2003) ont proposé une technique d'estimation qui consiste à étendre la méthode du pseudo-MPLSB en intégrant les poids dans l'estimation des composantes de la variance, ce que nous décrivons à la section suivante.

5.2 La méthode itérative à équations d'estimation pondérées

L'estimateur proposé par You et coll. (2003) est semblable à l'estimateur pseudo-MPLSB, excepté qu'il intègre les poids de sondage dans le calcul des composantes de la variance, et produit l'estimation des paramètres $\boldsymbol{\beta}$ et des composantes de la variance en se fondant sur une approche itérative à équation d'estimation pondérées (IEEP). Les auteurs ont calculé les estimateurs de σ_e^2 et σ_v^2 de la façon suivante :

$$\hat{\sigma}^{2(t)}_{ew} = \frac{\sum_{a=1}^{k}\sum_{h=1}^{n_a}\tilde{w}_{ah}[y_{ah} - \bar{y}_{aw} - (\mathbf{x}_{ah} - \bar{\mathbf{x}}_{aw})'\hat{\boldsymbol{\beta}}^{(t-1)}]^2}{\sum_{a=1}^{k}\left[(1-\delta_a^2)\sum_{h=1}^{n_a}\tilde{w}_{ah}\right]}$$

$$\equiv \tilde{\sigma}^{2(t)}_{ew}(\boldsymbol{\beta}) \qquad (21)$$

et

$$\hat{\sigma}^{2(t)}_{vw} = \frac{1}{k}\sum_{a=1}^{k}\tilde{v}^2_{aw} + \frac{\tilde{\sigma}^{2(t-1)}_{vw}}{k}\sum_{a=1}^{k}(\gamma_{aw}-1)^2 + \frac{\tilde{\sigma}^{2(t)}_{ew}}{k}\sum_{a=1}^{k}\delta_a^2\gamma^2_{aw}$$

$$\equiv \tilde{\sigma}^{2(t)}_{vw}(\tilde{v}_w, \sigma_e^2, \sigma_v^2). \qquad (22)$$

Les estimations pondérées de $\boldsymbol{\beta}$, σ_e^2 et σ_v^2 sont obtenues simultanément en suivant des étapes itératives de mise à jour, t représentant dans l'équation susmentionnée la t^{e} itération. Puisque les composantes de la variance σ_v^2 et σ_e^2 sont inconnues, les estimations de départ pour les étapes d'itération sont générées par la méthode de Henderson. De nouveau, comme pour le pseudo-MPLSB, dans la formule du modèle de régression ELL (7), l'indice inférieur a est remplacé par b.

Cette approche est semblable à la méthode des moindres carrés généralisés itérés pondérés par les probabilités de sélection (MCGPPS) proposée par Pfeffermann et coll. (1998) pour ajuster des modèles multiniveaux où le procédé d'estimation tient compte des probabilités de sélection inégales à chaque degré d'échantillonnage et comporte une itération entre le paramètre $\boldsymbol{\beta}$ et les composantes de la variance jusqu'à la convergence. Pfeffermann, Moura et Silva (2006) proposent aussi une approche fondée sur un modèle qui comprend le calcul du modèle hiérarchique pour des données d'échantillon données sous la forme d'une fonction du modèle de population et des probabilités de sélection, puis l'ajustement du modèle sur données d'échantillon selon une approche bayésienne en se servant de l'algorithme de Monte Carlo par chaîne de Markov.

5.3 Méthode de régression généralisée sur données d'enquête

Une autre approche pour produire l'estimateur du paramètre β et de sa variance est la méthode fondée sur le plan de sondage pour l'ajustement des modèles de régression (Lohr 1999). Cette technique est utilisée à l'heure actuelle dans les progiciels Stata, Sudaan et WesVar, par exemple. L'estimateur de $\boldsymbol{\beta}$ donné ci-après est l'estimateur par la régression pondérée par les poids de sondage pour un modèle avec structure de variance homoscédastique et observations non corrélées dans la population.

$$\hat{\boldsymbol{\beta}}_S = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \qquad (23)$$

Cet estimateur n'est pas dérivé sous le modèle spécifié par (7), même sous l'hypothèse de variances homoscédastiques pour les erreurs au niveau du ménage. L'estimation linéarisée/robuste de variance pour $\hat{\boldsymbol{\beta}}_S$ est basée sur l'estimateur de variance sous le plan de sondage pour un total, donné par

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_S) = \mathbf{D}\left\{\frac{m}{m-1}\sum_{b=1}^{m}\left(\sum_{h=1}^{n_b}w_{bh}\mathbf{d}_{bh}\right)'\left(\sum_{h=1}^{n_b}w_{bh}\mathbf{d}_{bh}\right)\right\}\mathbf{D} \qquad (24)$$

où $\mathbf{d}_{bh} = \hat{e}_{bh}\mathbf{x}_{bh}$; $\hat{e}_{bh}$ est le résidu de la régression par moindres carrés pondérés (MCP) ; $\mathbf{x}_{bh}$ est un vecteur de variables indépendantes ; w_{bh} est un poids de sondage ; $\mathbf{D} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$ et $\mathbf{W}$ est une matrice diagonale de poids de sondage.

La méthode de régression généralisée sur données d'enquête diffère des autres techniques en ce qui a trait au calcul des estimations et produit des estimations sans calcul des composantes de la variance, σ_v^2 et σ_e^2. Comme nous

l'avons montré plus haut, les équations pour l'estimateur du paramètre $\boldsymbol{\beta}$ et sa matrice de covariance estimée correspondante ne font intervenir que la matrice de poids de sondage $\mathbf{W}$. La matrice de covariance estimée donnée par (24) est souvent appelée estimateur sandwich.

6. Comparaison des méthodes d'ajustement de modèle

La méthode ELL est décrite comme une méthode d'estimation par moindres carrés généralisés (MCG) pondérée. Toutefois, comme nous l'avons souligné plus haut, les poids de sondage ne sont pas intégrés correctement dans le processus d'estimation, ce qui rend ininterprétables les éléments de certaines matrices intervenant dans l'estimation et rend asymétrique la matrice de covariance estimée. Dans la méthode ELL d'estimation des composantes de la variance, les poids ne sont pris en compte qu'au niveau de la grappe. Les deux moyens (calcul direct et calcul fondé sur un modèle d'hétéroscédasticité) utilisés par Elbers et coll. pour produire la composante de la variance au niveau du ménage n'intègrent pas les poids de sondage. Dans le calcul direct, la composante de la variance au niveau du ménage est déterminée d'après le résidu de la régression pondérée par les poids de sondage (MCP) réalisée à l'étape préliminaire et l'estimation pondérée de la composante au niveau de la grappe. Le calcul fondé sur l'hétéroscédasticité repose sur la modélisation du carré des résidus de la régression par MCP.

Alors que la méthode ELL suit une procédure d'estimation de type MCG, les méthodes pseudo-MPLSB et IEEP suivent la procédure des équations d'estimation généralisées (EEG) (Liang et Zeger 1986) avec utilisation d'une matrice de corrélation de travail échangeable, c'est-à-dire que tous les éléments hors diagonale de la matrice de corrélation à l'intérieur des grappes sont égaux, et dans le cas des méthodes pseudo-MPLSB et IEEP, sont égaux à $\sigma_v^2/(\sigma_v^2+\sigma_e^2)$. La matrice de corrélation de travail échangeable ou équicorrélée est l'une des matrices de corrélation de travail fréquentes présentées par Horton et Lipsitz (1999) dans un article où ils passent en revue divers progiciels pour l'ajustement de modèles de régression EEG.

Les deux méthodes, pseudo-MPLSB et IEEP, intègrent toutes deux les poids de sondage dans l'estimation du paramètre $\boldsymbol{\beta}$ et de l'erreur-type correspondante, quoique la méthode pseudo-MPLSB recourt à la méthode de Henderson dans l'estimation des composantes de la variance. Alors que la méthode de Henderson produit des estimations non pondérées des composantes de la variance, la méthode IEEP intègre les poids de sondage itérativement en partant de l'estimation des composantes de la variance pour le calcul de l'erreur-type de l'estimation du paramètre de régression.

Les publications portant sur l'application des méthodes pseudo-MPLSB et IEEP à des ensembles de données réelles sont très peu nombreuses. Celles qui existent traitent les grappes comme étant les petits domaines et s'appuient souvent sur l'ensemble de données présentées dans Battese, Harter et Fuller (1988), qui contient des renseignements sur les hectares de maïs et de soja par segments pour les comtés du Centre-Nord de l'Iowa et repose sur l'hypothèse d'un échantillonnage aléatoire simple à l'intérieur des domaines ou des grappes. Fait exception l'article récent de Militino, Ugarte, Goicoa et Gonzalez-Audicana (2006), dans lequel le pseudo-MPLSB est appliqué pour estimer la superficie totale occupée par les oliviers à Navarra, en Espagne, où, comme dans Battese et coll., les unités sont autopondérées. En général, pour l'estimation de la pauvreté, les méthodes pseudo-MPLSB et IEEP doivent être appliquées dans des situations plus complexes, puisque les grappes d'échantillonnage et les petits domaines ne sont pas identiques et que l'échantillon n'est pas autopondéré. Dans l'exemple de la section suivante, les grappes (barangays) diffèrent des petits domaines (municipalités), les grappes sont des sous-unités du petit domaine et le plan d'échantillonnage n'est pas autopondéré.

La méthode de régression RGE est l'une des méthodes d'estimation les moins compliquées, car elle emploie une procédure par les moindres carrés pondérés avec utilisation de l'estimateur sandwich pour estimer la variance de l'estimateur du paramètre de régression. Comme nous l'avons mentionné plus haut, cette méthode diffère des autres en ce que l'estimation des paramètres de régression et de leurs erreurs-types correspondantes est produite sans calculer les composantes de variance.

Selon la discussion qui précède, pour toutes les méthodes prises en considération, les procédures d'estimation fondées sur des données d'enquête pour le paramètre $\boldsymbol{\beta}$ et son erreur-type correspondante sont théoriquement valables étant donné les hypothèses sur lesquelles elles reposent, sauf dans le cas de la méthode ELL pour laquelle existent certaines incohérences dans l'estimation du paramètre $\boldsymbol{\beta}$ et de la covariance de $\hat{\boldsymbol{\beta}}$.

7. Application aux données réelles

À la présente section, nous comparons les quatre méthodes de régression étudiées (dont une contient deux variantes de la méthode ELL) en utilisant les données de la Family Income and Expenses Survey (FIES) de 2000 aux Philippines. La FIES est une enquête de portée nationale réalisée tous les trois ans par le National Statistics Office (NSO) des Philippines. L'enquête est conçue pour recueillir des données sur les revenus et les dépenses des familles, ainsi que des renseignements sur les facteurs ayant une

incidence sur les revenus et dépenses. Les ménages sélectionnés sont interviewés en deux opérations distinctes, couvrant chacune une période d'une demi-année, afin de tenir compte des variations saisonnières des revenus et des dépenses. Pour la FIES de 2000, les interviews ont été réalisées en juillet 2000 pour la période allant du 1er janvier au 30 juin, et en janvier 2001 pour la période allant du 1er juillet au 31 décembre. Le plan de sondage de la FIES s'appuyait sur l'échantillonnage aléatoire stratifié à plusieurs degrés. Les barangays, qui sont les unités primaires d'échantillonnage (UPE), sont répartis en une strate urbaine et une strate rurale dans chaque province et sélectionnés par échantillonnage systématique avec probabilité proportionnelle à la taille. Les grands barangays sont en outre subdivisés en secteurs de recensement et soumis à un échantillonnage supplémentaire avant l'étape finale durant laquelle les ménages sont échantillonnés systématiquement en se servant de la liste des ménages du Recensement de la population de 1995. La non-réponse à l'interview était de 3,4 % seulement, 39 615 ménages échantillonnés ayant pu être interviewés aux deux visites de l'enquête. La non-réponse partielle a été corrigée par imputation déterministe, c'est-à-dire qu'une entrée manquant pour une question particulière a été déduite d'après les réponses obtenues pour d'autres items du questionnaire.

Les variables auxiliaires utilisées dans le présent article sont celles incluses dans le modèle formulé par Haslett et Jones (2005), qui a été ajusté sans utiliser le logiciel POVMAP pour le projet de cartographie de la pauvreté par petit domaine aux Philippines. Les variables auxiliaires comprenaient les caractéristiques du ménage ainsi que les moyennes municipales (dans lesquelles les données sur les ménages utilisées ont la même valeur pour chaque ménage échantillonné dans une municipalité donnée, c'est-à-dire un petit domaine). Ces variables auxiliaires sont non seulement dérivées des données de la FIES, mais aussi de celles de l'Enquête sur la population active (EPA) de 2000 et du Recensement de la population et du logement (RPL) des Philippines. L'EPA est conçue pour recueillir des données sur les caractéristiques socioéconomiques de la population de plus de 15 ans. Le NSO la réalise trimestriellement par interview sur place en utilisant la semaine précédente comme période de référence. Comme elles faisaient partie de l'Integrated Survey of Households (NSCB 2000), les enquêtes de juillet 2000 et de janvier 2001 ont été réalisées auprès du même échantillon de ménages que la FIES de 2000. Donc, les deux ensembles de données peuvent être fusionnés pour former un ensemble plus riche de variables auxiliaires. Des variables auxiliaires supplémentaires ont également été tirées du Recensement de la population et du logement de 2000 sous forme de moyennes municipales. Les moyennes pour les variables de recensement du questionnaire abrégé ainsi que du questionnaire complet ont été calculées au niveau municipal pour créer de nouveaux ensembles de données pouvant être fusionnés avec l'ensemble de variables auxiliaires provenant de la FIES et de l'EPA.

Les tableaux 1, 2 et 3 donnent les estimations du paramètre ($\boldsymbol{\beta}$) et des erreurs-types correspondantes, ainsi que les estimations des composantes de la variance aux niveaux national, régional et provincial, respectivement. Le tableau 2 donne les résultats pour l'un des 16 modèles ajustés au niveau régional (il existait 16 régions aux Philippines en 2000). De même, le tableau 3 donne les résultats de l'un des 20 modèles provinciaux formulés pour les 20 provinces sélectionnées. Afin de normaliser les comparaisons, exactement le même ensemble de variables prédictives est utilisé pour toutes les méthodes d'ajustement du modèle. (Il existe cinq ensembles d'estimations du paramètre, bien que nous n'ayons examiné que quatre méthodes fondamentales, parce que la méthode ELL est utilisée avec et sans hétéroscédasticité.) Notons qu'en pratique, dans le cas de la méthode ELL, on subdivise souvent les données d'enquête et on ajuste des modèles distincts à chaque sous-échantillon, par exemple à chacune des strates définies selon les 16 régions des Philippines, voire même des modèles au niveau provincial. Cette approche peut produire des modèles surajustés et des erreurs-types présentant un biais par défaut pour les estimations sur petits domaines. Pour l'analyse présentée ici, nous n'avons ajusté qu'un seul modèle (c'est-à-dire le modèle au niveau national). En pratique, des modèles intermédiaires comportant certains effets régionaux, mais non tous, semblent être ceux qui donnent les meilleurs résultats. Voir par exemple Haslett et Jones (2005).

Afin d'évaluer les différences entre les estimations produites par les diverses méthodes, nous effectuons une comparaison informelle de la « signification » des diverses estimations de β en soustrayant de l'estimation produite par une méthode la moyenne des estimations obtenue par les autres méthodes, puis en divisant le résultat par l'erreur-type de la méthode en question. Au niveau national (tableau 1), les estimations des coefficients de régression produites par les diverses méthodes diffèrent de manière significative les unes des autres pour un certain nombre de variables indépendantes. La méthode RGE a tendance à produire, pour la majorité des variables, des estimations des coefficients de régression qui diffèrent de manière significative de celles obtenues par les autres méthodes. Comme nous l'avons souligné plus haut, l'estimateur RGE est l'estimateur par la régression pondérée par les poids de sondage pour un modèle avec structure de variance homoscédastique et observations non corrélées dans la population, de sorte que cet estimateur n'est pas dérivé sous le modèle spécifié par (7). Toutefois, il s'agit de l'estimateur le plus prudent, car il donne l'erreur-type la plus élevée pour toutes les caractéristiques au niveau du ménage. Par ailleurs, la méthode IEEP produit l'erreur-type estimée la plus grande

pour toutes les moyennes municipales. La méthode ELL_H (ELL avec hétéroscédasticité) peut être considérée comme étant la moins prudente, puisqu'elle produit les erreurs-types les plus faibles pour tous les coefficients de régression estimés des caractéristiques au niveau du ménage, ainsi que pour les moyennes municipales, sauf dans le cas de deux variables, pour lesquelles la méthode RGE a produit les estimations les plus faibles. En ce qui concerne les estimations des composantes de la variance, la méthode ELL produit la variance estimée au niveau de la grappe la plus faible, correspondant à environ 92 % de celle obtenue par la méthode du pseudo-MPLSB et à 86 % de celle obtenue par la méthode IEEP. Pour ce qui est de la variance au niveau du ménage, la méthode IEEP est celle qui produit l'estimation la plus faible.

Au niveau régional, les estimations des coefficients de régression sont généralement semblables pour les diverses méthodes d'estimation, si ce n'est que les méthodes RGE et (ou) ELL_H produisent, pour quelques variables, des estimations significativement différentes de celles obtenues par les autres méthodes. Comme dans le cas des erreurs-types estimées au niveau national, la méthode RGE a tendance à être la plus prudente pour la majorité des modèles de niveau régional, ayant donné les erreurs-types estimées les plus élevées pour la plupart des coefficients de régression des caractéristiques du ménage. La méthode IEEP produit l'erreur-type estimée la plus grande pour la plupart des coefficients des moyennes municipales. La méthode ELL_H produit les erreurs-types les plus faibles pour la majorité des coefficients de régression des caractéristiques du ménage et des moyennes municipales. La méthode ELL a tendance à produire l'estimation la plus faible de la variance au niveau de la grappe, les ratios par rapport au pseudo-MPLSB et à la méthode IEEP variant d'environ 82 % à 100 %. La méthode IEEP est encore celle donnant la variance au niveau du ménage la plus faible.

Tableau 1
Estimations au niveau national des paramètres de régression avec les erreurs-types et les composantes de la variance pour les quatre méthodes. *Valeur différente pour chaque ménage (moyenne = 0,1576633) **Basé sur les résultats pour la méthode ELL

Variables	ELL(sans hétérosc.)		ELL(avec hétérosc.)		Pseudo-MPLSB		IEEP		RGE	
explicatives	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	Erreur-type
famsize	-0,11867	0,00181	-0,12034	0,00165	-0,11875	0,00183	-0,11888	0,00180	-0,11405	0,00216
famsizesqc	0,00937	0,00039	0,00981	0,00036	0,00938	0,00039	0,00939	0,00038	0,00898	0,00044
type_mult	0,03876	0,01697	0,03703	0,01588	0,03699	0,01717	0,03466	0,01692	0,11460	0,02194
per_kids	-0,20342	0,01476	-0,20818	0,01322	-0,20293	0,01491	-0,20216	0,01467	-0,22864	0,01617
roof_light	-0,06314	0,01291	-0,05808	0,01056	-0,06263	0,01306	-0,06175	0,01287	-0,09251	0,01413
per_61up	-0,09402	0,01420	-0,08331	0,01371	-0,09392	0,01435	-0,09389	0,01412	-0,09705	0,01698
roof_strong	0,05882	0,01135	0,05633	0,00962	0,05944	0,01148	0,06030	0,01132	0,03118	0,01293
wall_light	-0,05459	0,01182	-0,04979	0,00975	-0,05426	0,01195	-0,05392	0,01178	-0,06286	0,01353
wall_salvaged	-0,10814	0,02505	-0,11327	0,02058	-0,10748	0,02533	-0,10607	0,02495	-0,15702	0,02925
wall_strong	0,14248	0,01051	0,12964	0,00910	0,14274	0,01063	0,14319	0,01047	0,12662	0,01284
fa_xs	-0,17052	0,00941	-0,16756	0,00782	-0,17144	0,00952	-0,17236	0,00939	-0,14213	0,01110
fa_s	-0,08368	0,00861	-0,08242	0,00725	-0,08403	0,00871	-0,08454	0,00857	-0,06667	0,00964
fa_l	0,09016	0,00908	0,08478	0,00792	0,09065	0,00918	0,09106	0,00904	0,07848	0,01047
fa_xl	0,16959	0,01104	0,15404	0,00992	0,17034	0,01117	0,17121	0,01100	0,14300	0,01334
fa_xxl	0,27072	0,01144	0,24485	0,01094	0,27172	0,01157	0,27274	0,01140	0,23913	0,01457
fa_xxxl	0,36190	0,01371	0,31369	0,01286	0,36270	0,01387	0,36382	0,01367	0,32123	0,02025
all_eled	0,19084	0,01535	0,20497	0,01307	0,19031	0,01551	0,18964	0,01527	0,21344	0,01831
all_hsed	0,42325	0,01250	0,43771	0,01083	0,42192	0,01263	0,42024	0,01244	0,48180	0,01475
all_coed	1,21591	0,01371	1,29368	0,01379	1,21324	0,01386	1,20935	0,01366	1,35022	0,01827
dom_help	0,60207	0,01629	0,61218	0,01886	0,60035	0,01645	0,59733	0,01620	0,70307	0,02656
head_male	-0,05878	0,00988	-0,04581	0,00932	-0,05862	0,00998	-0,05819	0,00982	-0,07410	0,01173
no_spouse	-0,09367	0,00987	-0,07376	0,00917	-0,09361	0,00997	-0,09351	0,00981	-0,09599	0,01123
hou_9600	0,28537	0,07654	0,25643	0,07375	0,28871	0,07911	0,28783	0,08066	0,31956	0,07941
hea_rel_mus	0,09058	0,02645	0,10859	0,02507	0,09753	0,02728	0,09731	0,02782	0,10196	0,02737
Per_eng	0,17273	0,06529	0,14561	0,06298	0,17782	0,06754	0,17799	0,06887	0,17076	0,06407
Hou_coelpg	0,37463	0,04348	0,39784	0,04210	0,37934	0,04494	0,37792	0,04581	0,42682	0,03711
Hou_own_ref	0,17716	0,10497	0,18342	0,10178	0,17189	0,10843	0,17329	0,11055	0,13791	0,09766
Hou_own_tel	1,39287	0,13356	1,42109	0,12987	1,38551	0,13723	1,38974	0,13989	1,23506	0,13019
Per_wor_prh	0,46957	0,15484	0,40302	0,14926	0,47517	0,16006	0,47208	0,16317	0,50814	0,15210
Per_ind_52	-0,76245	0,21708	-0,78120	0,21073	-0,76326	0,22410	-0,76307	0,22849	-0,73294	0,21214
const	9,54013	0,05525	9,54456	0,05290	9,53566	0,05698	9,53594	0,05791	9,52622	0,05613
Estimation des composantes de la variance	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage**	Niveau de la grappe**
	0,18461	0,04741	NA*	0,04741	0,18820	0,05172	0,18185	0,05498	0,18461	0,04741

Tableau 2
Estimations au niveau régional des paramètres de régression avec les erreurs-types et les composantes de la variance pour les quatre méthodes. *Valeur différente pour chaque ménage (moyenne = 0,18930) **Basé sur les résultats pour la méthode ELL

Variables	ELL(sans hétérosc.)		ELL(avec hétérosc.)		Pseudo-MPLSB		IEEP		RGE	
explicatives	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	explicatives	Bêta	Erreur-type	Bêta	Erreur-type
famsize	-0,12327	0,00760	-0,12934	0,00689	-0,12377	0,00752	-0,12380	0,00749	-0,11786	0,00997
famsizesqc	0,01096	0,00164	0,01190	0,00147	0,01101	0,00163	0,01102	0,00162	0,01030	0,00195
dom_help	0,81037	0,08873	0,75624	0,10986	0,80727	0,08784	0,80708	0,08751	0,84490	0,08911
wall_light	-0,06808	0,04289	-0,06390	0,03743	-0,06020	0,04272	-0,05973	0,04257	-0,14472	0,04226
wall_strong	0,13761	0,03745	0,15212	0,03469	0,14514	0,03737	0,14560	0,03725	0,06116	0,04249
fa_xs	-0,22074	0,04910	-0,22368	0,04518	-0,22723	0,04875	-0,22761	0,04858	-0,14856	0,05665
fa_s	-0,13540	0,03840	-0,12255	0,03344	-0,13775	0,03805	-0,13789	0,03791	-0,11059	0,04538
fa_l	0,09484	0,03709	0,08894	0,03429	0,09590	0,03676	0,09597	0,03663	0,08529	0,04122
fa_xl	0,16627	0,04315	0,15519	0,04072	0,16938	0,04284	0,16958	0,04269	0,13698	0,04897
fa_xxl	0,33706	0,04545	0,31196	0,04829	0,34173	0,04516	0,34201	0,04500	0,29156	0,05148
fa_xxxl	0,33103	0,06185	0,30377	0,06029	0,33762	0,06134	0,33801	0,06111	0,26052	0,06635
all_hsed	0,33987	0,05253	0,35591	0,04783	0,33807	0,05209	0,33796	0,05189	0,35776	0,04843
all_coed	1,21824	0,05734	1,24762	0,05842	1,20787	0,05692	1,20726	0,05671	1,32979	0,06227
per_kids	-0,24699	0,06440	-0,24047	0,05846	-0,24439	0,06371	-0,24424	0,06347	-0,27423	0,07050
per_61up	-0,14609	0,06126	-0,15938	0,05787	-0,14703	0,06063	-0,14708	0,06040	-0,13525	0,07124
hou_9600	1,13985	0,49103	1,27035	0,47888	1,14320	0,52137	1,14357	0,52172	1,07509	0,51937
Hou_own_ref	1,45233	0,24550	1,51020	0,23864	1,44986	0,26072	1,44985	0,26089	1,44779	0,23585
const	9,36877	0,20322	9,32363	0,19660	9,36597	0,21502	9,36569	0,21512	9,41385	0,21430
Estimation des composantes de la variance	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage**	Niveau de la grappe**
	0,19544	0,03073	NA*	0,03073	0,19052	0,03728	0,18902	0,03748	0,19544	0,03073

Tableau 3
Estimations au niveau provincial des paramètres de régression avec les erreurs-types et les composantes de la variance pour les quatre méthodes. *Valeur différente pour chaque ménage (moyenne = 0,23749) **Basé sur les résultats pour la méthode ELL

Variables	ELL(sans hétérosc.)		ELL(avec hétérosc.)		Pseudo-MPLSB		IEEP		RGE	
explicatives	Bêta	Erreur-type	Bêta	Erreur-type	Bêta	Bêta	Erreur-type	Bêta	Erreur-type	Bêta
famsize	-0,1450	0,0175	-0,1489	0,0156	-0,1452	0,0179	-0,1449	0,0171	-0,1413	0,0097
famsizesqc	0,0090	0,0063	0,0124	0,0067	0,0091	0,0065	0,0090	0,0062	0,0085	0,0055
fa_xs	-0,4549	0,1126	-0,3816	0,1010	-0,4552	0,1149	-0,4546	0,1095	-0,4479	0,0718
fa_s	-0,2550	0,0976	-0,2653	0,0794	-0,2545	0,0995	-0,2555	0,0951	-0,2693	0,1198
wall_light	-0,2055	0,0945	-0,1474	0,0778	-0,2057	0,0965	-0,2058	0,0919	-0,2063	0,1070
all_hsed	0,4007	0,1643	0,3531	0,1448	0,4015	0,1673	0,4006	0,1601	0,3891	0,1585
all_coed	1,5411	0,1677	1,8202	0,1769	1,5429	0,1709	1,5429	0,1635	1,5439	0,2326
Hou_own_tel	3,4373	1,0270	3,2630	1,0582	3,4265	1,0622	3,4274	0,9871	3,4392	0,5733
Per_wor_prh	-1,1075	1,1933	-1,5801	1,2008	-1,1049	1,2327	-1,1056	1,1483	-1,1150	0,8729
const	10,0976	0,1480	10,0798	0,1279	10,0988	0,1517	10,0981	0,1435	10,0872	0,1373
Estimation des composantes de la variance	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage	Niveau de la grappe	Niveau du ménage**	Niveau de la grappe**
	0,25753	0,01871	NA*	0,25753	0,26682	0,02079	0,24498	0,01671	0,25753	0,01871

Comme pour les estimations au niveau régional, les estimations des coefficients de régression au niveau provincial sont similaires, à l'exception de certaines différences pour les estimations par les méthodes RGE et ELL_H. Pour ce qui est des erreurs-types estimées des coefficients de régression, la méthode ELL_H produit encore les estimations les plus faibles pour la majorité des coefficients des caractéristiques du ménage ; toutefois, la méthode RGE (au lieu de la méthode ELL_H) produit maintenant les erreurs-types estimées les plus faibles pour la majorité des moyennes municipales. La méthode ELL continue d'avoir tendance à générer l'estimation la plus faible de la variance au niveau de la grappe pour la plupart des provinces, le ratio le plus petit par rapport au pseudo-MPLSB étant d'environ 53 % et par rapport à la méthode IEEP, d'environ 48 %. Pour un certain nombre de provinces, la méthode IEEP a tendance à produire l'estimation la plus faible de la variance au niveau de la grappe. En ce qui concerne la variance au niveau du ménage, la méthode IEEP produit encore l'estimation la plus faible. En général, les estimations de la variance au niveau de la grappe ont tendance à être plus variables au niveau provincial, ce qui est dû aux plus petites tailles d'échantillon.

En ce qui concerne les estimations de la pauvreté sur petits domaines, après l'application du modèle de régression aux données de recensement, les erreurs-types estimées dans la régression représentent une partie seulement des erreurs-types des estimations sur petits domaines. Il existe aussi une

variation au niveau de la grappe dans (7) qui doit être prise en considération (à divers degrés selon le niveau d'agrégation utilisé pour construire les petits domaines), ainsi qu'une variation au niveau du ménage. Ces sources supplémentaires de variation peuvent être évaluées par l'estimation des composantes de la variance. Comme il est montré ci-dessus, indépendamment du niveau (national, régional ou provincial) auquel le modèle est formulé, la méthode IEEP produit la variance au niveau du ménage la plus faible, tandis que la méthode ELL produit la variance au niveau de la grappe la plus faible. Puisque la variation au niveau de la grappe contribue habituellement plus fortement à l'erreur-type estimée au niveau du petit domaine, la méthode ELL est de nouveau la moins prudente. Nous notons que la variance au niveau du ménage sous la méthode ELL avec modèle d'hétéroscédasticité varie d'une unité à l'autre, si bien que nous présentons la valeur moyenne, et que le R^2 estimé pour le modèle d'hétéroscédasticité est négligeable, $R^2 = 0{,}03$, même au niveau national, de sorte qu'en ce qui concerne au moins l'ajustement du modèle de régression, cette méthode peut offrir quelques avantages pour l'ensemble de données examiné ici. Notre expérience de l'application de la méthode ELL nous porte à conclure que la modélisation de l'hétéroscédasticité n'est pas nécessaire.

Si nous revenons à la régression (c'est-à-dire les estimations produites pour $\boldsymbol{\beta}$ et l'erreur-type estimée pour les diverses méthodes), la méthode IEEP est celle qui intègre le mieux les poids de sondage provenant du calcul des composantes de la variance nécessaires pour produire les estimations sur petits domaines et leurs erreurs-types estimées. En ce qui concerne l'exécution, la méthode RGE serait généralement l'option la plus simple, parce qu'elle est disponible, par exemple, dans des progiciels tels que Stata, Sudaan ou WesVar. La méthode ELL combine les poids de sondage et la structure de covariance d'une façon non standard, en ce sens qu'elle utilise une estimation de $\mathbf{W}_b\mathbf{V}_b^{-1}$ dans (8) et (9) pour produire une matrice de covariance estimée asymétrique pour les estimations de $\boldsymbol{\beta}$ et pour estimer $\boldsymbol{\beta}$ proprement dit. Dans ce dernier cas, cette estimation serait acceptable si la matrice asymétrique était une inverse généralisée de la matrice de covariance correcte. Cependant, elle n'est manifestement pas acceptable comme matrice de covariance estimée, problème que la méthode ELL essaye de contourner (par exemple dans le logiciel POVMAP de la Banque mondiale) par calcul de la moyenne de chacune des paires pertinentes d'éléments hors diagonale afin de satisfaire la condition nécessaire qu'une matrice de covariance soit symétrique.

En général, dans la méthode ELL d'estimation de la pauvreté, seules peuvent être utilisées les variables dont la moyenne et l'écart-type concordent dans les moyennes sur données d'enquête ainsi que sur données de recensement. Cette contrainte tient au fait qu'après avoir ajusté le modèle de régression aux données d'enquête, à la deuxième phase, ce modèle est appliqué aux données de recensement en tant que prédicteur au niveau du ménage, autrement dit l'équation de régression (quelle que soit la façon dont elle a été estimée) est utilisée pour trouver les valeurs prévues du revenu et des dépenses par tête pour chaque ménage du recensement, produites au moyen de

$$\hat{Y}_{bh} = \mathbf{x}'_{bh}\hat{\boldsymbol{\beta}} + \hat{v}_b + \hat{e}_{bh} \tag{25}$$

en utilisant des valeurs imputées de v_b et e_{bh} (fondées, par exemple, sur un échantillonnage bootstrap de leurs estimations sur données d'enquête). Ici, les $\mathbf{x}_{bh}$ sont les variables auxiliaires provenant du recensement. Les indices de pauvreté sont habituellement fondés sur des fonctions non linéaires du logarithme du revenu ou du logarithme des dépenses, si bien que les prédictions issues de (25) sont transformées comme il convient avant de calculer la moyenne sur chaque petit domaine. Notons qu'en pratique, v_b peut être estimé pour les grappes échantillonnées, mais les codes d'échantillon et de recensement ne concordent habituellement pas, de sorte que celles-ci ne peuvent pas être identifiées dans le recensement et c'est donc le bootstrap (en sélectionnant parmi les barangays échantillonnés, c'est-à-dire les UPE) qui fournit les valeurs imputées pour tous les barangays ; un commentaire parallèle s'applique aux $\hat{e}_{bh}$ pour les ménages à l'intérieur des grappes. L'avantage général de l'utilisation de données de recensement de cette façon (comme le fait la méthode ELL) est que les variables prédictives peuvent être utilisées pour tous les ménages du recensement (qui sont nombreux) plutôt que simplement ceux de l'enquête, ce qui augmente la précision des estimations sur petits domaines (à condition que le modèle soit correct). Notons que les estimations données par (25) demeurent sans biais, même si v_b et e_{bh} ne sont pas inclus dans la prédiction proprement dites ; mais les estimations de la variance pour le petit domaine a doivent être calculées en se basant sur l'équation (25) afin que soit intégrée la variance supplémentaire nécessaire au niveau de la grappe et du ménage.

Dans l'estimation de la pauvreté, nous nous intéressons à des statistiques sommaires au niveau du domaine pour des fonctions non linéaires de $\hat{Y}_{bh}$, comme savoir si l'estimation est inférieure au seuil de pauvreté (prévalence de la pauvreté) et quel est l'écart de pauvreté, plutôt qu'à l'ajustement de la régression proprement dit. Il est intéressant, ici, d'examiner les effets de l'incertitude du modèle sur les estimations de la moyenne de domaine.

$$\bar{y}_a = \bar{\mathbf{x}}'_a\,\hat{\boldsymbol{\beta}} \tag{26}$$

où $\bar{\mathbf{x}}_a$ est la moyenne de population (c'est-à-dire du recensement) pour le domaine a des covariables incluant la constante 1, après avoir appliqué le modèle de régression aux données de recensement comme à la phase 2 de la

méthode ELL. En calculant de la même manière la moyenne de (7) pour obtenir la moyenne réelle $\bar{Y}_a$, en soustrayant le résultat de (26) et en appliquant l'opérateur de variance, nous obtenons l'équation de la variance de l'erreur de prédiction :

$$V(\bar{y}_a - \bar{Y}_a) = \bar{\mathbf{X}}_a \boldsymbol{\Phi}_w \bar{\mathbf{X}}'_a + \frac{1}{N_a^2} \sum_{b=1}^{m} N_b^2 \sigma_v^2 + \frac{1}{N_a} \sigma_e^2 \qquad (27)$$

où N_a est la taille de la population à un niveau particulier d'agrégation, N_b est la taille de la population dans chaque grappe, $\boldsymbol{\Phi}_w$ est la matrice de variance-covariance des estimations des coefficients de régression, et (σ_v^2, σ_e^2) sont les composantes de la variance au niveau de la grappe et du ménage, respectivement. Notons qu'estimer cette variance de l'erreur de prédiction requiert des estimations des composantes de la variance, mais que tout biais causé par l'incertitude dans ces dernières serait un effet de deuxième ordre (voir Prasad et Rao 1990).

Selon l'expression (27), l'importance de l'influence exercée par le modèle de régression fondé sur les données d'enquête et les autres composantes de la variance (au niveau de la grappe et du ménage) sur l'exactitude des estimations sur petits domaines finales peut être comparée pour toute méthode d'ajustement et (ou) tout niveau d'agrégation. En général, le modèle de régression (par la voie de l'estimation des paramètres de régression) ou l'effet de grappe est le facteur qui domine l'exactitude estimée de l'estimation sur petits domaines calculée. En utilisant le modèle au niveau national, dont les données sont présentées au tableau 1, et les variables auxiliaires de l'enquête, au lieu du recensement, pour estimer le premier terme de (27), nous constatons que la mesure dans laquelle l'effet du modèle de régression contribue à la variance des estimations sur petits domaines augmente appréciablement, à mesure que les données sur les ménages sont plus agrégées – environ 0,25 % au niveau municipal, 20 % au niveau provincial et 70 % au niveau régional. Autrement dit, la dominance de l'incertitude de l'estimation des paramètres du modèle de régression est d'autant plus importante que les données sont agrégées dans de plus grands domaines, indépendamment de la méthode d'ajustement de la régression. Ce résultat est conforme aux attentes, car même à des niveaux élevés d'agrégation, la contribution de l'effet du modèle à la variance globale dépend des valeurs moyennes des covariables, et non de la taille de la population. C'est pour cette raison qu'au niveau régional le plus agrégé, les méthodes d'estimation sur petits domaines offrent habituellement peu d'amélioration par rapport aux estimations directes. C'est également pour cela qu'il est important (comme nous l'avons fait dans le présent article) d'examiner en détail les procédures d'ajustement des régressions appliquées dans l'estimation sur petits domaines de la pauvreté dans le tiers monde.

L'effet de la variation au niveau de la grappe est différent : aux niveaux plus faibles d'agrégation (par exemple, municipalité), la variance calculée des estimations sur petits domaines est dominée par la composante au niveau de la grappe ou effet au niveau de la grappe, ce qui signifie que, pour l'estimation sur petits domaines (autres que les estimations régionales), la composante de la variance et non le modèle de régression a l'effet le plus important sur la valeur de l'erreur-type des estimations sur petits domaines. Par conséquent, l'exactitude des estimations des composantes de la variance, surtout au niveau de la grappe, est un élément essentiel à l'estimation exacte de l'erreur-type des estimations sur petits domaines au niveau d'agrégation auquel elles sont le plus utiles (par exemple au niveau municipal aux Philippines). De nouveau, la méthode utilisée pour l'ajustement de la phase 1 pour les composantes de la variance dont nous avons discuté dans le présent article joue un rôle crucial dans l'estimation sur petits domaines de la pauvreté.

Aux tableaux 4 à 6, nous présentons les résultats du test de Kruskal-Wallis (Siegel 1956) pour les diverses méthodes d'ajustement effectué sur les variances estimées au niveau municipal (tableau 4), provincial (tableau 5) et régional (tableau 6). Au tableau 4, des écarts significatifs existent entre les estimations de la variance produites par les diverses méthodes d'estimation sur petits domaines, comme l'indiquent les valeurs p de la statistique de Kruskal-Wallis (KW). La comparaison multiple des rangs moyens montre que les méthodes du pseudo-MPLSB et IEEP donnent des estimations de la variance au niveau de la grappe qui sont significativement plus élevées que celles obtenues pour les autres méthodes, mais qu'elles ne diffèrent pas significativement l'une de l'autre (quoique, pour la méthode IEEP, la valeur Z pour l'écart par rapport au rang moyen est en général assez bien plus élevée que toutes les autres).

La méthode ELL et la méthode RGE produisent des estimations des composantes de la variance significativement plus faibles et similaires. Ce résultat tient principalement au fait que nous avons utilisé la méthode ELL d'estimation des composantes de la variance pour produire des composantes de la variance pour la méthode RGE (parce que cette dernière ne comporte habituellement pas l'estimation de ces composantes), quoique les résidus que nous avons utilisés n'étaient pas identiques pour les deux méthodes d'ajustement de la régression. Comme prévu, au niveau municipal pour lequel les estimations sur petits domaines ont été utilisées en pratique, l'effet de grappe (plutôt que l'incertitude associée aux coefficients de régression) est généralement la partie dominante dans les estimations de la variance des estimations sur petits domaines. Puisque la variance au niveau de la grappe est la même pour les méthodes ELL et RGE, les estimations correspondantes de la variance au niveau du petit domaine ont également

tendance à être semblables. Explicitement, l'examen du tableau 4 montre que le classement des estimations de la variance concorde généralement avec le classement des effets de grappe.

Dans l'estimation de la pauvreté, les estimations aux niveaux élevés d'agrégation, telles que celles des tableaux 5 et 6, sont généralement produites en vue de les comparer aux estimations directes sur données d'enquête à ces niveaux d'agrégation plus élevés, même si elles ne sont pas particulièrement utiles pour la répartition de l'aide. Néanmoins, les résultats corroborent ceux présentés pour le niveau plus faible d'agrégation. Aux tableaux 5 et 6, les variances estimées des estimations de la pauvreté produites par les diverses méthodes ne diffèrent pas significativement les unes des autres aux niveaux provincial et régional, effet attribuable en partie au petit nombre de provinces et au nombre encore plus petit de régions. Les variances, et donc les erreurs-types, ne sont peut-être pas significativement différentes les unes des autres, mais il faut souligner que la méthode RGE a tendance à produire l'estimation la plus faible de l'erreur-type pour le modèle de régression et, à son tour, l'estimation la plus faible de la variance pour la pauvreté au niveau régional, même si cette méthode produit des erreurs-types plus élevées pour les coefficients de régression individuels (qui correspondent aux éléments diagonaux uniquement dans la matrice de covariance estimée de $\hat{\boldsymbol{\beta}}$). Comme prévu, à un niveau encore plus élevé d'agrégation, pour toutes les méthodes, l'effet relatif de la composante de régression est encore plus prononcé.

La conclusion générale est, que l'on ajuste un modèle à des données d'enquête uniquement ou que l'on utilise les estimations des paramètres de régression sur données d'enquête conjuguées à des données de recensement, qu'il est essentiel non seulement de trouver un modèle (c'est-à-dire, un ensemble de variables indépendantes) approprié fondé sur une taille d'échantillon adéquate, mais aussi d'obtenir de bonnes estimations des paramètres de régression et de leurs erreurs-types sous ce modèle, ainsi que de bonnes estimations et composantes de la variance à tous les niveaux pertinents d'agrégation. Habituellement, les niveaux pertinents d'agrégation sont déterminés par le plan de sondage, plutôt que simplement d'après le niveau auquel les estimations sur petits domaines sont souhaitées, quoique le nombre de niveaux ne doive pas nécessairement être limité à deux (par exemple niveau de la grappe et niveau du ménage).

Qu'elles soient utilisées pour l'estimation de la pauvreté ou dans un autre contexte, les données d'enquête introduisent aussi des problèmes ayant trait aux poids de sondage qui peuvent être importants non seulement pour l'estimation des paramètres de régression (et de leurs erreurs-types), mais aussi pour l'estimation des composantes de la variance. L'intégration des poids de sondage dans les modèles de régression avec données corrélées pose des problèmes, parce que c'est la corrélation de population, telle qu'elle est appliquée aux données d'enquête pondérées qui doit être modélisée correctement, de sorte que pondérer les matrices de corrélation par multiplication de matrices (comme cela est fait dans la méthode ELL) n'est pas une technique appropriée (voir l'annexe).

Pour les données sur les Philippines et pour la liste spécifiée de variables indépendantes, quelle que soit la méthode utilisée parmi les quatre étudiées, les estimations des paramètres sont très semblables, ce qui donne à penser qu'une question plus importante est la sous-estimation possible des erreurs-types des estimations des paramètres et des composantes de la variance, particulièrement au niveau de la grappe. La méthode ELL est la moins prudente en ce sens qu'elle donne les estimations les plus faibles des deux mesures de variance, et à cet égard (comme en ce qui concerne son utilisation de matrices de covariance estimées asymétriques) une certaine mise en garde pourrait être justifiée en ce qui concerne les aspects régression et composantes de la variance de cette méthode. La méthode RGE produit des estimations des erreurs-types des estimations sur petits domaines semblables à celles de la méthode ELL lorsque l'on utilise la même méthode pour les composantes de la variance, bien qu'elle donne des erreurs-types plus grandes (et utilise une bonne matrice de covariance) pour les paramètres de régression. Il en est ainsi parce qu'à un niveau moins agrégé, c'est-à-dire le niveau auquel la plupart des estimations sur petits domaines sont effectivement utilisées, les composantes de la variance dominent.

Dans les méthodes pseudo-MPLSB et IEEP, les poids de sondage sont intégrés correctement (étant donné un choix approprié de la pseudo-vraisemblance et, donc, des EEG) et les estimations de la composante de la variance au niveau de la grappe sont plus grandes (c'est-à-dire plus prudentes). Cela donne à penser que ces deux méthodes, particulièrement la méthode IEEP, comptent parmi les meilleures disponibles à l'heure actuelle, pas nécessairement pour l'estimation des équations de régression (pour laquelle l'existence de logiciels standard pourrait donner un avantage à la méthode RGE), mais pour estimer les composantes cruciales de la variance.

Tableau 4
Test de Kruskal-Wallis pour les variances estimées au niveau municipal (N = 1 243)

Méthodes	Effet de grappe			Effet bêta			Variance		
d'EPD	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z
ELL (sans hétérosc.)	0,002843	2 961,2(a)	-3,22	0,0002311	3 067,3(ab)	-0,89	0,00318	2 963,4(a)	-3,18
ELL (avec hétérosc.)	0,002843	2 961,2(a)	-3,22	0,0002128	2 802,0(c)	-6,72	0,00316	2 930,8(a)	-3,89
Pseudo-MPLSB	0,003094	3 229,4(b)	2,67	0,0002449	3 257,5(ad)	3,28	0,00346	3 241,3(b)	2,93
IEEP	0,003294	3 426,9(b)	7,01	0,0002529	3 364,5(d)	5,64	0,00366	3 441,3(b)	7,32
RGE (Stata)	0,002843	2 961,2(a)	-3,22	0,0002311	3 048,7(b)	-1,3	0,00317	2 963,1(a)	-3,18
Global		3,108			3,108			3,108	
Statistique de KW	H = 69,92	(P = 0,000)		H = 72,19	(P = 0,000)		H = 78,06	(P = 0,000)	

Tableau 5
Test de Kruskal-Wallis pour les variances estimées au niveau provincial (N = 83)

Méthodes	Effet de grappe			Effet bêta			Variance		
d'EPD	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z
ELL (sans hétérosc.)	0,0002518	200,3	-0,65	0,0001162	207,7	-0,03	0,00039	202,3	-0,48
ELL (avec hétérosc.)	0,0002518	200,3	-0,65	0,0001095	190,1	-1,52	0,00038	196,3	-0,99
Pseudo-MPLSB	0,000274	214,9	0,59	0,0001239	224,2	1,37	0,00042	217,1	0,78
IEEP	0,0002916	224,2	1,38	0,0001287	234,1	2,22	0,00045	227,8	1,68
RGE (Stata)	0,0002517	200,3	-0,65	0,00010	184	-2,04	0,00037	196,4	-0,98
Global		208			208			208	
Statistique de KW	H = 2,82	(P = 0,589)		H = 10,61	(P = 0,031)		H = 4,48	(P = 0,344)	

Tableau 6
Test de Kruskal-Wallis pour les variances estimées au niveau régional (N = 16)

Méthodes	Effet de grappe			Effet bêta			Variance		
d'EPD	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z	Médiane	Rang moyen	Z
ELL (sans hétérosc.)	0,000050	38,2	-0,45	0,000077	40,9	0,08	0,00013	39,3	-0,23
ELL (avec hétérosc.)	0,000050	38,2	-0,45	0,000073	35,1	-1,05	0,00012	37	-0,67
Pseudo-MPLSB	0,000055	42,6	0,4	0,000082	46,9	1,23	0,00014	44	0,67
IEEP	0,000058	45,3	0,93	0,000085	50,1	1,85	0,00015	46,6	1,17
RGE (Stata)	0,000050	38,2	-0,45	0,000070	29,6	-2,1	0,00013	35,6	-0,94
Global		40,5			40,5			40,5	
Statistique de KW	H = 1,30	(P = 0,861)		H = 8,36	(P = 0,079)		H = 2,58	(P = 0,630)	

Naturellement, ce genre de considérations (quoique essentielles) doivent s'appuyer sur une épuration appropriée des données, une bonne concordance des variables indépendantes possibles (en ce qui concerne la moyenne, la variance et la signification) de l'enquête et du recensement, dans le cas où des données de recensement sont également utilisées. Sont également nécessaires l'examen approprié, qui prend beaucoup de temps, d'une grande gamme de variables indépendantes possibles et la reconnaissance des limites qu'imposent les petites tailles d'échantillon à la subdivision des données d'enquête, puisque toutes les erreurs-types estimées pour les estimations des paramètres de régression ainsi que les estimations sur petits domaines (quelle que soit la méthode utilisée pour ajuster l'estimation des composantes de la variance) reposent sur la condition que le modèle de régression est correct.

8. Conclusion et recommandation

De bonnes statistiques sur la pauvreté sont nécessaires afin de pouvoir surveiller efficacement les interventions et l'aide offerte aux diverses localités appauvries. Les méthodes d'estimation sur petits domaines constituent l'une des méthodologies utilisées pour produire ce genre de statistiques. En ce sens, les questions soulevées dans le présent article au sujet de l'exactitude des estimations sur petits domaines ne sont pas simplement théoriques, mais sont essentielles à la réalisation des Objectifs du millénaire pour le développement et à la répartition de l'aide dans un secteur d'activités où les enjeux se chiffrent à plusieurs milliards de dollars.

Dans le présent article, nous avons examiné quatre méthodes d'estimation pour ajuster les modèles de régression en utilisant des données d'enquête et nous les avons reliées à l'estimation de la pauvreté sur petits domaines. Nous avons montré que, même si les écarts entre les estimations sont insuffisants pour rendre invalides les études à l'échelle nationale publiées, la méthode d'ajustement d'un modèle à des données d'enquête la plus fréquemment mise en œuvre, c'est-à-dire la méthode ELL avec hétéroscédasticité recommandée par la Banque mondiale, présente certaines limites, car, comme sa version homoscédastique,

elle ne repose pas sur un fondement théorique solide. Nous recommandons de remplacer la partie de la méthode ELL correspondant à l'ajustement d'un modèle aux données d'enquête. Les autres méthodes prises en considération (pseudo-MPLSB, IEEP et RGE) ont toutes un fondement mathématique théorique valide et les résultats produits peuvent être interprétés clairement une fois que les hypothèses ont été vérifiées. Appliquées à des données d'enquête complexes pondérées recueillies aux Philippines, les diverses méthodes indiquent que, pour l'estimation des composantes de la variance d'après des données d'enquête et, donc, pour l'estimation sur petits domaines à un fin niveau de détail, la méthode du pseudo-MPLSB et particulièrement la méthode IEEP donnent vraisemblablement de meilleurs résultats que les méthodes RGE ou ELL, quoique la méthode RGE est valable et facile à utiliser parce qu'elle est disponible dans les logiciels du commerce.

Nous avons également montré qu'au niveau où l'estimation sur petits domaines est effectivement utilisée pour répartir l'aide, l'estimation de la variance des estimations sur petits domaines a tendance à être dominée par la variance au niveau de la grappe plutôt que par l'exactitude des estimations des paramètres de régression. Donc, il est particulièrement important que la composante de la variance au niveau de la grappe (et, si l'ajustement du modèle est exécuté tel que recommandé, toute composante de la variance au niveau du petit domaine) soit estimée correctement. Il importe aussi que le modèle de régression utilisé pour produire les estimations sur petits domaines (y compris le choix de variables indépendantes pertinentes) soit approprié. Essentiellement, aux niveaux plus faibles d'agrégation, l'erreur-type des estimations sur petits domaines est dominée par les composantes de la variance, de sorte que l'estimation de ces dernières est cruciale, quel que soit le choix du niveau d'agrégation. Une méthode de régression sur données d'enquête valable, le bon choix des variables de régression et la détermination minutieuse de la taille d'échantillon (surtout si des modèles de régression distincts sont ajustés à des sous-ensembles de données d'enquête) demeurent aussi des éléments essentiels à une bonne estimation sur petits domaines de la pauvreté dans le tiers monde.

Remerciements

Les auteurs remercient les examinateurs et le rédacteur associé de leur lecture attentive du manuscrit et de leurs suggestions utiles.

Annexe

Dans la note en bas de page 8 du document de travail de la Banque mondiale rédigé par Elbers et coll. (2002) et, implicitement, dans l'article de Elbers et coll. (2003) publié dans *Econometrica*, la covariance du processus d'erreur est désignée par $\mathbf{\Omega}$ et il est déclaré que $\mathbf{W\Omega}^{-1} = \mathbf{P}^T\mathbf{P}$, où $\mathbf{W}$ est « une matrice de pondération de facteurs d'extension ». Dans la notation de la section 4 du présent article, $\mathbf{W}$ est diagonale par bloc, ou diagonale, avec blocs diagonaux $\mathbf{W}_b$, et $\mathbf{\Omega}$ est diagonale par bloc avec blocs diagonaux $\mathbf{V}_b$.

Cependant, soit $\mathbf{W}$ et $\mathbf{\Omega}$ (ou $\mathbf{\Omega}^{-1}$) sont non conformables (avec les facteurs de pondération dans $\mathbf{W}$ au niveau de la grappe et les observations, et donc $\mathbf{\Omega}^{-1}$, au niveau individuel), soit, si elles sont conformables, $\mathbf{W\Omega}^{-1}$ est généralement asymétrique (même si W est diagonale) à moins que $\mathbf{W}$ soit un multiple simple de la matrice identité, c'est-à-dire $\mathbf{W} = \sigma^2\mathbf{I}$.

D'où, $\mathbf{W\Omega}^{-1}$ n'est pas égale à $\mathbf{P}^T\mathbf{P}$ comme il l'a été soutenu, puisque $\mathbf{P}^T\mathbf{P}$ est symétrique en général et que $\mathbf{W\Omega}^{-1}$ ne l'est pas. Rendre $\mathbf{W\Omega}^{-1}$ symétrique en l'additionnant à sa transposée et en divisant le résultat par deux, comme cela est fait dans le logiciel PovMap de la Banque mondiale, n'est pas une solution techniquement adéquate de ce problème. (Notons que même dans le cas simple où $\mathbf{W}$ et $\mathbf{\Omega}^{-1}$ sont conformables, et que $\mathbf{W}$ est diagonale, mais que tous les éléments diagonaux ne sont pas égaux, $\mathbf{W\Omega}^{-1}$ n'est pas diagonale, parce que chaque élément de la ligne i de $\mathbf{\Omega}^{-1}$ est multiplié par w_i (où w_i est le i^{e} élément diagonal de $\mathbf{W}$), mais que, dans la i^{e} colonne, chaque élément *n'est pas* multiplié par un poids identique.)

En mettant de côté cette question de symétrie et en utilisant $\mathbf{P}^T\mathbf{P}$ à la place de $\mathbf{W\Omega}^{-1}$, Elbers et coll. semblent affirmer que comparer leur « estimateur MCG pondéré corrigé pour les données d'enquête » à l'estimateur MCG non corrigé implique qu'au lieu d'utiliser $\mathbf{\Omega}^{-1}$ comme mesure sous-jacente (c'est-à-dire l'inverse de la matrice de covariance pertinente), une version pondérée, à savoir $\mathbf{W\Omega}^{-1}\mathbf{W}^T$, devrait être utilisée. Cela ne crée aucun problème d'asymétrie en soi (à condition d'utiliser $\mathbf{P}^T\mathbf{P}$ à la place de $\mathbf{W\Omega}^{-1}$). Toutefois, même si $\mathbf{W}$ était diagonale et que $\mathbf{P}^T\mathbf{P}$ était utilisé, la matrice de pondération $\mathbf{W}$ ne peut même pas utiliser les poids diagonaux inégaux correspondant aux unités échantillonnées, disons w_i, parce que le ij^{e} élément de $\mathbf{\Omega}^{-1}$ (contrairement au ij^{e} élément de $\mathbf{\Omega}$) *ne* correspond *pas* aux i^{e} et j^{e} unités dans l'échantillon (ou dans la population), de sorte que l'on ne sait pas vraiment ce qu'est $\mathbf{W}$ ou comment $\mathbf{W}$ peut être définie de manière sensée comme une « matrice de pondération de facteurs d'extension ».

Cet argument reste applicable quand $\mathbf{V}_b$ est remplacée par son estimateur $\hat{\mathbf{V}}_b$ dans lequel sont utilisées des estimations à la place de σ_e^2 et σ_v^2.

Bibliographie

Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Chambers, R. (2006). What is poverty? Who asks? Who answers? *Poverty in Focus*, UNDP, 3 et 4 décembre 2006.

Elbers, C., Lanjouw, J. et Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355-364.

Elbers, C., Lanjouw, J. et Lanjouw, P. (2002). *Micro-level Estimation of Welfare.* Document de travail de recherche 2911, World Bank, Development Research Group, Washington, D.C.

Ghosh, M., et Rao J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Haslett, S., et Jones, G. (2004). *Local Estimation of Poverty and Malnutrition in Bangladesh*, Bangladesh Bureau of Statistics et United Nations World Food Programme.

Haslett, S., et Jones, G. (2005). *Local Estimation of Poverty in the Philippines*, Philippine National Statistics Co-ordination Board/World Bank Report. http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Local_Estimation_of_Poverty_Philippines.pdf.

Haslett, S., et Jones, G. (2005). Small area estimation using surveys and censuses: Some practical and statistical issues. *Statistics in Transition*, 7, 541-556.

Haslett, S., et Jones, G. (2006). *Small Area Estimation of Poverty, Caloric Intake and Malnutrition in Nepal.* Published: Nepal Central Bureau of Statistics/World Food Programme, United Nations/World Bank, septembre 2006, 184pp, ISBN 999337018-5.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

Horton, N.J., et Lipsitz, S.R. (1999). Review of software to fit Generalized Estimating Equation regression models. *The American Statistician*, 53, 160-169.

Liang, K.L., et Zeger, S. (1986). Longitudinal data analysis using Generalized Linear Models. *Biometrika*, 73, 13-22.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis.* Duxbury Press, Brooks/Cole Publishing Company.

Militino, A.F., Ugarte, M.D., Goicoa, T. et Gonzalez-Audicana, M. (2006). Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural*, *Biological*, *and Environmental Statistics*, 11, 450-461.

ONU site web. *http://www.un.org/fr/milleniumgoals/*.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society* B, 60, 23-40.

Pfeffermann, D., Moura, F.A. et Silva, P.L. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 949-959.

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, 163-171.

Rao, J.N.K. (1999). Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.

Rao, J.N.K. (2003). *Small Area Estimation*, Wiley Series in Survey Methodology. Wiley-Interscience, John Wiley & Sons, Inc.

NSCB (2000). *Profile of Censuses and Surveys*. National Statistical Coordination Board, Philippines.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill Series in Psychology. New York : McGraw-Hill.

Skinner, C.J., Holt, D. et Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Chichester : John Wiley & Sons.

You, Y., et Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *La Revue Canadienne de Statistique*, 30, 431-439.

You, Y., Rao, J.N.K. et Kovačević, M. (2003). Estimation des effets fixes et des composantes de la variance par un modèle à valeur aléatoire à l'origine en utilisant des données d'enquête. *Recueil* : *Symposium 2003*, *Défis reliés à la réalisation d'enquêtes pour la prochaine décennie*. Statistique Canada.

Zhao, Q. (2006). User manual for PovMap, The World Bank. http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao_ManualPovMap.pdf.