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

Article

Estimation de l'erreur quadratique moyenne robuste au biais pour les estimateurs sur petits domaines pseudo linéaires

par Ray Chambers, Hukum Chandra et Nikos Tzavidis

Décembre 2011

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 2011
Vol. 37, N° 2, pp. 163-181
Statistique Canada, N° 12-001-X au catalogue

Estimation de l'erreur quadratique moyenne robuste au biais pour les estimateurs sur petits domaines pseudo linéaires

Ray Chambers, Hukum Chandra et Nikos Tzavidis [1]

Résumé

Nous proposons une méthode d'estimation de l'erreur quadratique moyenne (EQM) pour les estimateurs des moyennes de domaine en population finie qui peuvent être exprimés sous une forme pseudo-linéaire, c'est-à-dire comme une somme pondérée des valeurs d'échantillon. En particulier, la méthode proposée peut être utilisée pour estimer l'EQM du meilleur prédicteur linéaire sans biais empirique, de l'estimateur direct fondé sur un modèle et du prédicteur du M-quantile. Elle représente une extension des idées de Royall et Cumberland (1978) et mène à des estimateurs de l'EQM qui sont plus simples à mettre en œuvre et éventuellement plus robustes au biais que ceux proposés dans la littérature sur les petits domaines. Cependant, il convient de souligner que les estimateurs de l'EQM définis en utilisant cette méthode peuvent également présenter une grande variabilité quand les tailles d'échantillon de domaine sont très petites. Nous illustrons les propriétés de la méthode à l'aide de simulations à grande échelle sous un modèle et sous un plan de sondage, dans ce dernier cas en nous fondant sur deux ensembles de données d'enquête réels contenant des données sur des petits domaines.

Mots clés : Meilleur prédicteur linéaire sans biais ; modèle du M-quantile ; estimation directe fondée sur un modèle ; modèle à effets aléatoires ; estimation sur petits domaines.

1. Introduction

Les modèles linéaires, et les prédicteurs linéaires fondés sur ces modèles, sont d'usage très répandu pour l'inférence d'après des données d'enquête. Cependant, ces modèles risquent d'être spécifiés incorrectement, particulièrement en ce qui concerne les moments de deuxième ordre et d'ordre plus élevé. Des méthodes d'estimation de l'erreur quadratique moyenne (EQM) des prédicteurs linéaires de quantités de population finie robustes au biais, c'est-à-dire des méthodes qui demeurent approximativement sans biais lorsque les hypothèses au sujet des moments de deuxième ordre et d'ordre plus élevé sont violées, ont été élaborées. Valliant, Dorfman et Royall (2000, chapitre 5) discutent de l'estimation de l'EQM robuste au biais pour de tels prédicteurs quand on suppose qu'une population suit un modèle linéaire.

Dans le présent article, nous abordons un problème subsidiaire, qui est celui de l'estimation de l'EQM robuste au biais des estimateurs des moyennes de domaine de population finie qui peuvent être exprimés sous une forme pseudo-linéaire, c'est-à-dire comme des sommes pondérées, mais dans lesquels les poids peuvent dépendre des valeurs d'échantillon de la variable d'intérêt. Une application importante, et qui motive notre approche, est l'inférence sur petits domaines. Par conséquent, dans la suite de l'exposé, le petit domaine sera le domaine d'intérêt. Notre approche, qui représente une extension des idées de Royall et Cumberland (1978), semble produire des estimateurs de l'EQM plus simples à mettre en œuvre que ceux qui ont été proposés dans la littérature sur les petits domaines.

Le plan de l'article est le suivant. À la section 2, nous discutons de l'estimation de l'EQM sous un modèle linéaire propre au domaine. Autrement dit, nous nous concentrons sur l'estimation de l'EQM conditionnelle. Puis, nous montrons comment notre approche pourrait être utilisée pour estimer l'EQM de trois prédicteurs linéaires sur petits domaines différents quand ils sont exprimés sous une forme pseudo-linéaire, à savoir a) le meilleur prédicteur linéaire sans biais empirique ou EBLUP (pour *empirical best linear unbiased predictor*) (Henderson 1953), b) l'estimateur direct fondé sur un modèle (EDFM) de Chandra et Chambers (2009) et c) le prédicteur du M-quantile (Chambers et Tzavidis 2006). À la section 3, nous présentons les résultats d'une série d'études en simulation qui illustrent les propriétés sous le modèle et sous le plan de notre approche d'estimation de l'EQM. Enfin, à la section 4, nous résumons nos principaux résultats. Tout au long de l'exposé, nous utilisons i ou h comme indice des D petits domaines d'intérêt, et j ou k comme indice des unités de population distinctes dans ces domaines.

2. Estimation de l'EQM robuste au biais pour les estimateurs pseudo-linéaires

2.1 Estimation de l'EQM sous un modèle linéaire propre au domaine

Nous considérons la situation où nous avons une population finie de taille N dont est tiré un échantillon de taille n. Nous supposons que cette population est constituée de D

1. Ray Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australie. Courriel : ray@uow.edu.au ; Hukum Chandra, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, Inde. Courriel : hchandra@iasri.res.in ; Nikos Tzavidis, Social Statistics and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, R.-U. Courriel : n.tzavidis@soton.ac.uk.

domaines non chevauchants, qui contiennent chacun des unités échantillonnées, les tailles d'échantillon réalisées dans chacun des domaines échantillonnés étant faibles. Nous supposons aussi qu'il existe un nombre connu N_i d'unités de la population dans le domaine *i* et qu'un nombre n_i de ces unités sont échantillonnées. Le nombre total d'unités dans la population est $N = \sum_{i=1}^{D} N_i$, et la taille d'échantillon totale correspondante est $n = \sum_{i=1}^{D} n_i$. Dans la suite, nous utilisons *s* pour désigner l'ensemble d'unités dans l'échantillon, s_i étant le sous-ensemble tiré du domaine *i*, et nous utilisons des expressions telles que $j \in i$ et $j \in s$ pour faire référence aux unités qui constituent le domaine *i* et l'échantillon *s*, respectivement.

Des modèles linéaires sont souvent utilisés pour fonder les estimateurs des moyennes de population. Cependant, quand sont requises des estimations des moyennes de domaines correspondantes, il n'est habituellement pas raisonnable de supposer qu'un modèle linéaire applicable à l'ensemble de la population s'applique aussi dans chaque domaine. Par conséquent, nous adoptons une approche conditionnelle et considérons l'estimation de l'EQM des estimateurs des moyennes de domaine quand les modèles linéaires à appliquer ne sont pas les mêmes dans les différents domaines. En particulier, nous nous penchons sur les estimateurs qui peuvent être exprimés comme une somme pondérée des valeurs d'échantillon, en les qualifiant de « linéaire » dans la suite pour indiquer qu'ils ont une structure linéaire.

Pour commencer, soit y_j la valeur de *Y* pour l'unité *j* de la population et supposons que cette unité est dans le domaine *i*. Nous émettons aussi l'hypothèse d'un modèle linéaire propre au domaine pour y_j de la forme

$$y_j = \boldsymbol{x}_j^T \boldsymbol{\beta}_i + e_j. \qquad (1)$$

Ici, $\boldsymbol{x}_j$ est un vecteur de dimension $p \times 1$ de variables auxiliaires au niveau de l'unité pour l'unité *j*, $\boldsymbol{\beta}_i$ est un vecteur de dimension $p \times 1$ de coefficients de régression propres au domaine et e_j est un effet aléatoire au niveau de l'unité de moyenne nulle et de variance σ_j^2 pour lequel il n'existe pas de corrélation entre les diverses unités de population. Nous ne faisons aucune hypothèse au sujet de σ_j^2 à ce stade. Notons que, tout au long de l'exposé, nous supposons que la méthode d'échantillonnage utilisée est non informative pour les valeurs de population de *Y* sachant les valeurs correspondantes des variables auxiliaires et les appartenances aux domaines des unités de population. Par conséquent, (1) s'applique à la fois au niveau de l'échantillon et de la population.

Soit $\boldsymbol{y}_s$ le vecteur colonne de valeurs d'échantillon de y_j et soit $\boldsymbol{w}_{is} = \{w_{ij} ; j \in s\}$ le vecteur colonne de poids fixes tels que $\hat{m}_i = \boldsymbol{w}_{is}^T \boldsymbol{y}_s = \sum_{j \in s} w_{ij} y_j$ est un estimateur linéaire de $m_i = N_i^{-1} \sum_{j \in i} y_j$. Par « fixes » nous entendons ici que ces poids ne dépendent pas des valeurs d'échantillon de *Y*. En outre, nous supposons que $w_{ij} = O(n_i^{-1})$ pour $j \in s_i$, $w_{ij} = o(n_i^{-1})$ pour $j \notin s_i$, et $\sum_{j \in s} w_{ij} = 1$. Ici, s_i désigne les n_i unités d'échantillon provenant du domaine *i*. Le biais de $\hat{m}_i$ sous (1) est alors

$$E(\hat{m}_i - m_i) = \left(\sum_{h=1}^{D} \sum_{j \in s_h} w_{ij} \boldsymbol{x}_j^T \boldsymbol{\beta}_h \right) - \bar{\boldsymbol{x}}_i^T \boldsymbol{\beta}_i, \qquad (2)$$

où $\bar{\boldsymbol{x}}_i$ désigne le vecteur de valeurs moyennes des variables auxiliaires dans le domaine *i*. De même, la variance de prédiction de $\hat{m}_i$ sous (1) est

$$\mathrm{Var}(\hat{m}_i - m_i) = N_i^{-2} \left\{ \sum_{h=1}^{D} \sum_{j \in s_h} a_{ij}^2 \sigma_j^2 + \sum_{j \in r_i} \sigma_j^2 \right\}, \qquad (3)$$

où r_i désigne les unités non échantillonnées dans le domaine *i* et $a_{ij} = N_i w_{ij} - I(j \in i)$. Nous utilisons $I(A)$ pour désigner la fonction indicatrice de l'évènement *A*, de sorte que $I(j \in i)$ prend la valeur 1 si l'unité de population *j* provient du domaine *i* et la valeur zéro autrement. Notons que, puisque a_{ij} est d'ordre $O(N_i n_i^{-1})$ pour $j \in s_i$, le premier terme entre parenthèses dans (3) est le terme principal de cette variance de prédiction si N_i est grand comparativement à n_i.

Soit $j \in h$. Nous considérons le cas particulier important où $\mu_j = E(y_j \mid \boldsymbol{x}_j) = \boldsymbol{x}_j^T \boldsymbol{\beta}_h$ est estimé par $\hat{\mu}_j = \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}_h = \sum_{k \in s} \phi_{kj} y_k$ avec les ϕ_{kj} correspondant aux poids appropriés. Alors

$$y_j - \hat{\mu}_j = (1 - \phi_{jj}) y_j - \sum_{k \in s(-j)} \phi_{kj} y_k$$

et donc

$$\mathrm{Var}(y_j - \hat{\mu}_j) = \sigma_j^2 \left\{ (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2 (\sigma_k^2 / \sigma_j^2) \right\} \qquad (4)$$

sous (1). Ici, $s(-j)$ désigne l'échantillon *s* dont l'unité *j* est exclue. Si, en outre, $\hat{\mu}_j$ est sans biais pour μ_j sous (1), c'est-à-dire

$$E(y_j - \hat{\mu}_j) = 0, \qquad (5)$$

nous pouvons adopter l'approche de Royall et Cumberland (1978) et estimer (3) par

$$\hat{V}(\hat{m}_i) = N_i^{-2} \left\{ \sum_{h=1}^{D} \sum_{j \in s_h} a_{ij}^2 \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2 + \sum_{j \in r_i} \hat{\sigma}_j^2 \right\}, \qquad (6)$$

où $\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \hat{\gamma}_{kj} \phi_{kj}^2$ et $\hat{\gamma}_{kj} = \hat{\sigma}_k^2 / \hat{\sigma}_j^2$. Habituellement, les estimations $\hat{\sigma}_j^2$ des variances résiduelles dans (6) sont calculées sous un perfectionnement de « modèle de travail » de (1). Dans la situation qui nous concerne le plus, où les tailles d'échantillon dans les différents domaines sont trop faibles pour estimer fiablement la variabilité propre au domaine, une hypothèse de groupement peut être faite, c'est-à-dire $\sigma_j^2 = \sigma^2$, auquel cas nous posons

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 = n^{-1}\sum_{j\in s}\left\{(1-\phi_{jj})^2 + \sum_{k\in s(-j)}\phi_{kj}^2\right\}^{-1}(y_j - \hat{\mu}_j)^2.$$

Dans ce cas, (6) devient

$$\hat{V}(\hat{m}_i) = N_i^{-2}\sum_{j\in s}\left\{a_{ij}^2 + (N_i - n_i)n^{-1}\right\}\hat{\lambda}_j^{-1}(y_j - \hat{\mu}_j)^2, \quad (7)$$

où, maintenant, $\hat{\lambda}_j = (1-\phi_{jj})^2 + \sum_{k\in s(-j)}\phi_{kj}^2$. Comme toute hypothèse concernant σ_j^2 dans l'extension de (1) à un modèle de travail n'affecte que les termes de deuxième ordre dans (3), l'estimateur (7) est robuste au biais, c'est-à-dire qu'il demeure approximativement sans biais sous l'erreur de spécification des moments de deuxième ordre de ce modèle de travail.

Un estimateur correspondant de l'EQM de $\hat{m}_i$ sous (1) s'ensuit directement. Il s'agit de

$$\hat{M}(\hat{m}_i) = \hat{V}(\hat{m}_i) + \hat{B}^2(\hat{m}_i), \quad (8)$$

où

$$\hat{B}(\hat{m}_i) = \sum_{h=1}^{D}\sum_{j\in s_h} w_{ij}\hat{\mu}_j - N_i^{-1}\sum_{j\in i}\hat{\mu}_j \quad (9)$$

est l'estimateur sans biais évident de (2).

L'utilisation du carré de l'estimateur sans biais (9) du biais de $\hat{m}_i$ dans l'estimateur conditionnel de l'EQM (8) peut être critiquée, parce que ce terme n'est pas lui-même sans biais pour le terme du carré du biais dans l'EQM. Cela peut être corrigé en remplaçant (9) par

$$\hat{M}(\hat{m}_i) = \hat{V}(\hat{m}_i) + \hat{B}^2(\hat{m}_i) - \hat{V}\{\hat{B}(\hat{m}_i)\}, \quad (10)$$

où $\hat{V}\{\hat{B}(\hat{m}_i)\}$ est un estimateur approprié de la variance de (9). Cependant, nous ne recommandons pas d'utiliser (10). Pour voir pourquoi, soit $\bar{\boldsymbol{\beta}} = D^{-1}\sum_{h=1}^{D}\hat{\boldsymbol{\beta}}_h$ et posons que $\boldsymbol{d}_h = \hat{\boldsymbol{\beta}}_h - \bar{\boldsymbol{\beta}}$, où $\hat{\boldsymbol{\beta}}_h$ est l'estimateur de $\boldsymbol{\beta}_h$ impliqué par les poids ϕ_{kj}. En outre, posons que $w_{hi} = \sum_{j\in s_h} w_{ij}$ et $\bar{\boldsymbol{x}}_{whi} = w_{hi}^{-1}\sum_{j\in s_h} w_{ij}\boldsymbol{x}_j$, de sorte que $\bar{\boldsymbol{x}}_{wi} = \sum_{h=1}^{D}\sum_{j\in s_h} w_{ij}\boldsymbol{x}_j = \sum_{h=1}^{D} w_{hi}\bar{\boldsymbol{x}}_{whi}$ est l'estimateur de $\bar{\boldsymbol{x}}_i$ fondé sur les poids w_{ij}. Enfin, soit $\delta_{hi} = \bar{\boldsymbol{x}}_h^T\boldsymbol{d}_h - \bar{\boldsymbol{x}}_i^T\boldsymbol{d}_i$, et posons que $\delta_i = \sum_{h=1}^{D} w_{hi}\delta_{hi}$. Alors, (9) peut s'écrire

$$\begin{aligned}
\hat{B}(\hat{m}_i) &= (\bar{\boldsymbol{x}}_{wi} - \bar{\boldsymbol{x}}_i)^T\bar{\boldsymbol{\beta}} \\
&\quad + \left(\sum_{h=1}^{D} w_{hi}\bar{\boldsymbol{x}}_{whi}^T\boldsymbol{d}_h - \bar{\boldsymbol{x}}_i^T\boldsymbol{d}_i\right) \\
&= (\bar{\boldsymbol{x}}_{wi} - \bar{\boldsymbol{x}}_i)^T\bar{\boldsymbol{\beta}} \\
&\quad + \left(\sum_{h=1}^{D} w_{hi}(\bar{\boldsymbol{x}}_{whi} - \bar{\boldsymbol{x}}_h)^T\boldsymbol{d}_h + \sum_{h=1}^{D} w_{hi}\bar{\boldsymbol{x}}_h^T\boldsymbol{d}_h - \bar{\boldsymbol{x}}_i^T\boldsymbol{d}_i\right) \\
&= (\bar{\boldsymbol{x}}_{wi} - \bar{\boldsymbol{x}}_i)^T\bar{\boldsymbol{\beta}} \\
&\quad + \sum_{h=1}^{D} w_{hi}(\bar{\boldsymbol{x}}_{whi} - \bar{\boldsymbol{x}}_h)^T\boldsymbol{d}_h + \delta_i. \quad (11)
\end{aligned}$$

Habituellement, D sera grand et le terme principal dans la variance de (9) sera la variance de δ_i dans (11). Si ce terme principal est grand, alors $\hat{V}\{\hat{B}(\hat{m}_i)\}$ sera également grand, et (10) pourrait prendre des valeurs négatives. Par conséquent, nous recommandons d'utiliser (8) plutôt que (10). Une conséquence directe est que (8) est alors un estimateur prudent de l'EQM de $\hat{m}_i$ sous (1). Cela pourrait être acceptable, à condition que la variance de δ_i soit faible. Toutefois, pour de très petites valeurs de n_i, cette variance peut être grande, de sorte que (8) surestimerait considérablement l'EQM réelle de $\hat{m}_i$. Nous recommandons donc une évaluation empirique préliminaire de la taille de la variance de δ_i par rapport à la valeur de (7) dans cette situation. Si cette évaluation indique que la variance de δ_i domine (7), alors l'expression (8) ne devrait pas être utilisée.

2.2 Estimation de l'EQM des estimateurs sur petits domaines pseudo-linéaires

L'approche de l'estimation conditionnelle de l'EQM décrite à la sous-section précédente s'appuie sur l'hypothèse selon laquelle les poids définissant l'estimateur linéaire $\hat{m}_i$ ne dépendent pas des valeurs d'échantillon de *Y*. Cependant, la plupart des estimateurs sur petits domaines ne satisfont pas cette condition, en ce sens qu'ils sont pseudo-linéaires en structure, avec des pondérations qui dépendent de ces valeurs d'échantillon. Par exemple, le meilleur prédicteur linéaire sans biais (BLUP, pour *Best Linear Unbiased Predictor*) de m_i sous la variante du modèle mixte linéaire de (1) où les paramètres de régression propres au domaine $\boldsymbol{\beta}_i$ sont des réalisations indépendantes et identiquement distribuées d'une variable aléatoire dont la valeur prévue $\boldsymbol{\beta}$ et la matrice de covariance $\boldsymbol{\Gamma}$ peuvent s'écrire comme une somme pondérée des valeurs d'échantillon de *Y* où les poids dépendent de $\boldsymbol{\Gamma}$ (voir Royall 1976). Par conséquent, la version empirique de ce prédicteur, c'est-à-dire l'EBLUP d'usage très répandu, est calculée en introduisant une estimation sur échantillon efficace de $\boldsymbol{\Gamma}$ (par exemple l'estimation du maximum de vraisemblance restreint ou MVR) par substitution dans les poids BLUP. Si l'hypothèse du modèle mixte linéaire est vraie, cet estimateur sur échantillon de $\boldsymbol{\Gamma}$ converge vers la valeur vraie et, par conséquent, les poids EBLUP convergent vers les poids BLUP. Autrement dit, pour de grandes valeurs de la taille globale d'échantillon *n*, nous pouvons traiter les poids EBLUP comme étant fixes et utiliser l'estimateur de l'EQM (8) pour l'EBLUP. Naturellement, les poids EBLUP ne sont pas vraiment fixes et (8) est par conséquent l'approximation de l'EQM vraie de l'EBLUP qui ne tient pas compte de la contribution de la variabilité de l'estimation de $\boldsymbol{\Gamma}$ à cette EQM. Cependant, cette sous-estimation éventuelle doit être mise en balance avec la robustesse au biais de (8) sous

l'erreur de spécification des moments de deuxième ordre de *Y*.

Un avantage important de l'expression (8) est qu'elle peut être utilisée avec une gamme d'estimateurs sur petits domaines pouvant être exprimés sous forme pseudo-linéaire. En particulier, de nombreux estimateurs sur petits domaines élaborés sous des modèles qui sont des variantes de (1) peuvent s'écrire sous cette forme, c'est-à-dire comme des sommes pondérées des valeurs d'échantillon de *Y*. Pour illustrer ce point, nous nous concentrons maintenant sur trois de ces estimateurs : l'EBLUP (Rao 2003, chapitre 6), l'estimateur direct fondé sur un modèle (EDFM) de Chandra et Chambers (2009) et le prédicteur du M-quantile de Chambers et Tzavidis (2006). Chacun de ces estimateurs peut s'écrire sous une forme pseudo-linéaire avec des poids qui satisfont $w_{ij} = O(n_i^{-1})$ pour $j \in s_i$ et $w_{ij} = o(n_i^{-1})$ pour $j \notin s_i$, et donc (8) peut être utilisé.

2.2.1 Estimation de l'EQM pour l'EBLUP

Nous considérons d'abord l'EBLUP bien connu pour m_i fondé sur l'extension du modèle (1) à un modèle mixte linéaire au niveau de l'unité de la forme

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i \qquad (12)$$

où $\boldsymbol{y}_i$ est le vecteur de dimension N_i des valeurs de population de y_j dans le domaine *i*, $\boldsymbol{X}_i$ est la matrice de dimensions $N_i \times p$ correspondante des valeurs des variables auxiliaires $\boldsymbol{x}_j$, $\boldsymbol{Z}_i$ est la composante de dimensions $N_i \times q$ des $\boldsymbol{X}_i$ correspondant aux *q* composantes aléatoires de $\boldsymbol{\beta}$, $\boldsymbol{u}_i$ est le vecteur de dimension *q* connexe d'effets aléatoires propres au domaine et $\boldsymbol{e}_i$ est le vecteur de dimension N_i des effets aléatoires individuels. On suppose habituellement que les effets de domaine et les effets individuels sont mutuellement indépendants, les effets de domaine étant indépendants et identiquement distribués suivant la loi $N(0, \boldsymbol{\Omega})$ et les effets individuels étant indépendants et identiquement distribués suivant la loi $N(0, \sigma^2)$. Voir Rao (2003, chapitre 6) pour le développement de la théorie qui sous-tend ce prédicteur. Nous notons que l'EBLUP peut s'écrire sous la forme pseudo-linéaire,

$$\hat{m}_i^{\text{EBLUP}} = \sum\nolimits_{j\in s} w_{ij}^{\text{EBLUP}} y_j = (\boldsymbol{w}_{is}^{\text{EBLUP}})^T \boldsymbol{y}_s \qquad (13)$$

où

$$\begin{aligned}\boldsymbol{w}_{is}^{\text{EBLUP}} &= (w_{ij}^{\text{EBLUP}}) \\ &= N_i^{-1}[\boldsymbol{\Delta}_{is} + \{\hat{\boldsymbol{H}}_s^T \boldsymbol{X}_r^T + (\boldsymbol{I}_n - \hat{\boldsymbol{H}}_s^T \boldsymbol{X}_s^T)\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\hat{\boldsymbol{\Sigma}}_{sr}\}\boldsymbol{\Delta}_{ir}].\end{aligned}$$

Ici, $\boldsymbol{\Delta}_{ir}$ est un vecteur de taille $N - n$ qui « isole » les unités non échantillonnées dans le domaine *i*, $\boldsymbol{X}_s$ et $\boldsymbol{X}_r$ sont les matrices d'ordre $n \times p$ et $(N - n) \times p$ respectivement des valeurs dans l'échantillon et hors échantillon des variables auxiliaires, $\boldsymbol{I}_n$ est la matrice identité d'ordre *n*, $\hat{\boldsymbol{H}}_s = (\boldsymbol{X}_s^T\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^T\hat{\boldsymbol{\Sigma}}_{ss}^{-1}$, $\hat{\boldsymbol{\Sigma}}_{ss} = \hat{\sigma}^2\boldsymbol{I}_n + \text{diag}\{\boldsymbol{Z}_{is}\hat{\boldsymbol{\Omega}}\boldsymbol{Z}_{is}^T ; i = 1, ..., D\}$ et $\hat{\boldsymbol{\Sigma}}_{sr} = \text{diag}\{\boldsymbol{Z}_{is}\hat{\boldsymbol{\Omega}}\boldsymbol{Z}_{ir}^T ; i = 1, ..., D\}$. Ici, $\boldsymbol{Z}_{is}(\boldsymbol{Z}_{ir})$ est la composante d'échantillon (hors échantillon) de $\boldsymbol{Z}_i$, et $\hat{\sigma}^2$ et $\hat{\boldsymbol{\Omega}}$ sont des estimations appropriées (par exemple MV ou MVR) des composantes de la variance de (12).

Étant donné ces conditions, l'estimation de l'EQM conditionnelle de l'EBLUP peut être effectuée en se servant de (8) avec les poids définis conformément à (13). À son tour, cela requiert que nous ayons accès à des estimateurs sans biais $\hat{\mu}_j$ des valeurs prévues individuelles propres au domaine μ_j. Cependant, ces estimateurs peuvent être instables quand les tailles d'échantillon de domaine sont faibles. Par conséquent, il est tentant de remplacer $\hat{\mu}_j$ par l'EBLUP pour y_j, c'est-à-dire $\hat{y}_j^{\text{EBLUP}} = \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}^{\text{EBLUE}} + \boldsymbol{z}_j^T\hat{\boldsymbol{u}}_i^{\text{EBLUP}}$, où $\hat{\boldsymbol{\beta}}^{\text{EBLUE}}$ désigne le meilleur estimateur linéaire sans biais empirique (EBLUE pour *Empirical Best Linear Unbiased Estimator*) de β dans le modèle linéaire mixte (12) et $\hat{\boldsymbol{u}}_i^{\text{EBLUP}}$ désigne l'effet de domaine prédit pour le domaine *i* qui contient l'observation *j*. Malheureusement, en raison de l'effet de rétrécissement bien connu associé aux EBLUP, cette approche n'est pas recommandée. Pour l'illustrer, nous notons que, dans (8), $\hat{V}(\hat{m}_i)$ utilise $(y_j - \hat{\mu}_j)^2$ comme estimateur de $E(y_j - \mu_j)^2$. Le biais dans cet estimateur est par conséquent

$$\begin{aligned}&E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2 \\ &\quad = -2E(y_j - \mu_j)(\hat{\mu}_j - \mu_j) + E(\hat{\mu}_j - \mu_j)^2 \\ &\quad = -E\{(\hat{\mu}_j - \mu_j)(2y_j - \mu_j - \hat{\mu}_j)\}\end{aligned}$$

de sorte que nous nous attendons à ce que $\hat{V}(\hat{m}_i)$ présente un biais négatif si $E\{(\hat{\mu}_j - \mu_j)(2y_j - \mu_j - \hat{\mu}_j)\}$ est positif et inversement. Maintenant, soit l'unité d'échantillon *j* provenant du domaine *i* et considérons le cas particulier d'un modèle à ordonnées à l'origine aléatoires pour y_j, c'est-à-dire $y_j = \boldsymbol{x}_j^T\boldsymbol{\beta} + u_i + e_j$, où u_i est l'effet aléatoire pour le domaine *i* et e_j est un effet individuel aléatoire non corrélé à u_i. Ici, $\mu_j = \boldsymbol{x}_j^T\boldsymbol{\beta} + u_i$. Supposons que nous ayons une grande taille globale d'échantillon, ce qui nous permet de remplacer $\hat{\boldsymbol{\beta}}^{\text{EBLUE}}$ par $\boldsymbol{\beta}$. L'EBLUP $\hat{\mu}_j = \hat{y}_j^{\text{EBLUP}}$ peut alors être approximé par $\tilde{\mu}_j = \boldsymbol{x}_j^T\boldsymbol{\beta} + \gamma_i u_i$, où γ_i est un facteur de « rétrécissement ». Il s'ensuit que

$$(\tilde{\mu}_j - \mu_j)(2y_j - \mu_j - \tilde{\mu}_j) = 2u_i(\gamma_i - 1)e_i - u_i^2(\gamma_i - 1)^2$$

de sorte que $E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2 \approx (\gamma_i - 1)^2\sigma_u^2$. Autrement dit, nous nous attendons à ce que $\hat{V}(\hat{m}_i)$ présente un biais positif si nous utilisons l'EBLUP rétréci $\hat{y}_j^{\text{EBLUP}}$ pour définir $\hat{\mu}_j$. Nous notons aussi que ce biais disparaît (approximativement) si nous « dérétrécissons » la composante résiduelle de cet EBLUP. Par exemple, dans le cas du modèle à ordonnées à l'origine aléatoires très répandu, nous utilisons

$$\hat{\mu}_j = \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}^{\mathrm{EBLUE}} + (\bar{y}_{is} - \bar{\boldsymbol{x}}_{is}^T\hat{\boldsymbol{\beta}}^{\mathrm{EBLUE}}) = \bar{y}_{is} + (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_{is})^T\hat{\boldsymbol{\beta}}^{\mathrm{EBLUE}}$$

où $\bar{y}_{is}$ et $\bar{\boldsymbol{x}}_{is}$ désignent les moyennes d'échantillon de Y et X respectivement, dans le domaine i. Sachant que (12) est le modèle de travail, une expression générale pour un tel estimateur « dérétréci » est

$$\hat{\mu}_j = \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}^{\mathrm{EBLUE}} + \boldsymbol{z}_j^T\tilde{\boldsymbol{u}}_i \qquad (14)$$

où $\tilde{\boldsymbol{u}}_i = (\boldsymbol{Z}_{is}^T\boldsymbol{Z}_{is})^{-1}\boldsymbol{Z}_{is}^T(y_{is} - \boldsymbol{X}_{is}\hat{\boldsymbol{\beta}}^{\mathrm{EBLUE}})$ est le prédicteur dérétréci de l'effet aléatoire pour le domaine i. Il n'est pas difficile de voir qu'alors, $\hat{\mu}_j = \sum_{k\in s}\phi_{kj}y_k$, où $\phi_{kj} = c_{ijsk} + b_{ijsk}I(k\in i)$, avec

$$\begin{aligned}\boldsymbol{c}_{ijs} &= (c_{ijsk}\ ;\ k\in s)\\ &= \hat{\boldsymbol{\Sigma}}_{ss}^{-1}\boldsymbol{X}_s(\boldsymbol{X}_s^T\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\boldsymbol{X}_s)^{-1}\{\boldsymbol{x}_j - \boldsymbol{X}_{is}^T\boldsymbol{Z}_{is}(\boldsymbol{Z}_{is}^T\boldsymbol{Z}_{is})^{-1}\boldsymbol{z}_j\}\end{aligned}$$

et $\boldsymbol{b}_{ijs} = (b_{ijsk}\ ;\ k\in s_i) = \boldsymbol{Z}_{is}(\boldsymbol{Z}_{is}^T\boldsymbol{Z}_{is})^{-1}\boldsymbol{z}_j$. Notons que ces ϕ_{kj} sont également utilisés pour calculer la valeur de $\hat{\lambda}_j$ définie juste après (7).

Enfin, nous observons que, quand (14) est utilisé dans (8), le biais estimé (9) devient

$$\hat{B}(\hat{m}_i) = \sum_{h=1}^{D}\left(\sum\nolimits_{j\in s_h} w_{ij}^{\mathrm{EBLUP}}\boldsymbol{z}_j\right)^T\tilde{\boldsymbol{u}}_h - \bar{\boldsymbol{z}}_i^T\tilde{\boldsymbol{u}}_i$$

puisque les poids EBLUP (13) sont « calés localement » sur X, c'est-à-dire $\sum_{j\in s} w_{ij}^{\mathrm{EBLUP}}\boldsymbol{x}_j = \bar{\boldsymbol{x}}_i$. Il s'ensuit que, dans ce cas, la variable δ_i définie juste avant (11) prend la forme

$$\delta_i = \sum\nolimits_{h=1}^{D} w_{hi}^{\mathrm{EBLUP}}\bar{\boldsymbol{z}}_h^T\tilde{\boldsymbol{u}}_h - \bar{\boldsymbol{z}}_i^T\tilde{\boldsymbol{u}}_i$$

où $w_{hi}^{\mathrm{EBLUP}} = \sum_{j\in s_h} w_{ij}^{\mathrm{EBLUP}}$. Pour une taille d'échantillon suffisamment grande, δ_i peut être approximé par

$$\begin{aligned}\delta_i &\approx \sum\nolimits_{h=1}^{D} w_{hi}^{\mathrm{BLUP}}\bar{\boldsymbol{z}}_h^T(\boldsymbol{Z}_{hs}^T\boldsymbol{Z}_{hs})^{-1}\boldsymbol{Z}_{hs}^T(y_{hs} - \boldsymbol{X}_{hs}\boldsymbol{\beta})\\ &\quad - \bar{\boldsymbol{z}}_i^T(\boldsymbol{Z}_{is}^T\boldsymbol{Z}_{is})^{-1}\boldsymbol{Z}_{is}^T(y_{is} - \boldsymbol{X}_{is}\boldsymbol{\beta})\\ &= \sum_{\substack{h=1\\h\neq i}}^{D} w_{hi}^{\mathrm{BLUP}}\bar{\boldsymbol{z}}_h^T\{\boldsymbol{u}_h + (\boldsymbol{Z}_{hs}^T\boldsymbol{Z}_{hs})^{-1}\boldsymbol{Z}_{hs}^T\boldsymbol{e}_{hs}\}\end{aligned}$$

où w_{hi}^{BLUP} est l'équivalent BLUP de w_{hi}^{EBLUP}. La variance de δ_i peut par conséquent être estimée par

$$\hat{V}(\delta_i) = \sum_{\substack{h=1\\h\neq i}}^{D}(w_{hi}^{\mathrm{EBLUP}})^2\bar{\boldsymbol{z}}_h^T\{\hat{\boldsymbol{\Omega}} + \hat{\sigma}^2(\boldsymbol{Z}_{hs}^T\boldsymbol{Z}_{hs})^{-1}\}\bar{\boldsymbol{z}}_h. \qquad (15)$$

Si $\hat{V}(\delta_i)$ est petit comparativement à la valeur de (7) dans ce cas, on peut utiliser (8) pour estimer l'EQM de l'EBLUP. Cependant, si n_i est très petit, cette condition pourrait ne pas être vérifiée. Le cas échéant, il pourrait être souhaitable de considérer des estimateurs de l'EQM dépendant davantage d'un modèle, tel que l'estimateur de l'EQM de Prasad et Rao (PR) (Prasad et Rao 1990 ; Rao 2003, section 7.2.3). Quand un modèle à moyennes aléatoires est utilisé, mais que la variabilité entre domaines est très faible comparativement à la variabilité à l'intérieur des domaines, cette recommandation s'étend aux tailles moyennes d'échantillon de domaine, comme nous allons le montrer.

2.2.2 Estimation de l'EQM pour l'EBLUP sous le modèle à moyennes aléatoires

Le modèle à moyennes aléatoires est le cas particulier de (12) où $y_j = \beta + u_i + e_j$, avec $u_i \sim N(0,\sigma_u^2)$ et $e_j \sim N(0,\sigma^2)$. L'EBLUE de β est alors $\hat{\beta} = \sum_{h=1}^{D}\hat{\alpha}_h\bar{y}_{hs}$ avec $\hat{\alpha}_i = (\hat{\varphi}+n_i^{-1})^{-1}\{\sum_{h=1}^{D}(\hat{\varphi}+n_h^{-1})^{-1}\}^{-1}$ et $\hat{\varphi} = \hat{\sigma}_u^2/\hat{\sigma}^2$, et l'EBLUP (13) est défini par des poids de la forme

$$\begin{aligned}w_{ij}^{\mathrm{EBLUP}} &= (1-f_i)(1-\hat{\gamma}_i)\sum\nolimits_{h=1}^{D}\hat{\alpha}_h n_h^{-1}I(j\in h)\\ &\quad + \{f_i + (1-f_i)\,\hat{\gamma}_i\}\,n_i^{-1}I(j\in i)\end{aligned}$$

avec $\hat{\gamma}_i = n_i\hat{\varphi}(1+n_i\hat{\varphi})^{-1}$. Pour $j\in h$, $\hat{\mu}_j = \sum_{k\in s}\phi_{kj}y_k = \bar{y}_{hs}$ et donc

$$\begin{aligned}\hat{\lambda}_j &= (1-\phi_{jj})^2 + \sum\nolimits_{k\in s(-j)}\phi_{kj}^2\\ &= (1-n_h^{-1})^2 + (n_h-1)n_h^{-2} = (n_h-1)n_h^{-1}.\end{aligned}$$

Il s'ensuit que l'estimateur (7) de la variance de prédiction conditionnelle de $\hat{m}_i^{\mathrm{EBLUP}}$ est dans ce cas

$$\begin{aligned}\hat{V}(\hat{m}_i^{\mathrm{EBLUP}}) = (1-f_i)^2\Big[&\sum\nolimits_{h=1}^{D}\{(1-\hat{\gamma}_i)^2\hat{\alpha}_h^2 n_h^{-2}\\ &+ (N_i-n_i)^{-1}n^{-1}\}n_h s_h^2\\ &+ \hat{\gamma}_i n_i^{-1}\{2(1-\hat{\gamma}_i)\hat{\alpha}_i + \hat{\gamma}_i\}s_i^2\Big],\end{aligned}$$

où $s_h^2 = (n_h-1)^{-1}\sum_{j\in s_h}(y_j - \bar{y}_{hs})^2$, tandis que selon (9), l'estimateur du biais conditionnel de prédiction de $\hat{m}_i^{\mathrm{EBLUP}}$ est $\hat{B}(\hat{m}_i^{\mathrm{EBLUP}}) = (1-f_i)(1-\hat{\gamma}_i)(\hat{\beta} - \bar{y}_{is})$. Pour $h\neq i$, nous avons alors également

$$\begin{aligned}w_{hi}^{\mathrm{EBLUP}} &= \sum\nolimits_{j\in s_h} w_{ij}^{\mathrm{EBLUP}}\\ &= (1-f_i)\hat{\alpha}_h(1+n_i\hat{\varphi})^{-1} \approx \hat{\alpha}_h(1+n_i\hat{\varphi})^{-1}\end{aligned}$$

quand nous ignorons les termes d'ordre $O(N_i^{-1})$. Une approximation semblable de (15) mène par conséquent à

$$\begin{aligned}\hat{V}(\delta_i) &= \sum_{\substack{h=1\\h\neq i}}^{D}(w_{hi}^{\mathrm{EBLUP}})^2(\hat{\sigma}_u^2 + n_h^{-1}\hat{\sigma}^2)\\ &\approx \hat{\sigma}^2\sum_{\substack{h=1\\h\neq i}}^{D}\left(\frac{\hat{\alpha}_h}{1+n_i\hat{\varphi}}\right)^2\left(\frac{1+n_h\hat{\varphi}}{n_h}\right).\end{aligned}$$

Supposons maintenant que la taille d'échantillon est la même dans chaque petit domaine, c'est-à-dire $n_i = m$. Alors, $n = mD$, $\hat{\alpha}_h = D^{-1}$ et l'approximation de $\hat{V}(\delta_i)$ susmentionnée prend la forme

$$\hat{V}(\delta_i) = \hat{\sigma}^2 \sum_{\substack{h=1 \\ h\neq i}}^{D} \left(\frac{D^{-1}}{1+m\hat{\varphi}} \right)^2 \left(\frac{1+m\hat{\varphi}}{m} \right) \approx n^{-1}\left(1+m\hat{\varphi}\right)^{-1}\hat{\sigma}^2,$$

tandis que l'approximation correspondante de $\hat{V}(\hat{m}_i^{\text{EBLUP}})$ est

$$\begin{aligned}\hat{V}(\hat{m}_i^{\text{EBLUP}}) &\approx \sum_{h=1}^{D}(1+m\hat{\varphi})^{-2}D^{-2}m^{-1}s_h^2 \\ &\quad + (1+m\hat{\varphi})^{-2}\hat{\varphi}(2D^{-1}+m\hat{\varphi})\,s_i^2 \\ &= n^{-1}(1+m\hat{\varphi})^{-2}\left\{\left(D^{-1}\sum_{h=1}^{D}s_h^2\right)+m\hat{\varphi}(2+n\hat{\varphi})\,s_i^2\right\}.\end{aligned}$$

En comparant ces approximations de $\hat{V}(\delta_i)$ et $\hat{V}(\hat{m}_i^{\text{EBLUP}})$, nous voyons que, si $m\hat{\varphi}$ est petit (par exemple quand m et $\hat{\varphi}$ sont tous les deux petits), ces termes seront du même ordre de grandeur. Dans cette situation, nous nous attendons à ce que (8) surestime l'EQM réelle de l'EBLUP. En particulier, l'approximation de (8) quand $m\hat{\varphi}$ est petit et que N_i est grand est

$$\hat{M}(\hat{m}_i^{\text{EBLUP}}) \approx n^{-1}\left(D^{-1}\sum_{h=1}^{D}s_h^2\right)+\left(\bar{y}_{is}-D^{-1}\sum_{h=1}^{D}\bar{y}_{hs}\right)^2. \quad (16)$$

Notons que l'espérance du résidu quadratique dans le second membre de (16) quand $m\hat{\varphi}$ est petit est $(1-D^{-1})(\sigma_u^2+m^{-1}\sigma^2)=O(1)$, et il s'agit donc du terme principal dans cet estimateur dans cette situation. Cette expression peut être comparée à l'expression correspondante pour l'estimateur de l'EQM de l'EBLUP proposé par Prasad et Rao (1990). Sous le modèle à moyennes aléatoires, l'estimateur PR de l'EQM est

$$\begin{aligned}\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}}) &= (1-f_i)^2\Bigg[\ \hat{\gamma}_i m^{-1}\hat{\sigma}^2 \\ &\quad + (1-\hat{\gamma}_i)^2\left(m\sum_{h=1}^{D}\hat{\tau}_h^{-1}\right)^{-1} + N_i^{-1}(1-f_i)^{-1}\hat{\sigma}^2 \\ &\quad + \frac{2}{T}m\hat{\tau}_i^{-3}\left\{\hat{\sigma}^4\left(\frac{n-D}{\hat{\sigma}^4}+\sum_{h=1}^{D}\hat{\tau}_h^{-2}\right)\right. \\ &\quad \left.\left. + \hat{\sigma}_u^4 m^2\sum_{h=1}^{D}\hat{\tau}_h^{-2} + 2\hat{\sigma}^2\hat{\sigma}_u^2 m\sum_{h=1}^{D}\hat{\tau}_h^{-2}\ \right\}\right]\end{aligned}$$

où $\hat{\tau}_i = n_i\hat{\sigma}_u^2+\hat{\sigma}^2$ et

$$\begin{aligned}T &= \frac{n-D}{\hat{\sigma}^4}\sum_{h=1}^{D}n_h^2\,\hat{\tau}_h^{-2} \\ &\quad + \left(\sum_{h=1}^{D}\hat{\tau}_h^{-2}\right)\left(\sum_{h=1}^{D}n_h^2\hat{\tau}_h^{-2}\right)-\left(\sum_{h=1}^{D}n_h\hat{\tau}_h^{-2}\right)^2.\end{aligned}$$

En supposant $n_i = m$, que $m\hat{\varphi}$ est petit et que N_i est grand, $\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}})$ est approximé par

$$\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}}) \approx \hat{\sigma}^2\{n^{-1}+2(n-D)^{-1}\}+\hat{\sigma}_u^2. \quad (17)$$

En comparant (16) et (17), nous voyons que l'instabilité et la surestimation associées à l'utilisation de (8) dans cette situation sont toutes deux dues à l'emploi du carré du résidu au niveau du domaine à un seul degré de liberté $\bar{y}_{is} - D^{-1}\sum_{h=1}^{D}\bar{y}_{hs}$ comme estimateur de σ_u^2. Cela renforce les commentaires antérieurs voulant que (8) ne devrait généralement pas être utilisé pour estimer l'EQM de l'EBLUP si les tailles d'échantillon de domaine sont très petites ou, dans le cas particulier du modèle à moyennes aléatoires, si les tailles d'échantillon de domaine sont modérées quand la variabilité entre domaines est très faible comparativement à la variabilité à l'intérieur des domaines.

2.2.3 Estimation de l'EQM pour l'EDFM

Le deuxième prédicteur de m_i que nous considérons est l'estimateur direct fondé sur un modèle (EDFM) décrit dans Chandra et Chambers (2009). Cet estimateur est fondé sur le même modèle linéaire mixte (12) que l'EBLUP, le prédicteur EDFM étant défini comme

$$\hat{m}_i^{\text{EDFM}} = \sum\nolimits_{j\in s} w_{ij}^{\text{EDFM}} y_j = (\boldsymbol{w}_{is}^{\text{EDFM}})^T \boldsymbol{y}_s \quad (18)$$

où

$$w_{ij}^{\text{EDFM}} = \frac{I(j\in s_i)\,w_j^{\text{EBLUP}}}{\sum_{k\in s} I(k\in s_i)\,w_k^{\text{EBLUP}}}. \quad (19)$$

Ici, $I(j \in s_i)$ est la fonction indiquant que l'unité j est dans l'échantillon du domaine i, et $\boldsymbol{w}_s^{\text{EBLUP}} = (w_j^{\text{EBLUP}})$ est le vecteur des poids qui définissent l'EBLUP pour le total de population y_j sous (12), c'est-à-dire

$$\boldsymbol{w}_s^{\text{EBLUP}} = (w_j^{\text{EBLUP}}) = \mathbf{1}_n + \{\hat{\boldsymbol{H}}_s^T\boldsymbol{X}_r^T + (\mathbf{I}_n - \hat{\boldsymbol{H}}_s^T\boldsymbol{X}_s^T)\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\hat{\boldsymbol{\Sigma}}_{sr}\}\mathbf{1}_{N-n}$$

où $\mathbf{1}_n(\mathbf{1}_{N-n})$ désigne le vecteur unitaire de taille $n(N-n)$ et $\hat{\boldsymbol{H}}_s$ a été défini à la section 2.2.1. Dans ce cas l'estimation par pseudo-linéarisation de l'EQM propre au domaine de l'EDFM est effectuée en se servant de (8), avec des poids définis par (19). Notons que les valeurs prévues estimées utilisées dans (8) lorsqu'on l'applique à l'EDFM sont les mêmes que les estimations dérétrécies (14) utilisées avec l'EBLUP, ce qui reflète le fait que l'EDFM et l'EBLUP sont tous deux fondés sur le même modèle linéaire mixte (12). Cependant, les poids pour l'EDFM (19) ne sont pas calés localement, de sorte que le terme quadratique de biais dans (8) ne peut pas être ignoré quand on estime l'EQM de ce prédicteur. En outre, puisque

$$w_{hi}^{\text{EDFM}} = \sum_{j \in s_h} w_{ij}^{\text{EDFM}} = 0$$

pour $h \neq i$, nous avons $\delta_i = 0$ pour l'EDFM et l'estimateur du biais (9) donne donc de bons résultats dans ce cas.

2.2.4 Estimation de l'EQM pour l'estimateur du M-quantile

Le troisième estimateur que nous envisageons est fondé sur l'approche de modélisation des M-quantiles décrite dans Chambers et Tzavidis (2006). Cette approche ne repose pas sur l'hypothèse d'un modèle mixte linéaire sous-jacent, s'appuyant plutôt sur la caractérisation de la relation entre y_j et $\boldsymbol{x}_j$ dans le domaine i en ce qui a trait au modèle de M-quantile linéaire qui est le mieux « adapté » aux valeurs d'échantillon y_j provenant de ce domaine. Autrement dit, cette approche consiste à remplacer (12) par un modèle de la forme

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}(q_i) + \boldsymbol{e}_i \qquad (20)$$

où $\boldsymbol{\beta}(q)$ désigne le vecteur de coefficients d'un modèle linéaire pour le M-quantile de régression d'ordre q pour les valeurs de population de Y et X, et q_i désigne le coefficient de M-quantile du domaine i. Étant donné une estimation $\hat{q}_i$ de q_i, un algorithme des moindres carrés repondérés itérativement (MCRI) est utilisé pour calculer une estimation

$$\hat{\boldsymbol{\beta}}(\hat{q}_i) = \left\{ \boldsymbol{X}_s' \boldsymbol{W}_s(\hat{q}_i) \boldsymbol{X}_s \right\}^{-1} \boldsymbol{X}_s' \boldsymbol{W}_s(\hat{q}_i) \boldsymbol{y}_s \qquad (21)$$

de $\boldsymbol{\beta}(q_i)$ dans (20), et une valeur hors échantillon de y_j dans le domaine i est alors prédite par $\hat{y}_j = \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}(\hat{q}_i)$. Ici, $\boldsymbol{W}_s(\hat{q}_i)$ est la matrice diagonale des poids finaux utilisée dans l'algorithme MCRI.

Tzavidis, Marchetti et Chambers (2010) notent que la valeur de l'estimateur du M-quantile proposé dans Chambers et Tzavidis (2006) peut être interprétée comme la valeur prévue de Y dans le domaine i par rapport à un estimateur biaisé et la densité de probabilité de cette variable dans le domaine. Par conséquent, ils élaborent un estimateur amélioré du M-quantile, en remplaçant cet estimateur biaisé de la densité de probabilité par l'estimateur de la densité de probabilité de Chambers et Dunstan (1986) sous le modèle propre au domaine (1). Cela revient à prédire m_i par

$$\hat{m}_i^{\text{MQ}} = \sum_{j \in s} w_{ij}^{\text{MQ}} y_j = (\boldsymbol{w}_{is}^{\text{MQ}})^T \boldsymbol{y}_s \qquad (22)$$

où

$$\boldsymbol{w}_{is}^{\text{MQ}} = n_i^{-1} \boldsymbol{\Delta}_{is}$$

$$+ (1 - N_i^{-1} n_i) \boldsymbol{W}_s(\hat{q}_i) \boldsymbol{X}_s \{ \boldsymbol{X}_s^T \boldsymbol{W}_s(\hat{q}_i) \boldsymbol{X}_s \}^{-1} (\bar{\boldsymbol{x}}_{ir} - \bar{\boldsymbol{x}}_{is}).$$

Ici, $\bar{\boldsymbol{x}}_{is}$ et $\bar{\boldsymbol{x}}_{ir}$ sont les vecteurs des moyennes d'échantillon et hors échantillon des x_j dans le domaine i. Il n'est pas difficile de montrer que les poids conformes à (22) sont calés localement. En outre, si nous posons alors que $\hat{\mu}_j = \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}(\hat{q}_i)$, où $\hat{\boldsymbol{\beta}}(\hat{q}_i)$ est défini par (21), il est facile de voir que (9) est nul et que l'EQM propre au domaine de l'estimateur du M-quantile corrigé du biais (22) peut donc être estimé en utilisant simplement la composante de la variance de prédiction estimée (7). Puisque, dans (7), la constante $\hat{\lambda}_j$ est habituellement très proche de l'unité sous l'estimation du M-quantile, nous la fixons égale à cette valeur. Nous calculons les valeurs de (7) reliées à l'estimation sur petits domaines (EPD) sous l'approche de modélisation des M-quantiles.

Comme nous l'avons déjà fait pour l'EBLUP, nous notons que l'utilisation de (7) revient à traiter implicitement les poids définissant (22) comme fixes, ce qui n'est pas le cas en réalité, puisque la matrice $\boldsymbol{W}_s(\hat{q}_i)$ est une fonction des valeurs d'échantillon de Y. Une conséquence directe est que la pseudo-linéarisation fondée sur l'estimation de l'EQM du prédicteur du M-quantile par la voie de (7) est une approximation de premier ordre de la valeur vraie de l'EQM de cet estimateur. Néanmoins, puisque tenir compte de la variabilité des poids dans la définition de l'estimateur du M-quantile complique considérablement l'estimation de son EQM – voir Street, Carroll et Ruppert (1988) pour un examen de cette question dans le contexte de la M-estimation « classique » des coefficients de régression – il est intéressant de voir comment l'estimateur (7) relativement simple se comporte quand il est utilisé pour estimer cette EQM.

2.3 L'estimation de l'EQM pour l'EBLUP synthétique pseudo-linéaire

Dans de nombreuses applications d'estimation sur petits domaines, il existe des domaines ne contenant pas d'unités échantillonnées et l'on utilise alors une estimation synthétique. Bien que ce genre d'estimateur n'entre pas dans la classe des estimateurs pseudo-linéaires examinés dans le présent article, les notions qui sous-tendent l'estimateur conditionnel de l'EQM (8) peuvent leur être appliquées également. Pour le voir, supposons que ces domaines sont numérotés les derniers, c'est-à-dire que si D^+ domaines ont un échantillon non nul, alors $n_h > 0$ pour $h \leq D^+$ et $n_h = 0$ pour $h > D^+$. Pour $i > D^+$ l'« EBLUP synthétique » pour m_i est

$$\hat{m}_i^{\text{SYN-EBLUP}} = \bar{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}^{\text{EBLUE}} = (\boldsymbol{w}_{is}^{\text{SYN-EBLUP}})^T \boldsymbol{y}_s$$

$$= \sum_{h=1}^{D^+} \sum_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} y_j \qquad (23)$$

où

$$\boldsymbol{w}_{is}^{\text{SYN-EBLUP}} = (w_{ij}^{\text{SYN-EBLUP}}) = \hat{\boldsymbol{H}}_s^T \bar{\boldsymbol{x}}_i.$$

Il est clair que (23) est un estimateur pseudo-linéaire et nous pouvons donc utiliser (7) pour estimer sa variance de prédiction, en notant que, puisque $n_i = 0$, $a_{ij} = N_i w_{ij}^{\text{EBLUP}}$ et donc (7) devient

$$\hat{V}(\hat{m}_i^{\text{SYN-EBLUP}}) = \sum_{j\in s}\{(w_{ij}^{\text{SYN-EBLUP}})^2 + N_i^{-1}n^{-1}\}\hat{\lambda}_j^{-1}(y_j - \hat{\mu}_j)^2. \quad (24)$$

Malheureusement, comme il n'y a pas d'échantillon dans le domaine *i*, nous ne pouvons pas utiliser (9) pour estimer le biais propre au domaine (2) de $\hat{m}_i^{\text{SYN-EBLUP}}$. Cependant, sous le modèle mixte linéaire (12), la valeur prévue de ce biais est

$$E(\hat{m}_i^{\text{SYN-EBLUP}} - m_i) = \sum_{h=1}^{D^+}\sum_{j\in s_h} w_{ij}^{\text{SYN-EBLUP}}\left(\boldsymbol{x}_j^T\boldsymbol{\beta} + \boldsymbol{z}_j^T\boldsymbol{u}_h\right) - \bar{\boldsymbol{x}}_i^T\boldsymbol{\beta} - \bar{\boldsymbol{z}}_i^T\boldsymbol{u}_i.$$

L'espérance conditionnelle du carré de ce biais prévu, sachant les effets de domaine $\boldsymbol{u}_s = (u_h\,;\, h = 1, \ldots, D^+)$ pour les domaines échantillonnés, est

$$E\{E^2(\hat{m}_i^{\text{SYN-EBLUP}} - m_i) \mid \boldsymbol{X}, \boldsymbol{u}_s\} = \left\{\sum_{h=1}^{D^+}\sum_{j\in s_h} w_{ij}^{\text{SYN-EBLUP}}(\boldsymbol{x}_j^T\boldsymbol{\beta} + \boldsymbol{z}_j^T\boldsymbol{u}_h) - \bar{\boldsymbol{x}}_i^T\boldsymbol{\beta}\right\}^2 + \bar{\boldsymbol{z}}_i^T\boldsymbol{\Omega}\bar{\boldsymbol{z}}_i,$$

ce qui suggère immédiatement que, pour un domaine *i* non échantillonné, nous estimions le carré du biais de l'estimateur synthétique $\hat{m}_i^{\text{SYN-EBLUP}}$ par

$$\hat{B}^2(\hat{m}_i^{\text{SYN-EBLUP}}) = \left\{\sum_{h=1}^{D^+}\sum_{j\in s_h} w_{ij}^{\text{SYN-EBLUP}}\left(\boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}^{\text{EBLUE}} + \boldsymbol{z}_j^T\tilde{\boldsymbol{u}}_h\right) - \bar{\boldsymbol{x}}_i^T\hat{\boldsymbol{\beta}}^{\text{EBLUE}}\right\}^2 + \bar{\boldsymbol{z}}_i^T\hat{\boldsymbol{\Omega}}\bar{\boldsymbol{z}}_i. \quad (25)$$

Ici, $\tilde{\boldsymbol{u}}_h$ est l'effet estimé « dérétréci » pour le domaine échantillonné *h* – voir (14). L'estimateur de l'EQM que nous proposons pour $\hat{m}_i^{\text{SYN-EBLUP}}$ est alors la somme de (24) et (25). Notons que, contrairement à (8), cet estimateur de l'EQM ne contient aucune information provenant du domaine *i*, et n'est donc pas un estimateur de l'EQM propre au domaine de (23). En particulier, sa validité dépend entièrement du fait que le modèle mixte (12) est vérifié, et il n'est donc pas robuste à l'erreur de spécification de ce modèle.

3. Études en simulation de l'estimateur de l'EQM proposé

À la présente section, nous décrivons les résultats de cinq études en simulation destinées à évaluer la performance de l'approche d'estimation conditionnelle de l'EQM décrite à la section précédente. Trois de ces études sont des simulations fondées sur un modèle dans lesquelles les données de population ont été générées au moyen du modèle mixte linéaire (12). Les deux autres sont des simulations fondées sur le plan de sondage, dans lesquelles les données de population ont été dérivées de deux ensembles de données d'enquête réelles pour lesquelles l'estimation sur petits domaines linéaire présente un intérêt.

Comme nous nous intéressons à la robustesse au biais, dans chacune des cinq études, le principal indicateur de performance d'un estimateur de l'EQM est son biais relatif médian, défini par

$$\text{BR}(M) = \underset{i}{\text{médiane}}\left\{M_i^{-1}K^{-1}\sum_{k=1}^{K}(\hat{M}_{ik} - M_i)\right\}\times 100.$$

Ici, l'indice inférieur *i* désigne les petits domaines et l'indice supérieur *k*, les *K* simulations Monte Carlo, avec $\hat{M}_{ik}$ désignant la valeur de l'estimateur de l'EQM pour la simulation *k* dans le domaine *i*, et M_i désignant l'EQM réelle (c'est-à-dire Monte Carlo) dans le domaine *i*. Puisque nous préférerions naturellement utiliser le plus stable de deux estimateurs de l'EQM approximativement sans biais, nous mesurons aussi la stabilité d'un estimateur de l'EQM par la médiane de la racine carrée de son erreur quadratique moyenne relative en pourcentage,

$$\text{REQMR}(M) = \underset{i}{\text{médiane}}\left\{\sqrt{K^{-1}\sum_{k=1}^{K}\left(\frac{\hat{M}_{ik} - M_i}{M_i}\right)^2}\right\}\times 100.$$

Bien que le but du présent article ne soit pas de comparer diverses méthodes d'estimation sur petits domaines, il est utile de relier la performance de l'estimation de l'EQM pour une méthode particulière d'estimation sur petits domaines à la performance réelle d'estimation de cette méthode. Par conséquent, nous fournissons deux mesures de la performance relative des méthodes d'estimation sur petits domaines qui ont été utilisées dans nos simulations. Il s'agit de la médiane du biais relatif en pourcentage

$$\text{BR}(m) = \underset{i}{\text{médiane}}\left\{\bar{m}_i^{-1}K^{-1}\sum_{k=1}^{K}(\hat{m}_{ik} - m_{ik})\right\}\times 100$$

et la médiane de la racine carrée de l'erreur quadratique moyenne relative en pourcentage

$$\text{REQMR}(m) = \underset{i}{\text{médiane}}\left\{\sqrt{K^{-1}\sum_{k=1}^{K}\left(\frac{\hat{m}_{ik} - m_{ik}}{m_{ik}}\right)^2}\right\}\times 100$$

des estimations $\hat{m}_{ik}$ générées par une méthode d'estimation. Notons qu'ici, $\bar{m}_i = K^{-1}\sum_{k=1}^{K} m_{ik}$.

3.1 Simulations fondées sur un modèle

La première étude en simulation fondée sur un modèle s'appuyait sur des données de population générées sous le modèle mixte (12) avec des effets aléatoires gaussiens. Une taille de population de $N = 15\,000$, avec $D = 30$ petits

domaines a été utilisée. Les tailles de population dans les petits domaines ont été réparties uniformément sur l'intervalle [443, 542] et maintenues fixes au cours des simulations. Dans chaque simulation, les valeurs de population de Y ont été produites sous le modèle à ordonnées à l'origine aléatoires $y_j = 500 + 1,5x_j + u_i + e_j$, avec x_j tiré d'une loi du khi-carré avec 20 degrés de liberté. Les effets de domaine u_i et les effets individuels e_j ont été tirés indépendamment des lois $N(0, \sigma_u^2)$ et $N(0, \sigma^2)$ respectivement, avec les valeurs de σ_u et σ présentées aux lignes SIM1-A et SIM1-B du tableau 1. Un échantillon de taille $n = 600$ a été sélectionné dans chaque population simulée, avec des tailles d'échantillon de domaine proportionnelles aux populations de domaine fixes, ce qui a produit une taille médiane d'échantillon de domaine de $n_i = 20$. Le tirage des échantillons a été effectué par échantillonnage aléatoire stratifié, les strates étant définies par les petits domaines. Un total de $K = 1\,000$ simulations ont été exécutées.

Les conditions pour la deuxième étude en simulation fondée sur un modèle étaient les mêmes que pour la première, excepté que les effets aléatoires au niveau du domaine et les effets aléatoires au niveau de l'individu ont été tirés indépendamment de lois du khi-carré corrigées par la moyenne, respectivement. Les valeurs correspondantes des variances au niveau du domaine et au niveau individuel sont présentées aux lignes SIM2-A et SIM2-B dans le tableau 1. Enfin, dans la troisième étude en simulation fondée sur un modèle, les conditions ont été maintenues les même que dans les simulations SIM1-A et SIM1-B pour les domaines 1 à 25, mais pour les domaines 26 à 30, les effets de domaine ont été tirés indépendamment d'une loi normale ayant une plus grande variance. Nous désignons ce cas comme un mélange dans le tableau 1, avec les variances pour les domaines 1 à 25 présentées aux lignes SIM3-A et SIM3-B, et les variances pour les domaines 26 à 30 présentées aux lignes SIM3-A* et SIM3-B*. Notre objectif dans cette troisième simulation était d'étudier le comportement des différentes méthodes d'estimation de l'EQM pour les domaines « aberrants » et nous présentons donc les valeurs pour les domaines 1 à 25 et pour les domaines 26 à 30 séparément dans les tableaux 2 et 4. Nous avons également reproduit chacun des trois scénarios susmentionnés en utilisant une plus petite taille globale d'échantillon de $n = 150$ (avec une taille médiane d'échantillon de domaine de $n_i = 5$). Ces simulations supplémentaires nous ont permis d'étudier l'effet de la réduction de la taille d'échantillon sur les propriétés des estimateurs de l'EQM.

Le tableau 2 montre le biais relatif médian BR(m) et la racine carrée de l'erreur quadratique moyenne relative médiane REQMR(m) des méthodes d'estimation sur petits domaines étudiées dans nos simulations pour les deux tailles d'échantillon ($n = 600$ et 150). Il s'agit de l'estimateur par la régression synthétique (voir Rao 2003, page 136), de l'EBLUP avec les poids définis par (13), de l'EDFM avec les poids définis par (18) et de l'estimateur du M-quantile défini par les poids (22). Les différences entre les divers estimateurs sur petits domaines présentés dans le tableau 2 sont essentiellement celles attendues. Le biais n'est pas vraiment un problème (ce à quoi il fallait s'attendre, puisque les données de population suivent un modèle linéaire dans tous les cas), tandis que pour les scénarios de simulation 1 et 2, l'estimateur indirect (EBLUP) est le plus efficace si l'on s'en tient à la REQMR. L'estimateur du M-quantile est celui qui donne les meilleurs résultats pour SIM3-A* et SIM3-B* avec $n_i = 20$, mais son écart par rapport à l'estimateur par la régression synthétique se réduit pour le scénario avec les tailles d'échantillon de domaine plus petites. Notons que, dans ce cas, les poids du M-quantile (22) sont fondés sur une estimation robuste aux valeurs aberrantes du coefficient du M-quantile $\hat{q}_i$ pour le domaine i, défini par la médiane (plutôt que la moyenne) des coefficients des M-quantiles des unités échantillonnées dans ce domaine. En outre, à mesure que les tailles d'échantillon diminuent, la REQMR de tous les estimateurs augmente, mais la performance relative de ces estimateurs reste la même. Sous l'hypothèse de normalité, l'EBLUP est meilleur que l'estimateur du M-quantile, mais les différences entre ces deux estimateurs deviennent plus faibles à mesure que nous nous écartons de la normalité, tandis que l'estimateur du M-quantile est plus efficace dans les scénarios avec modèle de mélange.

Le tableau 3 présente les divers estimateurs de l'EQM étudiés dans nos simulations qui sont fondés sur l'approche proposée dans le présent article. Ils sont désignés collectivement comme des estimateurs « conditionnels » de l'EQM ci-après. Dans le tableau 4, nous montrons les propriétés des estimateurs de l'EQM pour les estimateurs sur petits domaines considérés au tableau 2. Notons qu'en plus des estimateurs conditionnels de l'EQM, nous fournissons les résultats pour trois autres estimateurs de l'EQM pour l'EBLUP, parmi lesquels PR0 désigne l'estimateur proposé par Prasad et Rao (1990), voir Rao (2003, section 6.2.6). Il convient de souligner que PR0 n'est pas un estimateur de l'EQM propre au domaine de l'EBLUP, mais de son EQM sous le modèle linéaire mixte (12), c'est-à-dire dont la moyenne est calculée sur les réalisations possibles de l'effet de domaine. Par contre, les estimateurs PR1 et PR2 de l'EQM présentés au tableau 4 sont les versions propres au domaine de l'estimateur PR0 proposé dans Rao (2003, section 6.3.2, expressions 6.3.15 et 6.3.16 respectivement). Enfin, nous notons que l'estimateur de l'EQM de l'estimateur par la régression synthétique que nous avons utilisé dans nos simulations est son estimateur de variance fondé sur un modèle de régression de population à effets fixes. Nous le désignons par VReg.

Les résultats exposés au tableau 4 sont axés sur les biais médians BR(M) et la racine carrée de l'erreur quadratique moyenne relative médiane REQMR(M) des divers estimateurs de l'EQM. Naturellement, étant donné que toutes les hypothèses qui le sous-tendent sont satisfaites, l'estimateur PR0 et ses variantes propres au domaine, PR1 et PR2, donnent de très bons résultats dans les deux scénarios normaux (SIM1-A et SIM1-B) et dans les deux scénarios du khi-carré (SIM2-A et SIM2-B), le biais étant presque nul ($n_i = 20$) ou faible quand les tailles d'échantillon dans les domaines sont très faibles. Par contre, pour l'estimateur de l'EQM de l'estimateur par la régression synthétique, nous observons un biais relatif important sous tous les scénarios de simulation.

Tableau 1
Valeurs des paramètres utilisés dans les simulations fondées sur un modèle

Type	Simulation	σ_u^2	σ^2	$\rho = \sigma_u^2(\sigma_u^2 + \sigma^2)^{-1}$
Gaussien	SIM1-A	10,40	94,09	0,1
	SIM1-B	40,32	94,09	0,3
khi-carré	SIM2-A	2,0	10,0	0,1667
	SIM2-B	4,0	10,0	0,2857
Mélange (domaines 1 à 25)	SIM3-A	10,40	94,09	0,10
	SIM3-B	40,32	94,09	0,30
Mélange (domaines 26 à 30)	SIM3-A*	225,0	94,09	0,7051
	SIM3-B*	225,0	94,09	0,7051

Tableau 2
Biais relatif médian BR(m) et racine carrée de l'erreur quadratique moyenne relative médiane REQMR(m) des estimateurs des moyennes de petit domaine dans les simulations fondées sur un modèle

Méthode de pondération	Simulation							
	SIM1-A	SIM1-B	SIM2-A	SIM2-B	SIM3-A	SIM3-B	SIM3-A*	SIM3-B*
BR(m), n_i médian = 20								
Régression	0,005	0,005	0,000	0,000	0,004	0,004	0,006	0,006
EBLUP, (13)	0,005	0,006	0,004	-0,002	0,004	0,005	0,006	0,005
EDFM, (18)	0,006	0,006	0,005	-0,008	0,007	0,007	0,001	0,001
M-quantile, (22)	0,009	0,008	-0,002	0,002	0,015	0,015	-0,013	-0,013
REQMR(m), n_i médian = 20								
Régression	0,40	0,40	0,13	0,13	0,40	0,40	0,41	0,41
EBLUP, (13)	0,35	0,38	0,12	0,13	0,37	0,38	0,45	0,42
EDFM, (18)	0,55	0,55	0,41	0,43	0,56	0,56	0,55	0,55
M-quantile, (22)	0,41	0,41	0,13	0,13	0,41	0,41	0,36	0,36
BR(m), n_i médian = 5								
Régression	-0,002	-0,003	-0,001	0,002	-0,003	-0,004	0,011	0,011
EBLUP, (13)	0,001	0,005	-0,002	0,003	0,002	-0,001	0,008	0,011
EDFM, (18)	-0,002	-0,002	-0,005	0,004	-0,001	-0,002	-0,002	-0,002
M-quantile, (22)	-0,001	-0,001	-0,001	0,001	-0,003	-0,003	0,014	0,014
REQMR(m), n_i médian = 5								
Régression	0,81	0,81	0,26	0,26	0,82	0,82	0,80	0,80
EBLUP, (13)	0,53	0,69	0,19	0,22	0,61	0,71	1,00	0,87
EDFM, (18)	1,13	1,13	0,83	0,83	1,13	1,13	1,13	1,13
M-quantile, (22)	0,81	0,81	0,26	0,26	0,81	0,81	0,80	0,80

Tableau 3
Définitions des estimateurs conditionnels de l'EQM pour diverses méthodes de pondération

Méthode de pondération	Définition de $\hat{\mu}_j, j \in i$	Estimateur de l'EQM
EBLUP (13)	(14)	(8)
EDFM (18)	(14)	(8)
M-quantile (22)	$x_i^T\hat{\beta}(\hat{q}_i)$	(7) avec $\hat{\lambda}_i = 1$
EBLUP synthétique (23)	(14)	(24) + (25)

Tableau 4
Biais relatif médian BR(*M*) et racine carrée de l'erreur quadratique moyenne relative médiane REQMR(*M*) pour les estimateurs de l'EQM dans les simulations fondées sur un modèle

Méthode de pondération	Estimateur de l'EQM	Simulation							
		SIM1-A	SIM1-B	SIM2-A	SIM2-B	SIM3-A	SIM3-B	SIM3-A*	SIM3-B*
		BR(*M*), n_i médian = 20							
Régression	VReg	7,59	21,82	11,81	20,78	23,66	34,27	23,97	34,64
EBLUP, (13)	PR0	-0,83	-0,72	0,56	1,16	3,44	0,71	-15,65	-6,51
	PR1	-0,97	-0,72	0,64	1,08	2,94	0,56	-13,70	-5,81
	PR2	-0,92	-0,72	0,64	1,16	3,20	0,61	-14,65	-6,19
	Conditionnel	3,89	-0,89	3,06	0,93	-0,05	-0,54	-2,56	-1,59
EDFM, (18)	Conditionnel	-0,81	-0,80	-0,06	-0,42	-0,75	-0,75	-0,98	-0,98
M-quantile, (22)	Conditionnel	-3,10	-1,66	-0,09	-1,90	-5,04	-3,17	11,26	11,04
		REQMR(*M*), n_i médian = 20							
Régression	VReg	18	51	30	53	59	85	60	86
EBLUP, (13)	PR0	12	7	15	10	11	7	29	14
	PR1	14	7	17	11	10	7	27	13
	PR2	12	7	16	10	11	7	28	13
	Conditionnel	62	31	70	49	31	30	42	32
EDFM, (18)	Conditionnel	70	70	126	128	71	71	67	67
M-quantile, (22)	Conditionnel	32	34	49	48	31	32	48	48
		BR(*M*), n_i médian = 5							
Régression	VReg	5,59	19,17	10,35	19,12	20,92	30,91	22,93	33,00
EBLUP, (13)	PR0	3,51	-0,20	2,42	1,19	12,79	3,86	-30,64	-15,92
	PR1	3,04	-0,50	2,13	1,00	10,84	3,10	-25,77	-13,62
	PR2	3,16	-0,31	2,31	1,11	11,81	3,48	-28,16	-14,77
	Conditionnel	37,52	4,38	24,11	8,93	8,18	1,50	-0,66	-0,68
EDFM, (18)	Conditionnel	-0,24	-0,21	0,02	-0,09	-0,62	-0,33	1,29	1,24
M-quantile, (22)	Conditionnel	-7,60	-6,17	5,70	5,00	-5,95	-5,60	5,89	3,60
		REQMR(*M*), n_i médian = 5							
Régression	VReg	17	46	33	51	54	78	59	83
EBLUP, (13)	PR0	31	14	33	22	36	16	53	31
	PR1	48	18	44	28	34	16	48	29
	PR2	36	15	36	24	34	15	50	29
	Conditionnel	234	81	193	121	86	66	86	70
EDFM, (18)	Conditionnel	79	79	133	129	79	79	83	83
M-quantile, (22)	Conditionnel	62	63	90	97	63	63	122	102

L'estimateur conditionnel de l'EQM pour l'EBLUP présente un biais positif sous les scénarios normal (SIM1A) et du khi-carré (SIM2A), particulièrement dans le cas d'une corrélation intra-grappe modérée (3,89 % et 37,52 % pour la loi normale avec 20 et 5 unités dans chaque domaine respectivement, et 3,06 % et 24,11 % pour la loi du khi-carré avec 20 et 5 unités dans chaque domaine respectivement). Ce biais augmente lorsque la taille d'échantillon diminue. Cependant, la situation change quand nous examinons les résultats pour les composantes aberrantes des scénarios avec modèle de mélange (SIM3-A* et SIM3-B*). Ici, nous observons un biais négatif important pour les trois versions de l'estimateur PR (variant de -30,64 % à -5,81 % selon la taille d'échantillon de domaine). Comparativement, l'estimateur conditionnel de l'EQM pour l'EBLUP présente maintenant un biais négatif plus faible (-2,56 % et -0,66 %), tandis que le même estimateur de l'EQM appliqué à l'estimateur du M-quantile présente un biais vers le haut. Pour l'EDFM, l'estimateur conditionnel de l'EQM est essentiellement sans biais. Comme, en ce qui concerne l'estimation de l'EQM, un biais positif est préférable à un biais négatif, il semble évident que l'estimateur conditionnel de l'EQM proposé permet de mieux traiter cette situation de valeur aberrante. La figure 1 illustre graphiquement ce point pour une taille d'échantillon de $n = 600$. Ici, nous montrons les REQM propres au domaine et la moyenne (sur les simulations) des REQM estimées dans chacun des 30 domaines pour les simulations de mélange SIM3-A et SIM3-A*, la droite verticale définissant les cinq domaines « aberrants ». Dans la partie supérieure de ce graphique, nous voyons que l'estimateur PR0 n'arrive pas à déceler la hausse de l'EQM de l'EBLUP pour ces domaines « aberrants », puisqu'ils présentent un biais légèrement élevé dans les domaines « se comportant de la manière attendue », puis un biais assez

faible dans les domaines « aberrants ». Par contre, l'estimateur conditionnel de l'EQM pour l'EBLUP et pour l'EDFM suit d'assez près les REQM propres au domaine, tandis que le même estimateur de l'EQM fondé sur des poids M-quantile ont tendance à présenter un faible biais dans les domaines « ayant le comportement attendu » et un biais élevé dans les domaines « aberrants », ce qui représente, pourrait-on soutenir, un meilleur résultat que celui observé pour l'estimateur PR0 dans cette simulation. Il convient de souligner ici que, dans certaines circonstances, un modèle supposé peut être révisé après la détection de valeurs aberrantes. Cependant, cela nécessite un nombre suffisamment grand de valeurs aberrantes décelées pour permettre leur modélisation distincte, ce qui est peu probable en pratique. En outre, l'extrapolation de ces résultats au cas des très petites tailles d'échantillon de domaine doit se faire avec beaucoup de prudence, à cause de l'instabilité que peut manifester l'estimateur conditionnel de l'EQM dans ce cas.

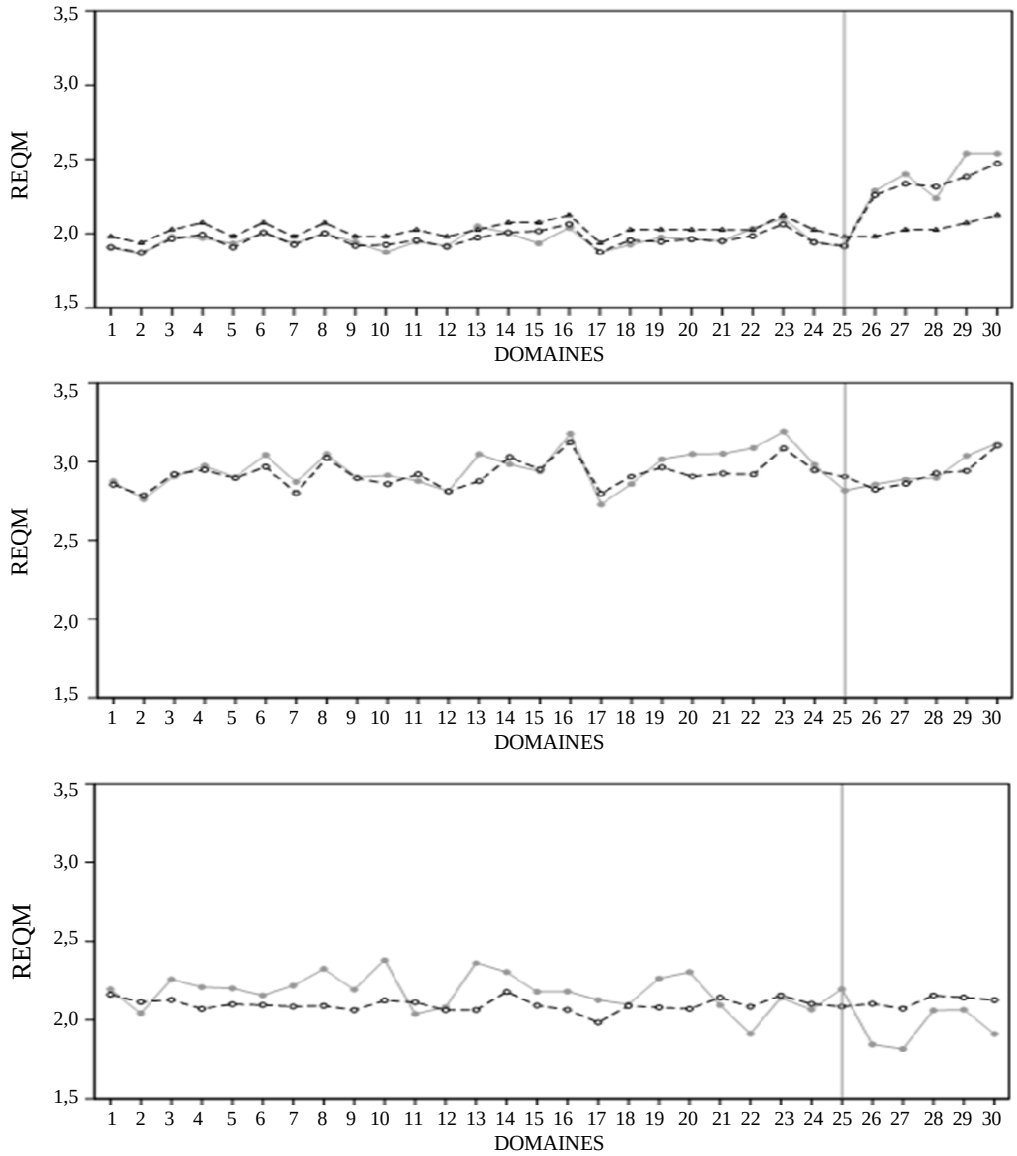

Figure 1 Valeurs propres au domaine de la REQM réelle (trait plein) et de la REQM estimée moyenne (trait interrompu) obtenues dans les simulations fondées sur un modèle de mélange SIM3-A et SIM3-A*. Les valeurs de l'estimateur PR0 sont indiquées par Δ, tandis que celles de l'estimateur conditionnel sont indiquées par o. Les graphiques montrent les résultats pour les estimateurs EBLUP (en haut), EDFM (au centre) et M-quantile (en bas). La droite verticale sépare les domaines 26 à 30 avec effets de « valeurs aberrantes » des domaines 1 à 25 se comportant « de la manière attendue ». La taille totale d'échantillon est de 600 avec des tailles d'échantillon de domaine égales à 20

Le tableau 4 montre aussi les REQM relatives des différents estimateurs de l'EQM pour les trois types de simulation fondée sur un modèle. Ici, nous voyons que les trois versions de l'estimateur PR de l'EQM de l'EBLUP sont plus stables que l'estimateur conditionnel de l'EQM de l'EBLUP (12 % pour PR *vs* 62 % pour l'EQM conditionnelle pour SIM1-A avec $n_i = 20$ et 31 % pour PR *vs* 234 % pour l'EQM conditionnelle pour SIM1-A avec $n_i = 5$). Ces différences diminuent toutefois sous le scénario SIM3-A* et SIM3-B*, bien que l'estimateur PR de l'EQM demeure plus stable (13 % pour PR *vs* 32 % pour l'estimateur conditionnel de l'EQM pour SIM3-B* avec $n_i = 20$ et 29 % pour l'estimateur PR de l'EQM *vs* 70 % pour l'estimateur conditionnel de l'EQM pour SIM3-B* avec $n_i = 5$). Il en est de même pour les estimateurs conditionnels de l'EQM de l'EDFM et des estimateurs des M-quantile. Essentiellement, quand les données d'échantillon suivent un modèle linéaire mixte, l'estimateur PR de l'EQM est très stable, tandis que l'estimateur conditionnel de l'EQM est plus variable.

En résumé, même si toutes les méthodes d'estimation de l'EQM que nous avons évaluées présentaient un certain biais pour les très petites tailles d'échantillon de domaine, les résultats de nos simulations fondées sur un modèle donnent des preuves que, pour des tailles d'échantillon de domaine plus grandes, la méthode d'estimation conditionnelle de l'EQM (8) est robuste au biais lorsqu'elle est appliquée aux trois estimateurs sur petits domaines pseudo-linéaires EBLUP, EDFM et M-quantile. Pour les très petites tailles d'échantillon de domaine, sa robustesse au biais est moins évidente. Comme il faut s'y attendre, l'estimateur PR0 de l'EQM « moyenne sur les domaines » dépendant du modèle pour l'EBLUP présente un biais en cas d'échec du modèle. Le fait que nous ayons observé un comportement assez semblable pour les variantes propres au domaine PR1 et PR2 de cet estimateur de l'EQM indique que « propre au domaine » ne signifie pas nécessairement « robuste au biais ». En particulier, le fait que les estimateurs PR1 et PR2 se comportent de manière très semblable à l'estimateur PR0 peut être dû au fait que les composantes propres au domaine de PR1 et PR2 sont d'un ordre plus faible et que les trois estimateurs de l'EQM ont le même terme principal, qui n'est pas propre au domaine. Nos résultats montrent aussi que l'estimateur conditionnel de l'EQM (8) est beaucoup plus variable que l'estimateur PR de l'EQM dépendant du modèle, même pour des tailles d'échantillon de domaine modérées.

3.2 Simulations fondées sur un plan de sondage

Que se passe-t-il quand, dans les conditions réelles, nous ne sommes pas certains que nos données suivent un modèle linéaire mixte? Afin d'étudier cette situation, nous présentons les résultats de deux études en simulation fondées sur un plan de sondage, toutes deux s'appuyant sur des populations réelles, pour lesquelles l'hypothèse d'un modèle linéaire est essentiellement une approximation. La première étude porte sur un échantillon de 3 591 ménages répartis entre $D = 36$ districts de l'Albanie qui ont participé à l'étude de mesure du niveau de vie de 2002 en Albanie. Cet échantillon a été répété par la méthode bootstrap pour créer une population réaliste de $N = 724\,782$ ménages par rééchantillonnage avec remise avec probabilité proportionnelle au poids de sondage du ménage. Un total de $K = 1\,000$ échantillons aléatoires stratifiés indépendants ont ensuite été tirés de cette population bootstrap, la taille totale d'échantillon étant égale à celle de l'échantillon original et les strates étant définies par les districts. Dans les districts, les tailles d'échantillon étaient les mêmes que dans l'échantillon original et variaient entre 8 et 688 (avec une taille médiane d'échantillon de district de 56). La variable d'intérêt Y était les dépenses de consommation par membre du ménage (DCM) et X était défini par trois variables binaires prenant les valeurs 0 ou 1 (propriété d'une télévision, d'une antenne parabolique et d'un terrain). Le but était d'estimer la valeur moyenne des DCM pour chaque district. Dans l'enquête originale de 2002, la relation linéaire entre les DCM et les trois variables qui constituaient X était assez faible, leur pouvoir prédicteur étant assez faible. En particulier, seule la propriété d'un terrain était reliée de manière significative au DCM au seuil de signification de 5 %. L'extension du modèle linéaire afin d'inclure des ordonnées à l'origine aléatoires, définie par les effets de district indépendants, améliorait considérablement l'ajustement. Ces ordonnées à l'origine aléatoires expliquaient environ 10 % de la variation résiduelle dans ce modèle.

La deuxième étude en simulation fondée sur un plan de sondage s'appuyait sur la version « exempte de valeurs aberrantes » de la population de fermes de culture à grande échelle australiennes utilisées dans les études en simulation décrites dans Chambers et Tzavidis (2006) et dans Chandra et Chambers (2009). En particulier, cette population a été définie en traitant par la méthode du bootstrap un sous-échantillon de 1 579 fermes « non aberrantes » qui avaient participé à l'Australian Agricultural et Grazing Industries Survey (AAGIS) pour créer une population de $N = 78\,072$ fermes par rééchantillonnage à partir de l'échantillon original de l'AAGIS avec probabilités proportionnelles aux poids de sondage des fermes. Dans ce cas, les petits domaines d'intérêt étaient les $D = 28$ régions d'exploitation agricole à grande échelle représentées dans ce sous-échantillon. La simulation fondée sur le plan de sondage a été exécutée en tirant $K = 1\,000$ échantillons aléatoires stratifiés indépendants de cette population bootstrap, les strates

étant définies par les régions et les tailles d'échantillon de strate étant définies par celles de l'échantillon original de l'AAGIS. Ces tailles d'échantillon varient de 6 à 117, avec une taille d'échantillon de région médiane de 53. Ici, Y représente le coût déboursé total (CDT) associé à l'exploitation de la ferme, et X est un vecteur qui comprend la superficie de la ferme (superficie), les effets pour six poststrates définies par trois zones climatiques et deux tranches de taille de ferme, ainsi que l'interaction de ces variables. Dans l'échantillon original de l'AAGIS, la relation entre le CDT et la superficie varie de manière significative entre les six poststrates, la valeur globale du R-carré étant d'environ 0,46 après la suppression de deux valeurs aberrantes. Par conséquent, dans le modèle de prédiction, les effets fixes ont été spécifiés comme correspondant à un ajustement linéaire distinct du CDT en fonction de la superficie dans chaque poststrate. Les effets aléatoires (nécessaires pour le calcul de l'EBLUP et de l'EDFM, mais non du prédicteur du M-quantile) ont été définis comme des effets régionaux indépendants (c'est-à-dire une spécification d'ordonnées à l'origine aléatoires) en se basant sur le fait que, dans l'échantillon original de l'AAGIS, la composante de variance entre régions explique environ 3 % de la variabilité résiduelle totale après suppression des deux valeurs aberrantes. Le but était d'estimer les moyennes régionales du CDT.

Les tableaux 5 et 6 montrent les biais relatifs médians et les REQM relatives médianes des divers estimateurs et des estimateurs correspondants de l'EQM de ces estimateurs en se basant sur les $K = 1\,000$ échantillons stratifiés indépendants tirés des populations albanienne et de l'AAGIS, respectivement. Il convient de souligner qu'en dépit du fait que les modèles linéaires mixtes ajustés aux données albaniennes ainsi que de l'AAGIS semblent raisonnables, les gains dus à l'adoption de méthodes d'estimation sur petits domaines fondées sur ces modèles ne donnent pas lieu à des améliorations considérables de l'efficacité étant donné les tailles originales des échantillons régionaux pour ces enquêtes. Par ailleurs, l'estimateur du M-quantile, qui n'est pas fondé sur une spécification d'effets aléatoires, donne de bons résultats en ce qui concerne tant le biais que l'EQM pour la population de l'AAGIS dans ce cas (tableau 6, n_i médian = 53), tandis que l'EBLUP, même s'il donne les meilleurs résultats en ce qui concerne l'EQM pour la population albanienne (tableau 5, n_i médian = 56), affiche aussi les biais les plus élevés (mais néanmoins faibles, la valeur la plus grande étant inférieure à 2 %) pour les deux populations sachant la taille d'échantillon de domaine originale. L'estimateur par régression des données d'enquête donne de bons résultats, quoique, pour les deux populations, il existe des estimateurs indirects dont les propriétés sont un peu meilleures. Les simulations fondées sur un plan de sondage s'appuyant sur les populations albanienne et de l'AAGIS ont également été exécutées en prenant des tailles d'échantillon de domaine plus petites que celles utilisées dans les enquêtes originales. En particulier, pour la population albanienne, la taille globale d'échantillon a été réduite à $n = 291$ (avec une taille d'échantillon de district médiane de 9). De même, pour la population de l'AAGIS, la taille globale d'échantillon a été réduite à $n = 233$ (avec une taille médiane d'échantillon régional de 8). Comme prévu, la REQM des estimateurs ponctuels augmente à mesure que les tailles d'échantillon de domaine diminuent. Dans l'ensemble, sous ces tailles d'échantillon plus petites, la REQM de l'EBLUP s'améliore comparativement à celle des autres estimateurs. Cependant, puisqu'il est douteux que ces plans à taille d'échantillon réduite soient raisonnables, nous n'accordons pas trop d'importance aux résultats qui en découlent, nous contentant seulement de mentionner qu'ils sont utiles pour évaluer la performance des estimateurs de l'EQM avec des données réelles et de très petites tailles d'échantillon.

Si nous nous penchons sur les résultats de simulation obtenus en utilisant les tailles originales d'échantillons régionaux, nous voyons que, pour l'EBLUP, les trois estimateurs PR de l'EQM présentent un biais vers le haut important dans les deux ensembles de simulations fondées sur un plan de sondage ainsi qu'une instabilité plus grande (population albanienne, tableau 5) que celle des estimateurs conditionnels de l'EQM, ou comparable à celle-ci (population de l'AAGIS, tableau 6). Pour la population albanienne, les trois versions de l'estimateur conditionnel de l'EQM sont essentiellement sans biais, tandis que pour la population de l'AAGIS, elles présentent un biais faible ou modéré.

Il convient de souligner que, pour la population albanienne (tableau 5), les propriétés relatives des estimateurs PR de l'EQM s'améliorent quand les échantillons deviennent plus petits. Cependant, cela tient au fait que les estimateurs conditionnels de l'EQM deviennent alors plus instables. Pour ces très petits échantillons de domaine, l'estimateur conditionnel de l'EQM comporte un biais moins important que l'estimateur PR de l'EQM (7,7 % *vs* 10,5 %), mais il est également plus instable (REQMR de l'estimateur conditionnel de l'EQM égal à 99 % *vs* 50 % pour l'estimateur PR de l'EQM). Néanmoins, ce n'est pas le cas pour la population de l'AAGIS, avec une taille médiane $n_i = 8$. Dans ce cas, les estimateurs PR de l'EQM donnent de mauvais résultats, les estimateurs conditionnels de l'EQM étant à la fois moins biaisés et plus stables.

Tableau 5
Propriétés des estimateurs des moyennes régionales et des estimateurs de leur EQM – Population de ménages de l'Albanie

Méthode de pondération	n_i **médian = 56**		n_i **médian = 9**	
Estimateur	BR(m)	REQMR(m)	BR(m)	REQMR(m)
Régression	0,04	6,25	-0,13	16,56
EBLUP, (13)	0,42	5,90	1,62	12,42
EDFM, (18)	0,03	6,14	0,04	16,92
M-quantile, (22)	0,04	6,07	-0,05	16,60
Méthode/EQM	BR(M)	REQMR(M)	BR(M)	REQMR(M)
Régression /VReg	17,6	42	11,2	42
EBLUP/PR0	14,6	44	10,5	50
EBLUP/PR1	14,4	43	8,8	48
EBLUP/PR2	14,5	43	9,7	48
EBLUP/Conditionnel	0,1	24	7,7	99
EDFM/Conditionnel	-0,8	25	-5,5	64
M-quantile/Conditionnel	2,9	27	-2,0	75

Tableau 6
Propriétés des estimateurs des moyennes régionales et des estimateurs de leur EQM – Population de fermes de l'AAGIS

Méthode de pondération	n_i **médian = 53**		n_i **médian = 8**	
Estimateur	BR(m)	REQMR(m)	BR(m)	REQMR(m)
Régression	0,03	13,36	0,08	29,83
EBLUP, (13)	1,64	13,53	0,92	25,82
EDFM, (18)	-0,73	14,26	-1,02	37,77
M-quantile, (22)	-0,04	11,68	-0,15	32,22
Méthode/EQM	BR(M)	REQMR(M)	BR(M)	REQMR(M)
Régression /VReg	74,1	406	54,7	867
EBLUP/PR0	22,4	131	17,7	374
EBLUP/PR1	19,5	137	19,0	367
EBLUP/PR2	21,0	123	31,1	444
EBLUP/Conditionnel	5,5	132	17,8	255
EDFM/Conditionnel	-0,5	181	0,9	318
M-quantile/Conditionnel	-0,7	69	-1,9	212

L'estimateur de l'EQM de l'estimateur par la régression présente un biais modéré ou élevé pour les deux populations et tous les scénarios de simulation. Dans le cas de la population albanienne, il semble soutenir la concurrence des autres estimateurs de l'EQM en ce qui concerne la REQMR, mais dans le cas de la population de l'AAGIS, il est clairement moins stable que les autres estimateurs de l'EQM. Enfin, l'estimateur conditionnel de l'EQM de l'estimateur du M-quantile donne de bons résultats, son biais relatif étant faible et sa stabilité, bonne, pour tous les scénarios de simulation et les deux populations, sauf dans le cas de la population albanienne avec une taille médiane d'échantillon $n_i = 9$ auquel cas la REQMR est de 75 %.

L'examen des valeurs de la REQM propre au domaine présentées à la figure 2 pour la population albanienne et à la figure 3 pour la population de l'AAGIS donne une idée des raisons de ces différences de comportement. Notons que, dans les deux cas, les tailles d'échantillon sont celles tirées des enquêtes originales. Donc, à la figure 2, nous constatons que les estimateurs conditionnels de l'EQM suivent tous trois exceptionnellement bien les REQM sous le plan propre aux districts de leurs estimateurs respectifs. En revanche, l'estimateur PR0 ne semble pas capable de saisir les différences entre districts de la REQM sous le plan de l'estimateur EBLUP. À la figure 3, nous voyons que l'estimateur conditionnel de l'EQM de l'estimateur du M-quantile donne de très bons résultats dans toutes les régions, tandis que l'estimateur correspondant de l'EQM de l'estimateur EDFM donne aussi de bons résultats dans toutes les régions sauf une (région 6), dans laquelle il surestime considérablement la REQM sous le plan de ce prédicteur. Cette région mérite d'être mentionnée, parce que les échantillons qui sont déséquilibrés en ce qui a trait à la superficie dans cette région produisent des poids négatifs sous le modèle mixte linéaire supposé. Le tableau se complique lorsque l'on considère les propriétés de l'estimation de la REQM propre à la région pour l'EBLUP à la figure 3. Ici, nous voyons clairement que l'estimateur conditionnel de l'EQM de l'EBLUP suit mieux la REQM sous le plan propre à la région de ce prédicteur que l'estimateur PR0 de l'EQM. Sauf dans le cas de la région 6 (pour laquelle l'équilibre de l'échantillon pose problème), la variation régionale de la valeur de l'estimateur PR0 de la REQM de l'EBLUP semble faible, ce qui indique un problème de biais sérieux.

Comme nous l'avons mentionné plus haut, il n'est pas rare de vouloir produire une estimation pour un petit domaine pour lequel il n'existe pas d'échantillon. Le cas échéant, on doit s'appuyer entièrement sur l'exactitude de la spécification du modèle. Au tableau 7, nous illustrons

l'importance de cette hypothèse en comparant les propriétés de l'estimation de l'EBLUP et de l'estimation de son EQM pour les domaines échantillonnés à celles observées pour l'EBLUB synthétique pour les domaines pour lesquels aucune donnée d'échantillon n'est disponible. Nous présentons deux situations. La première est une modification de la simulation SIM1-A fondée sur un modèle avec une petite taille moyenne d'échantillon et avec cinq domaines dont l'échantillon est nul. La deuxième est une modification similaire, en prenant une petite taille d'échantillon, de la simulation fondée sur un plan de sondage appliquée à la population de l'AAGIS, avec quatre domaines dont l'échantillon est nul. Il est évident que, quand le modèle qui sous-tend l'EBLUP est effectivement vérifié (c'est-à-dire SIM1-A), l'estimation ainsi que l'estimation de son EQM (fondée sur le modèle PR0 ou sur l'option conditionnelle) donne de bons résultats. Le problème tient au fait que, s'il existe un doute quant à la façon dont ce modèle tient (comme dans la population de l'AAGIS), l'EBLUP peut échouer et notre estimateur de son EQM peut aussi ne pas arriver à déceler ce problème. Cela est bien illustré par les résultats pour la population de l'AAGIS présentés au tableau 7, où nous voyons que l'estimateur PR0 ainsi que l'estimateur conditionnel de l'EQM pour l'EBLUP synthétique n'arrivent pas du tout à déceler l'important biais positif de ce dernier et présentent donc un grand biais vers le bas.

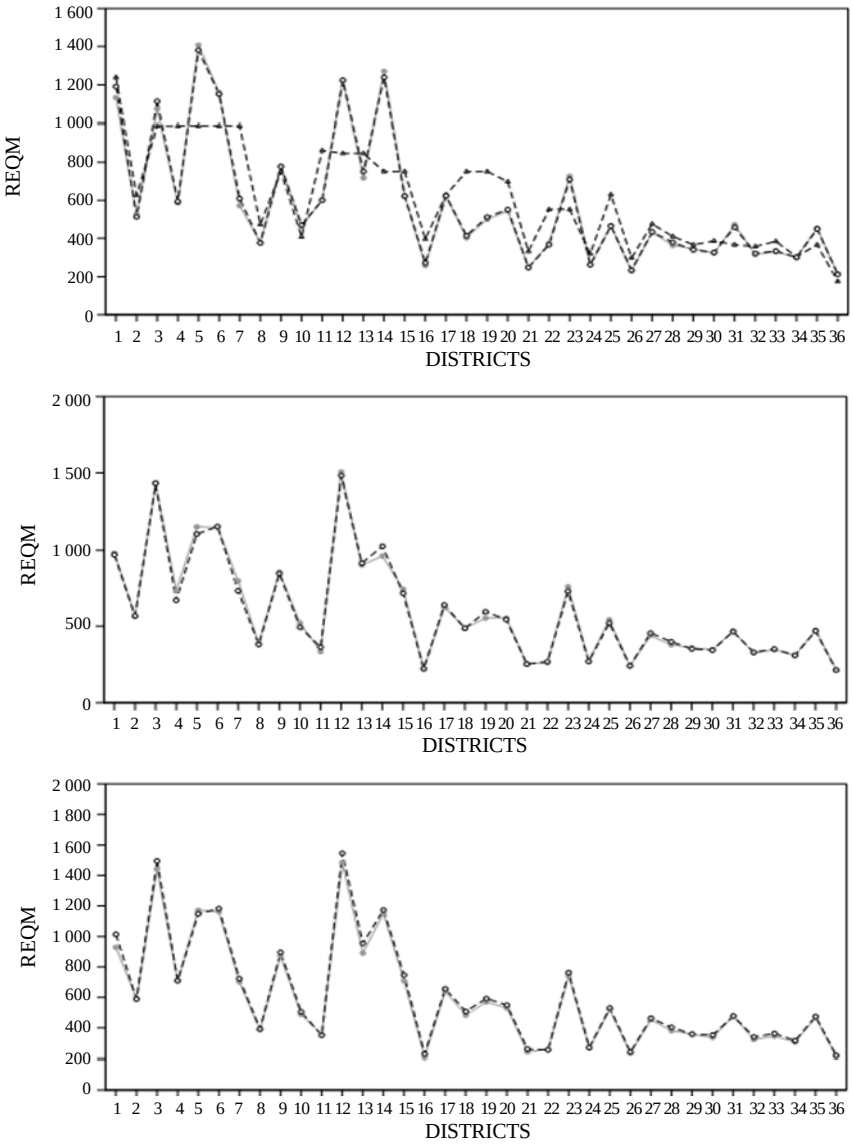

Figure 2 Valeurs au niveau du district de la REQM sous le plan réel (trait plein) et de la REQM estimée moyenne (trait interrompu) obtenues dans les simulations fondées sur le plan en utilisant la population de ménages de l'Albanie. Les districts sont classés par ordre de taille de population croissante. Les valeurs pour l'estimateur PR0 sont indiquées par Δ, tandis que celles pour l'estimateur conditionnel sont indiquées par o. Les graphiques montrent les résultats pour les estimateurs EBLUP (en haut), EDFM (au centre) et M-quantile (en bas)

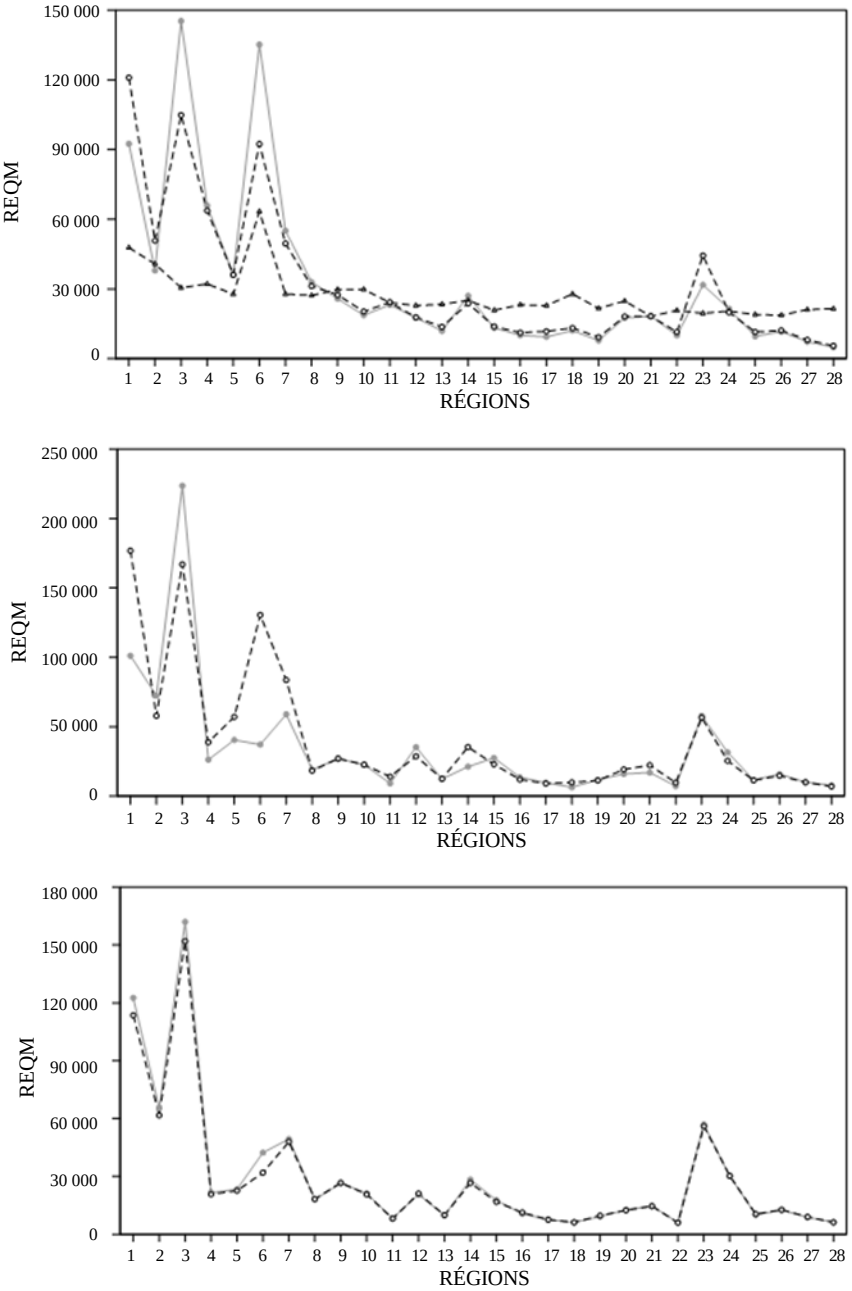

Figure 3 Valeurs au niveau de la région de la REQM sous le plan réel (trait plein) et de la REQM estimée moyenne (trait interrompu) obtenues dans les simulations fondées sur le plan en utilisant la population de fermes de l'AAGIS. Les régions sont classées par ordre de taille de population croissante. Les valeurs pour l'estimateur PR0 sont indiquées par Δ, tandis que celles pour l'estimateur conditionnel sont indiquées par o. Les graphiques montrent les résultats pour les estimateurs EBLUP (en haut), EDFM (au centre) et M-quantile (en bas)

Tableau 7
Propriétés des estimateurs EBLUP et de l'EQM en présence de domaines dont l'échantillon est nul

	Méthode de pondération/Estimateur	SIM1-A, n_i médian = 10		AAGIS, n_i médian = 9	
		BR(*m*)	REQMR(*m*)	BR(*m*)	REQMR(*m*)
Domaines avec $n_i > 0$	(13)/EBLUP	0,00	0,52	2,29	24,94
Domaines avec $n_i = 0$	(23)/EBLUP synthétique	-0,05	1,25	87,45	96,46
	EQM Estimateur	BR(*M*)	REQMR(*M*)	BR(*M*)	REQMR(*M*)
Domaines avec $n_i > 0$	(13)/PR0	0,5	11	29,91	760
	(13)/Conditionnel	0,7	50	23,87	298
Domaines avec $n_i = 0$	(23)/PR0	-1,8	35	-29,07	601
	(23)/Conditionnel	-3,6	34	-31,45	101

4. Conclusion et discussion

Dans le présent article, nous proposons une méthode robuste au biais et facile à appliquer d'estimation de l'EQM conditionnelle des estimateurs pseudo-linéaires de moyennes (et de totaux) de petits domaines. Nos résultats empiriques montrent que cette méthode d'estimation de l'EQM donne des résultats raisonnablement bons en ce qui concerne le biais quand elle est utilisée pour estimer l'EQM sous un modèle ainsi que sous un plan de sondage des trois estimateurs pseudo-linéaires assez différents pris en considération dans la présente étude. Cependant, ces résultats améliorés pour ce qui est du biais ont pour prix une augmentation de la variabilité. En particulier, quand les tailles d'échantillon de domaine sont très petites, nous ne recommandons pas d'appliquer la méthode d'estimation de l'EQM que nous proposons à un estimateur conditionnellement biaisé tel que l'EBLUP.

L'usage de l'EBLUP est très répandu pour l'estimation sur petits domaines et, dans ce contexte, l'estimateur de l'EQM dépendant d'un modèle PR0 proposé pour l'EBLUP par Prasad et Rao (1990) est sans biais quand les hypothèses qui sous-tendent ce modèle sont valides (SIM1-A/B et SIM2-A/B dans nos simulations fondées sur un modèle), mais il contient un biais en présence d'effets de domaine aberrants (SIM3-A/A* et SIM3-B/B*). Il s'agit aussi de l'estimateur de l'EQM le plus stable dans les simulations fondées sur un modèle. Toutefois, étant donné sa construction fondée sur la moyenne de domaines, il ne détecte pas l'EQM propre au domaine de l'EBLUP dans nos deux simulations fondées sur un plan de sondage, où nous pouvons seulement considérer que le modèle linéaire mixte supposé n'est qu'approximativement correct. Cela donne à penser que la méthode d'estimation conditionnelle de l'EQM que nous proposons devrait être envisagée comme une alternative au modèle PR0 dans les situations où il existe un doute quant à l'exactitude de la spécification du modèle mixte linéaire au niveau du petit domaine ou dans celles où les tailles d'échantillon de domaine ne sont pas faibles. Les résultats empiriques présentés ici permettent de se faire une idée de ce qui constitue une petite taille d'échantillon.

Si l'utilisateur doute de la validité du modèle mixte linéaire supposé, il pourrait envisager l'estimation fondée sur un modèle de rechange d'application plus générale, par exemple le modèle du M-quantile, ou remplacer l'EBLUP par une option plus robuste aux valeurs aberrantes (Sinha et Rao 2009). Dans le premier cas, l'approche que nous proposons ici est, à l'heure actuelle, la seule approche analytique de l'estimation de l'EQM, tandis que dans le second cas, l'approche proposée fournit une option analytique pour remplacer les méthodes bootstrap d'estimation de l'EQM demandant beaucoup de calculs. Notons toutefois que, pour des tailles d'échantillon de domaine très petites, l'estimateur de l'EQM robuste au biais proposé dans le présent article reste instable.

Une future piste de recherche pourrait consister à comparer la méthode analytique d'estimation de l'EQM proposée ici aux estimateurs de l'EQM fondés sur le bootstrap, par exemple l'estimateur bootstrap non paramétrique de l'EQM de l'estimateur du M-quantile proposé par Tzavidis, Marchetti et Chambers (2010), et l'estimateur bootstrap de l'EQM pour l'estimateur EBLUP robuste proposé par Sinha et Rao (2009). Une question clé dans cette étude sera de déterminer si d'autres options d'estimateur bootstrap de l'EQM sont plus stables, surtout pour les petites tailles d'échantillon de domaine.

L'extension de l'approche d'estimation conditionnelle de l'EQM aux situations non linéaires d'estimation sur petits domaines reste à faire. Cependant, puisque cette approche est étroitement reliée à l'estimation robuste de l'EQM en population fondée sur la linéarisation par développement en série de Taylor (ainsi que l'estimation jackknife de l'EQM, voir Valliant, Dorfman et Royall 2000, section 5.4.2), il devrait être possible de développer des extensions appropriées pour les méthodes d'estimation non linéaires sur petits domaines correspondantes. Bien que nous ne présentions pas les résultats pertinents ici, nous voyons pour preuve le fait que la méthode d'estimation conditionnelle de l'EQM décrite dans le présent article a déjà été utilisée pour estimer l'EQM de l'estimateur EDFM quand il est appliqué à des variables qui ne se prêtent pas à la modélisation linéaire mixte, c'est-à-dire celles pour lesquelles la proportion de valeurs nulles est élevée (Chandra et Chambers 2009), et à des variables catégoriques (Chandra, Chambers et Salvati 2011). Plus récemment, l'approche a également été utilisée pour estimer l'EQM des estimateurs sur petits domaines des M-quantile géographiquement pondérés dans des situations où les valeurs de petit domaine sont spatialement corrélées (Salvati, Tzavidis, Pratesi et Chambers 2011). Il a également été employé par Salvati, Chandra, Ranalli et Chambers (2010) pour estimer l'EQM d'estimateurs sur petits domaines fondés sur un modèle de petit domaine non paramétrique (Opsomer, Claeskens, Ranalli, Kauermann et Breidt 2008).

Comme le montre clairement le présent exposé, notre approche préférée d'estimation de l'EQM repose sur l'hypothèse que l'EQM qui nous intéresse vraiment est celle définie par le modèle propre au domaine (1). Cette approche diffère de celle habituellement suivie pour définir l'EQM dans l'estimation sur petits domaines, c'est-à-dire en adoptant un concept de moyenne de l'EQM sur les domaines comme mesure appropriée de l'exactitude d'un estimateur sur petits domaines. Comme l'a fait remarquer Longford

(2007), le but ultime de l'estimation sur petits domaines est de faire des inférences au sujet des caractéristiques des petits domaines conditionnellement aux valeurs réalisées (mais inconnues) des effets de petit domaine, c'est-à-dire par rapport à (1). On peut considérer cela comme étant un objectif fondé sur le plan (comme dans Longford 2007) ou, comme nous le préférons, un objectif fondé sur un modèle qui ne rentre pas vraiment dans le cadre habituel des effets aléatoires pour l'estimation sur petits domaines. Dans l'un et l'autre cas, nous nous intéressons à la variabilité par rapport à des valeurs prévues fixes propres au domaine, ce qui est conforme au concept de variabilité qui est habituellement appliqué en inférence au niveau de la population.

Remerciements

Les auteurs remercient le rédacteur en chef, le rédacteur associé et deux examinateurs de leurs commentaires et suggestions précieux qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

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

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

Chandra, H., et Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistics*, 25(3), 379-395.

Chandra, H., Chambers, R. et Salvati, N. (2011). Small area estimation of proportions in business surveys. Paraîtra dans le *Journal of Statistical Computation and Simulation*.

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

Longford, N.T. (2007). De l'erreur-type des estimateurs pour petits domaines fondés sur un modèle. *Techniques d'enquête*, 33, 81-92.

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

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

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

Salvati, N., Chandra, H., Ranalli, M.G. et Chambers, R. (2010). Small area estimation using a nonparametric model based direct estimator. *Computational Statistics and Data Analysis*, 54, 2159-2171.

Salvati, N., Tzavidis, N., Pratesi, M. et Chambers, R. (2011). Small area estimation via M-quantile geographically weighted regression. À venir dans *TEST*, DOI 10.1007/s11749-010-0231-1.

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

Street, J.O, Carroll, R.J. et Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42, 152-154.

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

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York : John Wiley & Sons, Inc.