

## Techniques d'enquête 41-1

# Estimation sur petits domaines sous un modèle de Fay-Herriot avec test préliminaire pour la présence d'effets aléatoires de domaine

par Isabel Molina, J.N.K. Rao et Gauri Sankar Datta

Date de diffusion : le 29 juin 2015



Statistique  
Canada

Statistics  
Canada

Canada

---

## Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à [www.statcan.gc.ca](http://www.statcan.gc.ca).

Vous pouvez également communiquer avec nous par :

**Courriel** à [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)

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

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

**Programme des services de dépôt**

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site [www.statcan.gc.ca](http://www.statcan.gc.ca) sous « À propos de nous » > « Notre organisme » > « [Offrir des services aux Canadiens](#) »

## Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

## Signes conventionnels dans les tableaux

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

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

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2015

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'[entente de licence ouverte](#) de Statistique Canada.

Une [version HTML](#) est aussi disponible.

*This publication is also available in English.*

---

# Estimation sur petits domaines sous un modèle de Fay-Herriot avec test préliminaire pour la présence d'effets aléatoires de domaine

Isabel Molina, J.N.K. Rao et Gauri Sankar Datta<sup>1</sup>

## Résumé

Le modèle de Fay-Herriot est un modèle au niveau du domaine d'usage très répandu pour l'estimation des moyennes de petit domaine. Ce modèle contient des effets aléatoires en dehors de la régression linéaire (fixe) basée sur les covariables au niveau du domaine. Les meilleurs prédicteurs linéaires sans biais empiriques des moyennes de petit domaine s'obtiennent en estimant les effets aléatoires de domaine, et ils peuvent être exprimés sous forme d'une moyenne pondérée des estimateurs directs propres aux domaines et d'estimateurs synthétiques de type régression. Dans certains cas, les données observées n'appuient pas l'inclusion des effets aléatoires de domaine dans le modèle. L'exclusion de ces effets de domaine aboutit à l'estimateur synthétique de type régression, autrement dit un poids nul est appliqué à l'estimateur direct. L'étude porte sur un estimateur à test préliminaire d'une moyenne de petit domaine obtenu après l'exécution d'un test pour déceler la présence d'effets aléatoires de domaine. Parallèlement, elle porte sur les meilleurs prédicteurs linéaires sans biais empiriques des moyennes de petit domaine qui donnent toujours des poids non nuls aux estimateurs directs dans tous les domaines, ainsi que certains estimateurs de rechange basés sur le test préliminaire. La procédure de test préliminaire est également utilisée pour définir de nouveaux estimateurs de l'erreur quadratique moyenne des estimateurs ponctuels des moyennes de petit domaine. Les résultats d'une étude par simulation limitée montrent que, si le nombre de domaines est petit, la procédure d'essai préliminaire mène à des estimateurs de l'erreur quadratique moyenne présentant un biais relatif absolu moyen considérablement plus faible que les estimateurs de l'erreur quadratique moyenne usuels, surtout quand la variance des effets aléatoires est faible comparativement aux variances d'échantillonnage.

**Mots-clés :** Modèle au niveau du domaine; meilleur prédicteur linéaire sans biais empirique; erreur quadratique moyenne; test préliminaire; estimation sur petits domaines.

## 1 Introduction

Un modèle de base au niveau du domaine, appelé modèle de Fay-Herriot (FH), est souvent utilisé pour obtenir des estimateurs efficaces des moyennes de domaine quand les tailles d'échantillon dans les domaines sont petites. Ce modèle comprend des effets aléatoires de domaine non observables, et le meilleur prédicteur linéaire sans biais empirique (EBLUP pour *empirical best linear unbiased predictor*) d'une moyenne de petit domaine s'obtient en estimant l'effet aléatoire associé. L'EBLUP est une combinaison pondérée d'un estimateur direct propre au domaine et d'un estimateur synthétique de type régression qui utilise toutes les données. Un estimateur de l'erreur quadratique moyenne (EQM) de l'EBLUP a été obtenu pour la première fois par Prasad et Rao (1990) en utilisant un estimateur par la méthode des moments de la variance des effets aléatoires, et plus tard par Datta et Lahiri (2000) pour l'estimateur du maximum de vraisemblance restreint (MVRE) de la variance. Rao (2003, Chapitre 7) donne un compte rendu détaillé des EBLUP et des estimateurs de leur EQM pour les modèles FH.

Parfois, les données observées n'appuient pas l'inclusion des effets de domaine dans le modèle. L'exclusion de ces effets mène à l'estimateur synthétique de type régression. Partant de cette idée, Datta,

1. Isabel Molina, Département de statistiques, Université Carlos III de Madrid, C/Madrid 126, 28903 Getafe (Madrid), Espagne et Instituto de Ciencias Matemáticas (ICMAT), (Madrid), Espagne. Courriel : isabel.molina@uc3m.es; J.N.K. Rao, École de mathématiques et de statistiques, Université Carleton, Ottawa, Canada; Gauri Sankar Datta, Département de statistiques, Université de Géorgie, Athens, États-Unis.

Hall et Mandal (2011) ont proposé d'effectuer un test préliminaire pour déterminer la présence d'effets aléatoires de domaine à un seuil de signification spécifié, et de définir l'estimateur pour petits domaines en fonction du résultat du test. Si l'hypothèse nulle de l'absence d'effets aléatoires de domaine n'est pas rejetée, le modèle sans effets de domaine est pris en considération pour estimer les moyennes de petit domaine, c'est-à-dire que l'estimateur synthétique de type régression est utilisé. Si l'hypothèse nulle est rejetée, l'EBLUP habituel sous le modèle FH avec effets de domaine est utilisé. Datta et coll. (2011) ont remarqué que l'estimateur à test préliminaire (ETP) pouvait aboutir à des gains d'efficacité importants comparativement à l'EBLUP, particulièrement quand le nombre de petits domaines est modéré. En guise de test préliminaire, ils ont considéré un test fondé sur la normalité, ainsi qu'un test de type bootstrap qui permet d'éviter l'hypothèse de normalité.

Quand la variance estimée des effets de domaine est nulle, l'EBLUP devient automatiquement l'estimateur synthétique de type régression. Cependant, l'EQM estimée obtenue par Prasad et Rao (1990) ou par Datta et Lahiri (2000) ne se réduit pas à l'EQM estimée de l'estimateur synthétique de type régression. Donc, les estimateurs habituels de l'EQM sont biaisés pour une faible variance des effets aléatoires. Par conséquent, nous proposons des estimateurs de l'EQM de l'EBLUP basés sur la procédure de test préliminaire (TP). Si le test indique que la variance des effets aléatoires n'est pas importante, nous considérons l'estimateur de l'EQM de l'estimateur synthétique. Sinon, nous considérons les estimateurs habituels de l'EQM de l'EBLUP.

L'EBLUP applique un poids nul aux estimations directes pour tous les domaines quand la variance estimée des effets de domaine est nulle. Par ailleurs, les praticiens des sondages préfèrent souvent appliquer un poids strictement positif aux estimations directes, parce que ces dernières utilisent les données au niveau de l'unité propres au domaine disponibles et intègrent aussi le plan de sondage. Li et Lahiri (2010) ont présenté un estimateur du maximum de vraisemblance ajusté (MVA) de la variance des effets aléatoires qui est toujours positif et, par conséquent, mène à des EBLUP donnant des poids strictement positifs aux estimateurs directs. Comme nous le verrons, un biais est payé sous forme de biais lorsqu'on utilise l'EBLUP basé sur l'estimateur MVA. Nous proposons ici d'autres options d'estimateurs pour petits domaines qui donnent toujours un poids positif aux estimateurs directs, mais avec un biais plus faible.

Dans le présent article, nous étudions empiriquement les propriétés des estimateurs ETP des moyennes de petit domaine comparativement aux EBLUP habituels et à d'autres estimateurs proposés. En particulier, nous étudions le choix du seuil de signification pour les estimations de domaine et pour les estimations de l'EQM basées sur le test préliminaire (TP). Les EBLUP basés sur l'estimateur MVA de la variance des effets aléatoires de Li et Lahiri (2010), qui donnent des poids non nuls aux estimateurs directs dans tous les domaines, sont également étudiés et comparés aux versions TP de l'estimateur MVA (TP-MVA). Différents estimateurs de l'EQM de ces estimateurs ETP-MVA sont également étudiés en ce qui a trait au biais relatif. En nous fondant sur les résultats de simulation, nous recommandons les EBLUP et les estimateurs de l'EQM qui ont de bonnes propriétés. Enfin, nous examinons la couverture et la longueur des intervalles de prédiction fondés sur l'hypothèse de normalité, obtenus en utilisant les EBLUP et les estimateurs de l'EQM associés.

La présentation de l'article est la suivante. À la section 2, nous décrivons le modèle FH et les EBLUP des moyennes de petit domaine. À la section 3, nous commentons l'estimation de l'EQM. À la section 4, nous présentons les estimateurs ETP des moyennes de petit domaine et les estimateurs de l'EQM basés sur

le TP. À la section 5, nous décrivons les estimateurs pour petits domaines et les estimateurs de l'EQM associés sous estimation MVA de la variance des effets de domaine. À la section 6, nous présentons des estimateurs de rechange qui appliquent également des poids positifs aux estimateurs directs, ainsi que les estimateurs de l'EQM proposés. À la section 7, nous présentons les résultats de l'étude par simulation. Enfin, à la section 8, nous tirons certaines conclusions.

## 2 Estimation des moyennes de petit domaine

Considérons une population partitionnée en  $m$  domaines et soit  $\theta_i$  la moyenne de la variable d'intérêt pour le domaine  $i, i = 1, \dots, m$ . Nous supposons qu'un échantillon est tiré indépendamment dans chaque domaine. Soit  $y_i$  un estimateur direct sans biais sous le plan de sondage de  $\theta_i$  obtenu en utilisant des données d'enquête provenant du domaine échantillonné  $i$ . Les estimateurs directs sont très inefficaces pour les domaines dont l'échantillon est de petite taille. Nous étudions l'estimation sur petits domaines sous un modèle au niveau du domaine, dans lequel les valeurs des covariables au niveau du domaine sont disponibles pour tous les domaines. Le modèle fondamental de ce type est le modèle de Fay-Herriot, introduit par Fay et Herriot (1979) pour estimer le revenu par habitant dans de petites localités aux États-Unis. Ce modèle comprend deux parties. La première repose sur l'hypothèse que les estimateurs directs,  $y_i$ , des moyennes de petit domaine,  $\theta_i$ , sont sans biais sous le plan de sondage, et satisfont

$$y_i = \theta_i + e_i, \quad e_i \stackrel{\text{ind}}{\sim} N(0, D_i), \quad i = 1, \dots, m. \quad (2.1)$$

Ici, la variance d'échantillonnage  $D_i = \text{Var}(y_i | \theta_i)$  est supposée connue pour tous les domaines  $i = 1, \dots, m$ . En pratique, les valeurs de  $D_i$  sont déterminées à partir de sources externes ou en lissant les variances d'échantillonnage estimées en utilisant une méthode à fonction de variance généralisée (Fay et Herriot 1979).

Dans la deuxième partie du modèle de Fay-Herriot,  $\theta_i$  est traité comme étant aléatoire et l'on suppose qu'un vecteur  $p$  de covariables au niveau du domaine,  $\mathbf{x}_i$ , relié linéairement à  $\theta_i$ , est disponible pour chaque domaine  $i$ , c'est-à-dire

$$\theta_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i, \quad v_i \stackrel{\text{iid}}{\sim} N(0, A), \quad i = 1, \dots, m, \quad (2.2)$$

où  $v_i$  est l'effet aléatoire du domaine  $i$ , supposé indépendant de  $e_i$ , et  $A \geq 0$  est la variance des effets aléatoires. Observons que, marginalement,

$$y_i \stackrel{\text{ind}}{\sim} N(\mathbf{x}_i' \boldsymbol{\beta}, D_i + A), \quad i = 1, \dots, m. \quad (2.3)$$

En posant que  $\mathbf{y} = (y_1, \dots, y_m)'$ ,  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)'$  et  $\mathbf{D} = \text{diag}(D_1, \dots, D_m)$ , le modèle (2.3) peut être exprimé en notation matricielle sous la forme  $\mathbf{y} \sim N\{\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(A)\}$  avec  $\boldsymbol{\Sigma}(A) = \mathbf{D} + A\mathbf{I}_m$ , où  $\mathbf{I}_m$  désigne la matrice identité de dimensions  $m \times m$ . Si  $A$  est connue, le meilleur prédicteur linéaire sans biais (BLUP) par composante de  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$  est donné par

$$\tilde{\boldsymbol{\theta}}(A) = (\tilde{\theta}_1(A), \dots, \tilde{\theta}_m(A))' = \mathbf{X}\tilde{\boldsymbol{\beta}}(A) + A\boldsymbol{\Sigma}^{-1}(A)\{\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}(A)\}, \quad (2.4)$$

où

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(A) &= \{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\mathbf{X}\}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\mathbf{y} \\ &= \left\{ \sum_{i=1}^m (A + D_i)^{-1} \mathbf{x}_i \mathbf{x}_i' \right\}^{-1} \sum_{i=1}^m (A + D_i)^{-1} \mathbf{x}_i y_i \end{aligned} \quad (2.5)$$

est l'estimateur des moindres carrés pondérés (MCP) de  $\boldsymbol{\beta}$ . Toutefois, en pratique,  $A$  est inconnue. En substituant un estimateur convergent  $\hat{A}$  à  $A$  dans le BLUP (2.4), nous obtenons l'EBLUP donné par

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_m)' = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{A}\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \quad (2.6)$$

où  $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{A})$  et  $\hat{\boldsymbol{\Sigma}} = \mathbf{D} + \hat{A}\mathbf{I}_m$ . Pour le  $i^{\text{e}}$  domaine, l'EBLUP de  $\theta_i$  peut être exprimé comme une combinaison linéaire convexe de l'estimateur synthétique de type régression  $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$  et de l'estimateur direct  $y_i$ , sous la forme

$$\hat{\theta}_i = B_i(\hat{A})\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \{1 - B_i(\hat{A})\}y_i, \quad (2.7)$$

où le poids appliqué à l'estimateur synthétique de type régression  $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$  est donné par  $B_i(\hat{A})$ , où  $B_i(A) = D_i/(A + D_i)$ . Notons que le poids augmente avec la variance d'échantillonnage  $D_i$ . Donc, quand l'estimateur direct n'est pas fiable, c'est-à-dire que  $D_i$  est grande comparativement à la variance totale  $\hat{A} + D_i$ , un poids plus important est appliqué à l'estimateur synthétique de type régression  $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ . Par ailleurs, si l'estimateur direct est efficace,  $D_i$  est petite comparativement à  $\hat{A} + D_i$ , et un plus grand poids est alors donné à l'estimateur direct  $y_i$ .

Plusieurs estimateurs de  $A$  ont été proposés dans la littérature, y compris des estimateurs des moments sans hypothèse de normalité, l'estimateur du maximum de vraisemblance (MV) et l'estimateur du maximum de vraisemblance restreint (ou résiduel) (MVRE). L'estimateur MV de  $A$  est  $\hat{A}_{\text{MV}} = \max(0, \hat{A}_{\text{MV}}^*)$ , où  $\hat{A}_{\text{MV}}^*$  peut être obtenu en maximisant la fonction de vraisemblance profil donnée par

$$L_p(A) = c|\boldsymbol{\Sigma}(A)|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}(A)\mathbf{y}\right\},$$

où  $c$  désigne une constante générique et

$$\mathbf{P}(A) = \boldsymbol{\Sigma}^{-1}(A) - \boldsymbol{\Sigma}^{-1}(A)\mathbf{X}\{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\mathbf{X}\}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A).$$

L'estimateur MVRE de  $A$  est  $\hat{A}_{\text{RE}} = \max(0, \hat{A}_{\text{RE}}^*)$ , où  $\hat{A}_{\text{RE}}^*$  s'obtient en maximisant la vraisemblance restreinte/résiduelle, donnée par

$$L_{\text{RE}}(A) = c |\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\mathbf{X}|^{-1/2} |\boldsymbol{\Sigma}(A)|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}(A)\mathbf{y}\right\}.$$

Dans le présent article, nous nous concentrons sur l'estimateur MVRE  $\hat{A}_{\text{RE}}$  qui est fréquemment utilisé en pratique, et nous désignons par  $\hat{\boldsymbol{\theta}}_{\text{RE}} = (\hat{\theta}_{\text{RE},1}, \dots, \hat{\theta}_{\text{RE},m})'$  l'EBLUP donné en (2.6) obtenu avec  $\hat{A} = \hat{A}_{\text{RE}}$ .

### 3 Erreur quadratique moyenne

Notons que le BLUP  $\tilde{\theta}_i(A)$  de la moyenne de petit domaine  $\theta_i$  est une fonction linéaire de  $\mathbf{y}$ . Donc, son EQM peut être calculée facilement et est donnée par la somme de deux termes :

$$\text{EQM}\{\tilde{\theta}_i(A)\} = g_{1i}(A) + g_{2i}(A),$$

où  $g_{1i}(A)$  est dû à l'estimation de l'effet aléatoire de domaine  $v_i$  et  $g_{2i}(A)$  est dû à l'estimation du paramètre de régression  $\boldsymbol{\beta}$ , avec

$$\begin{aligned} g_{1i}(A) &= D_i \{1 - B_i(A)\}, \\ g_{2i}(A) &= B_i^2(A) \mathbf{x}_i' \{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\mathbf{X}\}^{-1} \mathbf{x}_i. \end{aligned}$$

Cependant, l'EBLUP  $\hat{\theta}_i$  donné en (2.7) n'est pas linéaire en  $\mathbf{y}$  en raison de l'estimation de la variance des effets aléatoires  $A$ . En utilisant un estimateur des moments de  $A$ , Prasad et Rao (1990) ont obtenu une approximation d'ordre deux correcte de l'EQM de l'EBLUP. Plus tard, Datta et Lahiri (2000) et Das, Jiang et Rao (2004) ont obtenu une approximation d'ordre deux correcte de l'EQM sous estimation du MV et du MVRE de  $A$ . En utilisant l'estimateur MVRE de  $A$ , leur approximation de l'EQM, pour une grande valeur de  $m$ , est donnée par

$$\text{EQM}(\hat{\theta}_{\text{RE},i}) = g_{1i}(A) + g_{2i}(A) + g_{3i}(A) + o(m^{-1}), \quad (3.1)$$

où

$$g_{3i}(A) = B_i^2(A) \frac{V_{\text{RE}}(A)}{A + D_i} \quad \text{et} \quad V_{\text{RE}}(A) = \frac{2}{\sum_{i=1}^m (A + D_i)^{-2}}.$$

Notons que, quand  $m \rightarrow \infty$ ,  $g_{1i}(A) = O(1)$ ,  $g_{2i}(A) = O(m^{-1})$  et  $g_{3i}(A) = O(m^{-1})$ , de sorte que  $g_{1i}(A)$  est le terme principal dans l'EQM quand  $m$  est grand. Cependant, si  $A$  est petite, le terme  $g_{1i}(A)$  est approximativement nul et  $g_{3i}(A)$  pourrait alors devenir le terme principal quand  $m$  est petit. Par exemple, en ne prenant qu'une seule covariable ( $p = 1$ ) avec des valeurs constantes  $x_i = 1$  et des variances d'échantillonnage constantes  $D_i = D, i = 1, \dots, m$ , et en posant que  $A = 0$ , nous obtenons  $g_{1i}(0) = 0$ ,  $g_{2i}(0) = D/m$  et  $g_{3i}(0) = 2D/m$ ; autrement dit,  $g_{3i}(0)$  est deux fois plus grand que  $g_{2i}(0)$ .

Datta et Lahiri (2000) ont obtenu un estimateur de l'EQM de l'EBLUP  $\hat{\theta}_{\text{RE},i}$  donné par

$$\text{eqm}(\hat{\theta}_{\text{RE},i}) = g_{1i}(\hat{A}_{\text{RE}}) + g_{2i}(\hat{A}_{\text{RE}}) + 2g_{3i}(\hat{A}_{\text{RE}}). \quad (3.2)$$

L'estimateur de l'EQM (3.2) est sans biais d'ordre deux en ce sens que

$$E\{\text{eqm}(\hat{\theta}_{\text{RE},i})\} = \text{EQM}(\hat{\theta}_{\text{RE},i}) + o(m^{-1}).$$

Dans le cas où  $A = 0$ , le BLUP  $\tilde{\theta}_{\text{RE},i}$  de  $\theta_i$  devient l'estimateur synthétique de type régression  $\hat{\theta}_{\text{SYN},i} = \mathbf{x}'_i \tilde{\boldsymbol{\beta}}(0)$ . Mais étonnamment, l'approximation de l'EQM de l'EBLUP donnée en (3.1) peut être très différente de l'EQM de l'estimateur synthétique. Notons que cette dernière est donnée par

$$\text{EQM}(\hat{\theta}_{\text{SYN},i}) = g_{2i}(0) < g_{2i}(0) + g_{3i}(0),$$

parce que le terme  $g_{3i}(0)$  est strictement positif, même pour  $A = 0$ . En fait, dans l'exemple simple d'une seule covariable ( $p = 1$ ) avec valeurs constantes  $x_i = 1$  et variances d'échantillonnage constantes  $D_i = D, i = 1, \dots, m$ , nous avons  $\text{EQM}(\hat{\theta}_{\text{SYN},i}) = g_{2i}(0) = D/m$ , tandis que l'approximation de l'EQM de l'EBLUP donnée en (3.1) avec  $A = 0$  donne  $\text{EQM}(\hat{\theta}_{\text{RE},i}) \approx g_{2i}(0) + g_{3i}(0) = 3D/m$ , qui est trois fois plus grande. Il se fait que (3.1) n'est pas une bonne approximation de l'EQM de l'EBLUP quand  $A = 0$  et, nous devrions plutôt utiliser  $\text{EQM}(\hat{\theta}_{\text{RE},i}) = g_{2i}(0)$ . En outre, puisque pour  $A = 0$ , cette quantité ne dépend d'aucun paramètre inconnu, nous pouvons aussi la prendre comme estimateur de l'EQM, c'est-à-dire que nous pouvons prendre  $\text{eqm}(\hat{\theta}_{\text{RE},i}) = g_{2i}(0)$ .

En pratique, la vraie valeur de  $A$  est inconnue, mais nous avons un estimateur convergent  $\hat{A}_{\text{RE}}$ . Quand  $\hat{A}_{\text{RE}} = 0$ , l'EBLUP devient l'estimateur synthétique de type régression pour tous les domaines, c'est-à-dire

$$\hat{\theta}_{\text{RE},i} = \hat{\theta}_{\text{SYN},i} = \mathbf{x}'_i \tilde{\boldsymbol{\beta}}(0), i = 1, \dots, m.$$

Dans ce cas,  $g_{1i}(\hat{A}_{\text{RE}}) = 0$  pour tous les domaines et l'estimateur de l'EQM donné en (3.2) se réduit à

$$\text{eqm}(\hat{\theta}_{\text{RE},i}) = g_{2i}(0) + 2g_{3i}(0) > g_{2i}(0) = \text{EQM}(\hat{\theta}_{\text{SYN},i}), i = 1, \dots, m.$$

Donc, l'estimateur de l'EQM donné en (3.2) peut gravement surestimer l'EQM pour  $\hat{A}_{\text{RE}} = 0$ . Afin de réduire la surestimation, nous considérons un estimateur modifié de l'EQM de  $\hat{\theta}_{\text{RE},i}$  donné par

$$\text{eqm}_0(\hat{\theta}_{\text{RE},i}) = \begin{cases} g_{2i} & \text{si } \hat{A}_{\text{RE}} = 0, \\ g_{1i}(\hat{A}_{\text{RE}}) + g_{2i}(\hat{A}_{\text{RE}}) + 2g_{3i}(\hat{A}_{\text{RE}}) & \text{si } \hat{A}_{\text{RE}} > 0, \end{cases} \quad (3.3)$$

où  $g_{2i} = g_{2i}(0) = \mathbf{x}'_i (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} \mathbf{x}_i, i = 1, \dots, m$ .

En fait, pour une valeur de  $A$  proche de zéro, il se peut que  $g_{2i}$  soit plus proche de la vraie EQM que l'estimateur de l'EQM complet  $\text{eqm}(\hat{\theta}_{\text{RE},i})$ , mais la question qui se pose est celle de savoir quand  $A$  est suffisamment proche de zéro. Cette question motive le recours à une procédure de test préliminaire de  $A = 0$  pour définir des estimateurs de rechange de l'EQM de l'EBLUP à la section 4.

## 4 Estimateurs à test préliminaire

L'estimateur de  $A$  utilisé dans l'EBLUP de  $\theta_i$  introduit une incertitude qui pourrait ne pas être négligeable quand  $m$  est petit. En effet, dans l'estimateur (3.2) de l'EQM, le terme  $g_{3i}$  découle de l'estimation de  $A$ . Cependant, quand la valeur de  $A$  est suffisamment faible par rapport aux variances d'échantillonnage, cette incertitude pourrait être évitée en utilisant l'estimateur synthétique de type régression  $\mathbf{x}'\tilde{\boldsymbol{\beta}}(0)$  au lieu de l'EBLUP. Datta et coll. (2011) ont proposé un estimateur pour petits domaines basé sur une procédure de test préliminaire de l'hypothèse  $H_0 : A = 0$  contre  $H_1 : A > 0$ . Si  $H_0$  n'est pas rejetée, l'estimateur synthétique de type régression est utilisé comme estimateur de  $\theta_i$ ; sinon, l'EBLUP habituel est utilisé. Ils ont proposé la statistique de test

$$T = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{TP}})' \mathbf{D}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{TP}}),$$

où  $\hat{\boldsymbol{\beta}}_{\text{TP}} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}^{-1}\mathbf{y}$  est l'estimateur MCP de  $\boldsymbol{\beta}$  obtenu en supposant que  $H_0 : A = 0$  est vérifiée. La statistique de test  $T$  suit une loi  $X_{m-p}^2$  avec  $m - p$  degrés de liberté sous  $H_0$ . Alors, pour un seuil de signification spécifié  $\alpha$ , l'estimateur ETP de  $\boldsymbol{\theta}$  défini par Datta et coll. (2011) est donné par

$$\hat{\boldsymbol{\theta}}_{\text{TP}} = (\hat{\theta}_{\text{TP},1}, \dots, \hat{\theta}_{\text{TP},m})' = \begin{cases} \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{TP}} & \text{si } T \leq X_{m-p,\alpha}^2; \\ \hat{\boldsymbol{\theta}}_{\text{RE}} & \text{si } T > X_{m-p,\alpha}^2, \end{cases}$$

où  $X_{m-p,\alpha}^2$  est la valeur critique supérieure au seuil  $\alpha$  de  $X_{m-p}^2$ . L'estimateur ETP est conçu spécialement pour traiter les cas où le nombre de petits domaines est modeste, disons  $m = 15$ .

Ici, nous proposons d'utiliser la procédure TP pour l'estimation de l'EQM de l'EBLUP, en ne considérant que l'EQM de l'estimateur synthétique  $g_{2i}$  quand l'hypothèse nulle n'est pas rejetée, et l'estimation complète de l'EQM autrement. Mais soulignons que la statistique de test  $T$  dans la procédure TP ne dépend pas de l'estimateur de  $A$ . Cela signifie que, même quand  $H_0$  est rejetée, il peut arriver que  $\hat{A}_{\text{RE}} = 0$ . Donc, ici, nous définissons l'estimateur ETP de l'EQM de l'EBLUP  $\hat{\theta}_{\text{RE},i}$  sous la forme

$$\text{eqm}_{\text{TP}}(\hat{\theta}_{\text{RE},i}) = \begin{cases} g_{2i} & \text{si } T \leq X_{m-p,\alpha}^2 \quad \text{ou } \hat{A}_{\text{RE}} = 0, \\ g_{1i}(\hat{A}_{\text{RE}}) + g_{2i}(\hat{A}_{\text{RE}}) + 2g_{3i}(\hat{A}_{\text{RE}}) & \text{si } T > X_{m-p,\alpha}^2 \quad \text{et } \hat{A}_{\text{RE}} > 0. \end{cases} \quad (4.1)$$

## 5 Maximum de vraisemblance ajusté

Les méthodes d'estimation de  $A$  décrites à la section 2 pourraient produire des estimations nulles. Le cas échéant, les EBLUP attribueront un poids nul aux estimateurs directs dans tous les domaines, quelle que soit l'efficacité de l'estimateur direct dans chaque domaine. Par ailleurs, les praticiens des sondages préfèrent souvent attribuer systématiquement un poids strictement positif aux estimateurs directs, parce qu'ils sont fondés sur des données au niveau de l'unité propres au domaine pour la variable d'intérêt, sans l'hypothèse d'un modèle de régression. Pour cette situation, Li et Lahiri (2010) ont proposé l'estimateur du maximum de vraisemblance ajusté (MVA) qui donne un estimateur strictement positif de  $A$ . Cet estimateur, désigné ici  $\hat{A}_{\text{MVA}}$ , s'obtient en maximisant la vraisemblance ajustée définie par

$$L_{MVA}(A) = A \times L_p(A).$$

L'EBLUP donné en (2.6) avec  $\hat{A} = \hat{A}_{MVA}$  sera noté ci-après sous la forme  $\hat{\theta}_{MVA} = (\hat{\theta}_{MVA,1}, \dots, \hat{\theta}_{MVA,m})'$ .

Notons que  $\hat{\theta}_{MVA}$  attribue des poids strictement positifs aux estimateurs directs.

Li et Lahiri (2010) ont proposé un estimateur sans biais d'ordre deux de l'EQM de  $\hat{\theta}_{MVA,i}$  donné par

$$\begin{aligned} \text{eqm}(\hat{\theta}_{MVA,i}) &= g_{1i}(\hat{A}_{MVA}) + g_{2i}(\hat{A}_{MVA}) + 2g_{3i}(\hat{A}_{MVA}) \\ &- B_i^2(\hat{A}_{MVA})b_{MVA}(\hat{A}_{MVA}), \end{aligned} \quad (5.1)$$

où  $b_{MVA}(A)$  est le biais de  $\hat{A}_{MVA}$  qui est donné par

$$b_{MVA}(A) = \frac{\text{trace}\{\mathbf{P}(A) - \Sigma^{-1}(A)\} + 2/A}{\text{trace}\{\Sigma^{-2}(A)\}}.$$

## 6 Estimateurs combinés

L'estimateur MVA strictement positif de  $A$  présente habituellement un plus grand biais que les estimateurs MV ou MVRE quand  $A$  est relativement petite par rapport aux  $D_i$ . Donc, si nous voulons encore obtenir un estimateur pour petits domaines qui applique un poids strictement positif à l'estimateur direct, afin de réduire le biais susmentionné, il sera préférable de n'utiliser l'estimateur MVA que quand cela est strictement nécessaire; c'est-à-dire, quand les données ne fournissent pas suffisamment de preuves que l'égalité  $A = 0$  n'est pas vraie ou que l'estimateur MVRE résultant de  $A$  est nul. Nous présentons ici deux estimateurs pour petits domaines de  $\theta$  donnant un poids strictement positif à l'estimateur direct, qui ont été obtenus sous forme d'une combinaison de l'EBLUP basé sur la méthode du MVA et de l'EBLUP basé sur l'estimation du MVRE.

Dans la première combinaison proposée, la méthode du MVA est utilisée pour estimer  $A$  quand le test préliminaire ne donne pas lieu au rejet de l'hypothèse nulle et dans la deuxième combinaison proposée, elle est utilisée quand l'estimation du MVRE n'est pas positive. Plus précisément, le premier estimateur combiné, appelé ci-après TP-MVA, est défini par

$$\hat{\theta}_{TPMVA} = \begin{cases} \hat{\theta}_{MVA} & \text{si } T \leq X_{m-p,\alpha}^2 \text{ ou } \hat{A}_{RE} = 0, \\ \hat{\theta}_{RE} & \text{si } T > X_{m-p,\alpha}^2 \text{ et } \hat{A}_{RE} > 0. \end{cases} \quad (6.1)$$

Le deuxième estimateur combiné, appelé MVRE-MVA, est donné par

$$\hat{\theta}_{REMVA} = \begin{cases} \hat{\theta}_{MVA} & \text{si } \hat{A}_{RE} = 0, \\ \hat{\theta}_{RE} & \text{si } \hat{A}_{RE} > 0, \end{cases} \quad (6.2)$$

voir Rubin-Bleuer et Yu (2013). Pour l'estimation de l'EQM de  $\hat{\theta}_{REMVA}$ , ces auteurs ont proposé

$$\text{eqm}(\hat{\theta}_{\text{REMVA},i}) = \begin{cases} \text{eqm}(\hat{\theta}_{\text{MVA},i}) & \text{si } \hat{A}_{\text{RE}} = 0, \\ \text{eqm}(\hat{\theta}_{\text{RE},i}) & \text{si } \hat{A}_{\text{RE}} > 0. \end{cases} \quad (6.3)$$

L'utilisation de  $\text{eqm}(\hat{\theta}_{\text{MVA},i})$  quand  $\hat{A}_{\text{RE}} = 0$  donne lieu à une surestimation importante si la valeur vraie de  $A$  est faible, parce que  $\hat{\theta}_{\text{MVA},i}$  sera plus proche de l'estimateur synthétique de type régression. Donc, nous proposons l'estimateur de l'EQM de rechange

$$\text{eqm}_0(\hat{\theta}_{\text{REMVA},i}) = \begin{cases} g_{2i} & \text{si } \hat{A}_{\text{RE}} = 0, \\ \text{eqm}(\hat{\theta}_{\text{RE},i}) & \text{si } \hat{A}_{\text{RE}} > 0. \end{cases} \quad (6.4)$$

De nouveau, puisque, quand la variance  $A$  est petite,  $\text{eqm}(\hat{\theta}_{\text{RE},i})$  pourrait encore surestimer la vraie valeur de l'EQM de  $\hat{\theta}_{\text{REMVA},i}$ , nous considérons également l'estimateur ETP suivant

$$\text{eqm}_{\text{TP}}(\hat{\theta}_{\text{REMVA},i}) = \begin{cases} g_{2i} & \text{si } T \leq X_{m-p,\alpha}^2 \quad \text{ou} \quad \hat{A}_{\text{RE}} = 0, \\ \text{eqm}(\hat{\theta}_{\text{RE},i}) & \text{si } T > X_{m-p,\alpha}^2 \quad \text{et} \quad \hat{A}_{\text{RE}} > 0. \end{cases} \quad (6.5)$$

## 7 Expériences de simulation

Une étude par simulation a été conçue en vue de répondre aux objectifs suivants :

- Étudier les propriétés, en termes de biais et d'EQM, des estimateurs ETP quand  $\alpha$  varie pour une valeur fixe de  $A$ , et quand  $A$  varie pour une valeur fixe de  $\alpha$ . Nous souhaitons déterminer quelles valeurs de  $\alpha$  sont adéquates pour une valeur donnée de  $A$ .
- Comparer les estimateurs ETP aux EBLUP basés sur le MVRE et aux EBLUP basés sur le MVA.
- Étudier les propriétés des estimateurs proposés de l'EQM en ce qui concerne le biais relatif, ainsi que la couverture et la longueur des intervalles de prédiction.
- Comparer les trois estimateurs pour petits domaines présentés qui attribuent un poids strictement positif à l'estimateur direct pour tous les domaines, à savoir l'EBLUP fondé sur les estimateurs MVA, TP-MVA et MVRE-MVA.

Pour réaliser les objectifs susmentionnés, nous avons généré des données à partir du modèle de Fay-Herriot donné par les équations (2.1) et (2.2) avec une moyenne constante, c'est-à-dire avec  $p = 1$ ,  $\beta = \mu$  et  $\mathbf{x}_i = 1, i = 1, \dots, m$ . Nous posons que  $\mu = 0$  sans perte de généralité, que le nombre de domaines est  $m = 15$  et que  $D_i = 1, i = 1, \dots, m$ . L'étude par simulation a été répétée pour des valeurs croissantes de la variance du modèle,  $A \in \{0,01; 0,02; 0,05; 0,1; 0,2; 1\}$ , ainsi que pour six seuils de signification du test de  $H_0 : A = 0$  contre  $H_0 : A > 0$ , à savoir  $\alpha = \{0,05; 0,1; 0,2; 0,3; 0,4; 0,5\}$ . Pour chaque combinaison de  $A$  et  $\alpha$ , nous avons procédé aux étapes qui suivent pour chaque exécution de la simulation  $\ell = 1, \dots, L$  avec  $L = 10\,000$  exécutions :

1. Générer les données au moyen du modèle hypothétique de moyenne nulle constante; c'est-à-dire

$$\begin{aligned}\theta_i^{(o)} &= v_i^{(o)}, \quad v_i^{(o)} \stackrel{\text{ind}}{\sim} N(0, A), \\ y_i^{(o)} &= \theta_i^{(o)} + e_i^{(o)}, \quad e_i^{(o)} \stackrel{\text{ind}}{\sim} N(0, D_i), \quad i = 1, \dots, m.\end{aligned}$$

2. Calculer les estimateurs suivants de  $\theta$  : l'EBLUP basé sur l'estimation du MVRE de  $A$ ,  $\hat{\theta}_{RE}^{(o)}$ , l'estimation ETP,  $\hat{\theta}_{TP}^{(o)}$ , l'EBLUP basé sur l'estimation du MVA de  $A$ ,  $\hat{\theta}_{MVA}^{(o)}$ , l'estimation combinée TP-MVA  $\hat{\theta}_{TPMVA}^{(o)}$  et l'estimation MVRE-MVA  $\hat{\theta}_{REMVA}^{(o)}$ .
3. Pour chaque domaine  $i = 1, \dots, m$ , calculer : les trois estimations de l'EQM de l'EBLUP  $\hat{\theta}_{RE,i}$  données dans (3.2), (3.3) et (4.1), désignées respectivement par  $eqm^{(o)}(\hat{\theta}_{RE,i})$ ,  $eqm_0^{(o)}(\hat{\theta}_{RE,i})$  et  $eqm_{TP}^{(o)}(\hat{\theta}_{RE,i})$ , et les trois estimations (6.3), (6.4) et (6.5) de l'EQM de l'estimateur combiné pour petits domaines  $\hat{\theta}_{REMVA,i}$ , désignées  $eqm^{(o)}(\hat{\theta}_{REMVA,i})$ ,  $eqm_0^{(o)}(\hat{\theta}_{REMVA,i})$  et  $eqm_{TP}^{(o)}(\hat{\theta}_{REMVA,i})$ , respectivement.
4. Pour chaque domaine  $i = 1, \dots, m$ , obtenir les intervalles de prédiction  $1 - \alpha$  fondés sur l'hypothèse de normalité pour la moyenne de petit domaine  $\theta_i$  basée sur les trois estimateurs considérés de l'EQM de l'EBLUP :

$$\begin{aligned}IC_i^{(o)} &= \hat{\theta}_{RE,i}^{(o)} \mp Z_{\alpha/2} \sqrt{eqm^{(o)}(\hat{\theta}_{RE,i})}, \\ IC_{0,i}^{(o)} &= \hat{\theta}_{RE,i}^{(o)} \mp Z_{\alpha/2} \sqrt{eqm_0^{(o)}(\hat{\theta}_{RE,i})}, \\ IC_{TP,i}^{(o)} &= \hat{\theta}_{RE,i}^{(o)} \mp Z_{\alpha/2} \sqrt{eqm_{TP}^{(o)}(\hat{\theta}_{RE,i})},\end{aligned}$$

où  $Z_{\alpha/2}$  est la valeur critique supérieure au seuil  $\alpha/2$  d'une loi normale centrée réduite.

5. Répéter les étapes 1 à 4 pour  $\ell = 1, \dots, L$ , pour  $L = 10\,000$ . Puis, pour chaque estimateur pour petits domaines  $\hat{\theta}_i \in \{\hat{\theta}_{RE,i}, \hat{\theta}_{TP,i}, \hat{\theta}_{MVA,i}, \hat{\theta}_{TPMVA,i}, \hat{\theta}_{REMVA,i}\}$ ,  $i = 1, \dots, m$ , calculer le biais et l'EQM empiriques sous la forme

$$B(\hat{\theta}_i) = \frac{1}{L} \sum_{\ell=1}^L (\hat{\theta}_i^{(\ell)} - \theta_i^{(o)}), \quad EQM(\hat{\theta}_i) = \frac{1}{L} \sum_{\ell=1}^L (\hat{\theta}_i^{(\ell)} - \theta_i^{(o)})^2.$$

Obtenir ensuite la moyenne sur les domaines des biais et des EQM absolus sous la forme

$$\overline{BA}(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^m |B(\hat{\theta}_i)|, \quad \overline{EQMA}(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^m EQM(\hat{\theta}_i).$$

6. Calculer le biais relatif de chaque estimateur de l'EQM,  $eqm(\hat{\theta}_i)$ , comme il suit

$$BR\{eqm(\hat{\theta}_i)\} = \left\{ \frac{1}{L} \sum_{\ell=1}^L eqm^{(o)}(\hat{\theta}_i) - EQM(\hat{\theta}_i) \right\} / EQM(\hat{\theta}_i).$$

Calculer la moyenne sur les domaines des biais relatifs absolus sous la forme

$$\overline{\text{BRA}} \{ \text{eqm}(\hat{\theta}) \} = \frac{1}{m} \sum_{i=1}^m | \text{BR} \{ \text{eqm}(\hat{\theta}_i) \} |.$$

7. Pour chaque type d'intervalle de prédiction  $\text{IC}_i^{(l)} = (L_i^{(l)}, U_i^{(l)})$ , pour  $\text{IC}_i^{(l)} \in \{ \text{IC}_i^{(l)}, \text{IC}_{0,i}^{(l)}, \text{IC}_{\text{TP},i}^{(l)} \}$  donné à l'étape 4, calculer le taux de couverture (TC) et la longueur moyenne (LM) empiriques comme il suit

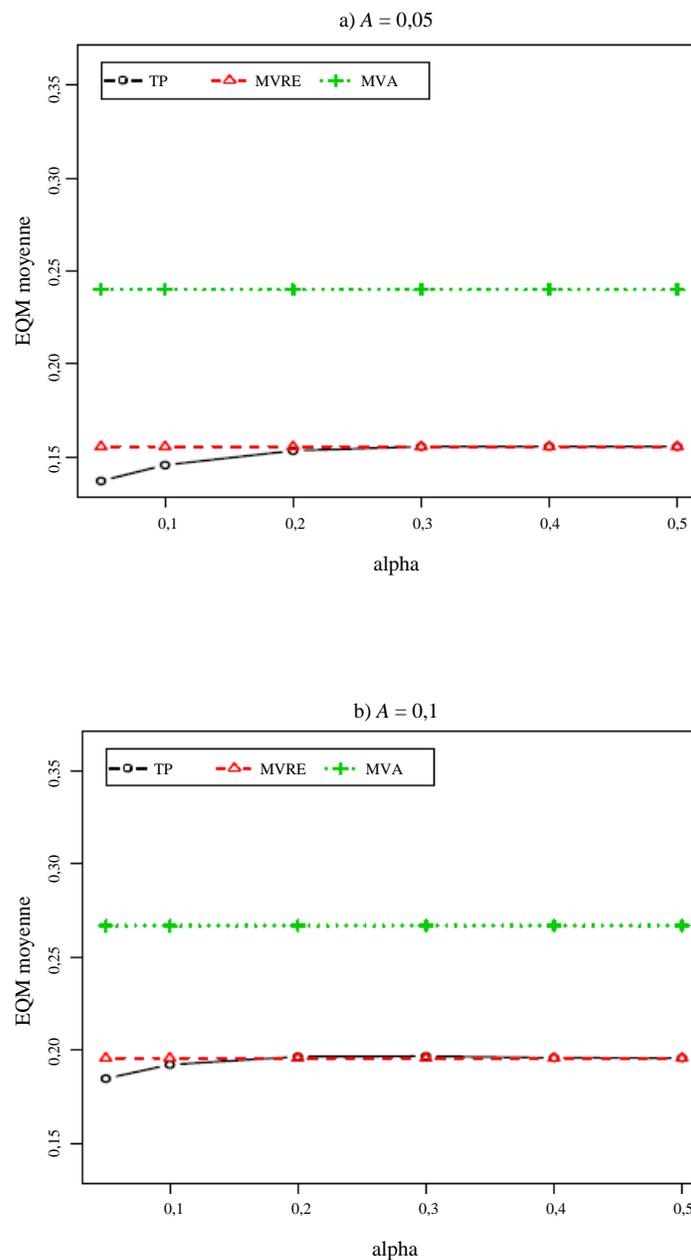
$$\text{TC}(\text{IC}_i) = \frac{\# \{ \theta_i^{(l)} \in \text{IC}_i^{(l)} \}}{L}, \quad \text{LM}(\text{IC}_i) = \frac{1}{L} \sum_{l=1}^L (U_i^{(l)} - L_i^{(l)}).$$

Enfin, calculer la moyenne sur les domaines des taux de couverture et des longueurs moyennes, comme il suit

$$\overline{\text{TC}}(\text{IC}) = \frac{1}{m} \sum_{i=1}^m \text{TC}(\text{IC}_i), \quad \overline{\text{LM}}(\text{IC}) = \frac{1}{m} \sum_{i=1}^m \text{LM}(\text{IC}_i).$$

Les figures 7.1 et 7.2 représentent graphiquement les EQM moyennes des estimateurs ETP pour chaque valeur de  $A \in \{0,05; 0,1; 0,2\}$ , ainsi que l'EQM moyenne des EBLUP basés sur le MVRE et le MVA en fonction du seuil de signification  $\alpha$ . Notons que, quand la valeur de  $A$  est petite, pour une grande valeur de  $\alpha$ , la procédure TP donne lieu plus souvent au rejet de  $H_0$  et par conséquent l'estimateur ETP devient plus fréquemment l'EBLUP usuel, tandis que si la valeur de  $\alpha$  est faible, la procédure TP donne lieu moins fréquemment au rejet de  $H_0$  et l'estimateur synthétique de type régression est alors utilisé plus souvent. Par contre, pour une grande valeur de  $A$ , l'estimateur ETP devient plus fréquemment l'EBLUP quelle que soit la valeur de  $\alpha$ . Les biais absolus des estimateurs ne sont pas présentés ici, parce qu'ils sont à peu près les mêmes pour tous les estimateurs ETP pour les différentes valeurs de  $\alpha$ . Il en est ainsi parce que, quand le modèle est vérifié, les deux composantes de l'estimateur ETP, l'estimateur synthétique et l'EBLUP, sont sans biais pour le paramètre étudié. Notons que l'estimateur synthétique est sans biais même quand  $A > 0$ . La première conclusion qui se dégage des figures 7.1 et 7.2 est que l'EQM de l'estimateur ETP est pratiquement constante pour les diverses valeurs de  $\alpha \geq 0,1$ . Nous voyons aussi que l'EQM moyenne de l'estimateur ETP pour une valeur donnée de  $\alpha$  augmente avec  $A$ , parce que l'estimateur ETP se réduit plus fréquemment à l'EBLUP quand  $A$  augmente et que l'EQM de l'EBLUP augmente avec  $A$ . Observons aussi que l'estimateur ETP et l'EBLUP basé sur le MVRE donnent des résultats très similaires pour  $\alpha \geq 0,2$ . Cependant, pour  $\alpha < 0,2$ , l'estimateur ETP devient plus efficace que l'EBLUP aussitôt que  $A$  s'approche de l'hypothèse nulle ( $A < 0,1$ ), ce qui concorde avec la remarque de Datta et coll. (2011).

Pour l'EBLUP basé sur le MVA, les figures 7.1 et 7.2 montrent que l'EQM moyenne est considérablement plus grande que celles des deux autres estimateurs, mais que les écarts par rapport aux autres diminuent à mesure que  $A$  augmente. Cette situation est attribuable au biais de l'estimateur MVA de  $A$  quand la valeur de  $A$  est petite. Nous étudierons plus loin les estimateurs pour petits domaines combinés TP-MVA et MVRE-MVA, qui n'utilisent l'EBLUP basé sur le MVA que si l'hypothèse nulle n'est pas rejetée ou que l'estimation réalisée de  $A$  est nulle.



**Figure 7.1** EQM moyennes de l'ETP, de l'EBLUP basé sur le MVRE et de l'EBLUP basé sur le MVA en fonction de  $\alpha$ , pour a)  $A = 0,05$  et b)  $A = 0,1$ .

Datta et coll. (2011, page 366) ont recommandé d'utiliser  $\alpha \geq 0,2$  pour l'ETP. En outre, selon la littérature sur l'estimation TP pour les modèles à effets fixes, un bon choix de  $\alpha$  en ce qui concerne le biais et l'EQM est  $\alpha = 0,2$  (Bancroft 1944; Han et Bancroft 1968). Cependant, les résultats susmentionnés donnent à penser que, pour  $\alpha \geq 0,2$ , l'estimateur ETP est pratiquement le même que l'EBLUP et qu'on pourrait par conséquent choisir de toujours utiliser l'EBLUP.

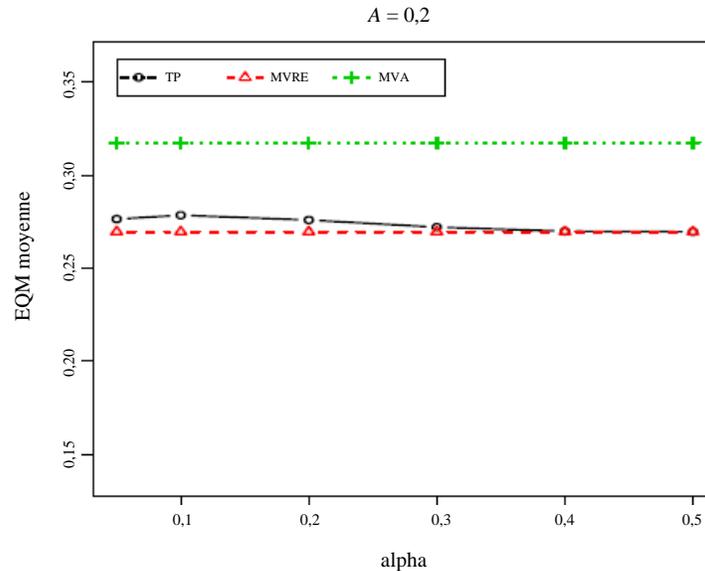


Figure 7.2 EQM moyennes de l'ETP, de l'EBLUP basé sur le MVRE et de l'EBLUP basé sur le MVA en fonction de  $\alpha$ , pour  $A = 0,2$ .

Nous allons maintenant étudier les propriétés de l'estimateur ETP pour l'estimation de l'EQM en fonction de  $\alpha$ . La figure 7.3 représente graphiquement le biais relatif absolu moyen des estimateurs de l'EQM  $eqm_{TP}(\hat{\theta}_{RE,i})$  étiqueté TP en fonction du seuil de signification  $\alpha$  pour chaque valeur  $A \in \{0,05; 0,1; 0,2; 1\}$ . Lorsque l'on choisit  $\alpha$  très petit  $\alpha < 0,1$ , l'hypothèse nulle  $H_0 : A = 0$  est rejetée moins fréquemment et  $eqm_{TP}(\hat{\theta}_{RE,i})$  devient souvent égal à  $g_{2i}$ , ce qui entraîne une sous-estimation. Pour une grande valeur de  $\alpha$  ( $\alpha > 0,2$ ), l'hypothèse nulle est rejetée plus fréquemment et  $eqm_{TP}(\hat{\theta}_{RE,i})$  devient l'estimateur usuel de l'EQM de l'EBLUP, qui surestime fortement la valeur de l'EQM quand  $A$  est petite. La valeur  $\alpha = 0,2$  semble être un bon compromis, avec un biais relatif absolu moyen de l'ordre de 10 % pour  $A \geq 0,1$  et de 20 % pour  $A = 0,05$ .

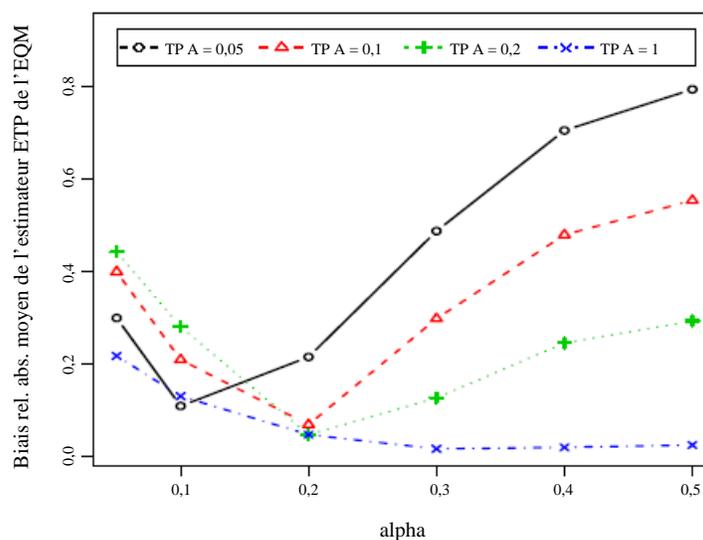
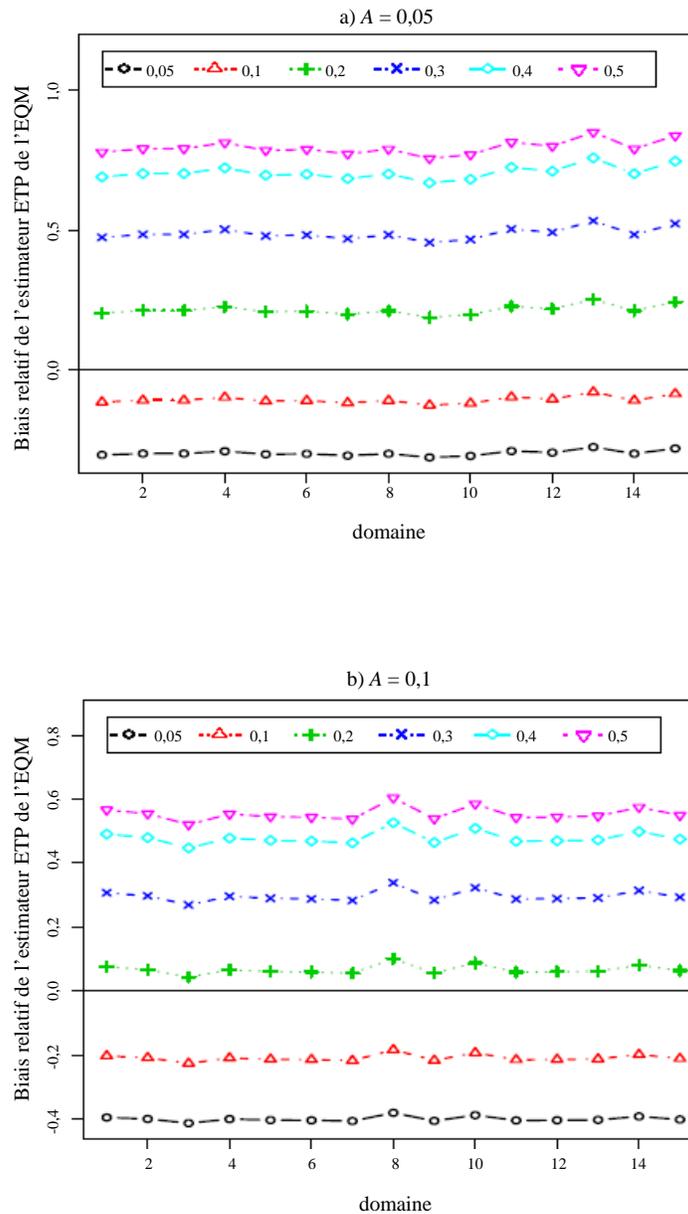
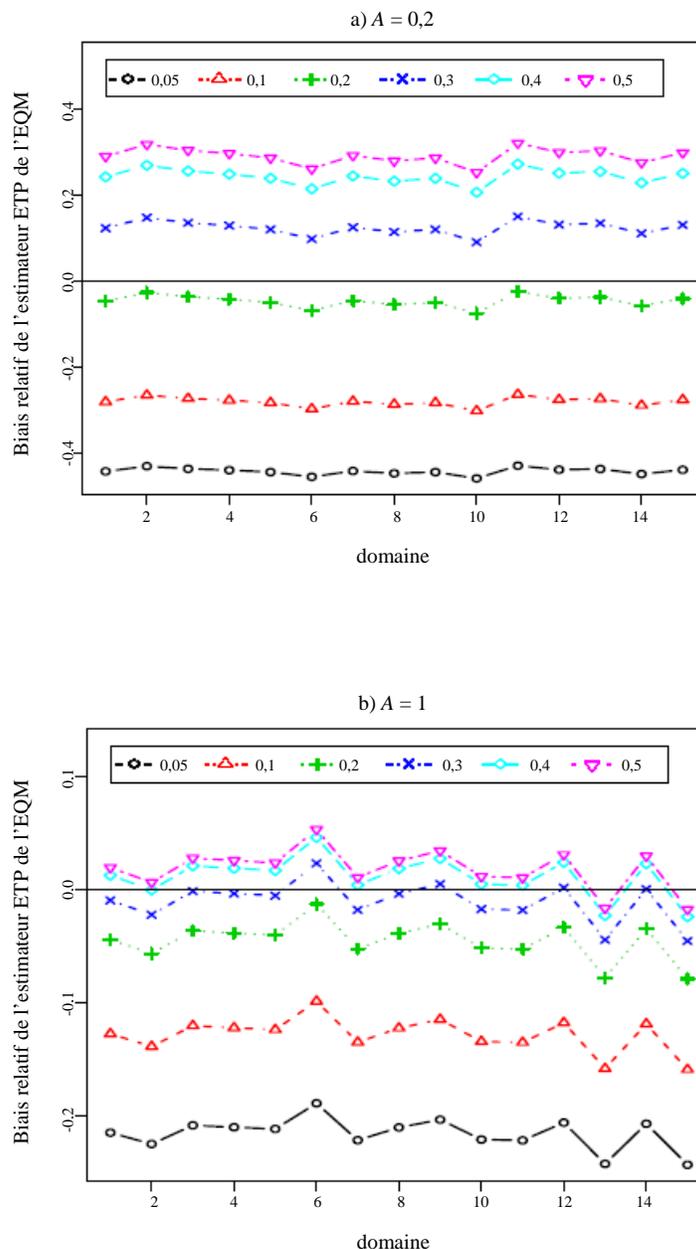


Figure 7.3 Moyenne sur les domaines des biais relatifs absolus de l'estimateur de l'EQM  $eqm_{TP}(\hat{\theta}_{RE,i})$ , étiqueté TP, pour  $A \in \{0,05; 0,1; 0,2; 1\}$  en fonction du seuil de signification  $\alpha$ .

Les résultats susmentionnés donnent à penser que  $\alpha = 0,2$  est un bon choix lorsqu'on utilise la procédure TP pour estimer l'EQM de l'EBLUP usuel. Cette constatation a été étudiée de manière plus approfondie en examinant les biais relatifs (affectés d'un signe) de  $eqm_{TP}(\hat{\theta}_{RE,i})$  pour chaque domaine. Ces résultats sont représentés graphiquement aux figures 7.4 et 7.5, avec quatre graphiques, un pour chaque valeur de  $A \in \{0,05; 0,1; 0,2; 1\}$ . Les chiffres qui figurent dans les légendes de ces graphiques sont les seuils de signification  $\alpha$  pour l'estimateur ETP de l'EQM  $eqm_{TP}(\hat{\theta}_{RE,i})$ . Ces graphiques confirment nos observations antérieures, à savoir que l'estimateur de l'EQM fondé sur l'ETP,  $eqm_{TP}(\hat{\theta}_{RE,i})$ , sous-estime EQM  $(\hat{\theta}_{RE,i})$  pour les faibles valeurs de  $\alpha$  et la surestime pour les grandes valeurs de  $\alpha$ . Il s'avère que  $eqm_{TP}(\hat{\theta}_{RE,i})$  avec  $\alpha = 0,2$  convient bien pour toutes les valeurs de  $A$ .



**Figure 7.4** Biais relatif de  $eqm_{TP}(\hat{\theta}_{RE,i})$  pour chaque seuil de signification  $\alpha \in \{0,05; 0,1; 0,2; 0,3; 0,4; 0,5\}$  en fonction du domaine  $i$ , pour a)  $A = 0,05$  et b)  $A = 0,1$ .



**Figure 7.5** Biais relatif de  $eqm_{TP}(\hat{\theta}_{RE,i})$  pour chaque seuil de signification  $\alpha \in \{0,05; 0,1; 0,2; 0,3; 0,4; 0,5\}$  en fonction du domaine  $i$ , pour a)  $A = 0,2$  et b)  $A = 1$ .

Comparons maintenant  $eqm_{TP}(\hat{\theta}_{RE,i})$  pour le seuil de signification choisi de  $\alpha = 0,2$  aux deux autres estimateurs de l'EQM,  $eqm_0(\hat{\theta}_{RE,i})$  et  $eqm(\hat{\theta}_{RE,i})$ , donnés par (3.3) et (3.2), respectivement. La figure 7.6 représente graphiquement les biais relatifs absolus moyens des trois estimateurs de l'EQM, étiquetés respectivement TP, MVRE0 et MVRE. Nous constatons que  $eqm_0(\hat{\theta}_{RE,i})$  donne de meilleurs résultats que  $eqm(\hat{\theta}_{RE,i})$  pour tous les domaines, mais que  $eqm_{TP}(\hat{\theta}_{RE,i})$  demeure meilleur que  $eqm_0(\hat{\theta}_{RE,i})$  pour toutes les valeurs considérées de  $A$  sauf  $A = 1$ , valeur pour laquelle les différences entre les trois estimateurs sont négligeables. Les écarts diminuent à mesure que  $A$  augmente, mais

soulignons que l'estimateur de l'EQM usuel,  $eqm(\hat{\theta}_{RE,i})$ , peut être sévèrement biaisé si la valeur de  $A$  est petite, avec un biais relatif absolu moyen supérieur à 50 % pour  $A < 0,2$  et croissant exponentiellement quand  $A$  tend vers zéro. La conclusion est que, quand  $H_0$  n'est pas rejetée, même si l'estimation réalisée de  $A$  est positive, il semble préférable d'omettre le terme  $g_{3i}$  dans l'estimateur de l'EQM et de ne considérer que  $g_{2i}$ .

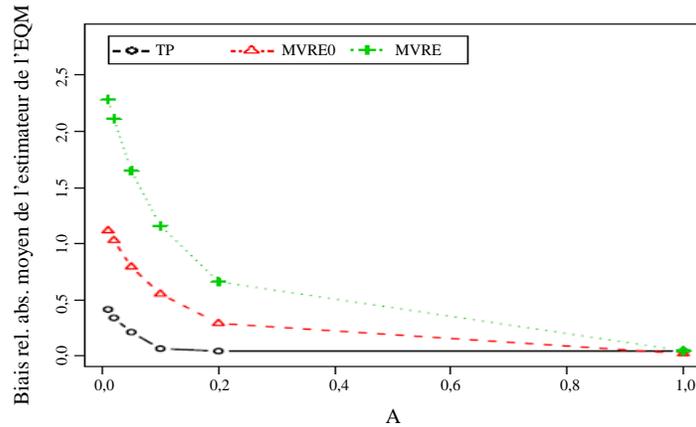


Figure 7.6 Moyenne sur les domaines des biais relatifs absolus des estimateurs de l'EQM  $eqm_{TP}(\hat{\theta}_{RE,i})$  avec  $\alpha = 0,2$ , étiqueté TP,  $eqm(\hat{\theta}_{RE,i})$  étiqueté MVRE et  $eqm_0(\hat{\theta}_{RE,i})$  étiqueté MVRE0, en fonction de  $A$ .

Examinons maintenant les estimateurs pour petits domaines qui appliquent un poids strictement positif à l'estimateur direct pour tous les domaines, à savoir l'EBLUP basé sur le MVA,  $\hat{\theta}_{MVA}$ , et les deux estimateurs combinés, TP-MVA donné en (6.1) et MVRE-MVA donné en (6.2). Les EQM moyennes sont représentées graphiquement à la figure 7.7 pour ces trois estimateurs. Dans ce graphique,  $\hat{\theta}_{MVA}$  semble être un peu moins efficace, et est suivi par TP-MVA. L'estimateur combiné MVRE-MVA semble donner d'un peu meilleurs résultats que les deux autres pour une faible valeur de  $A$ , quoique pour  $A \geq 0,2$ , l'estimateur TP-MVA est très proche. Pour l'estimation de l'EQM, nous nous concentrons sur l'estimateur MVRE-MVA en raison de sa meilleure performance.

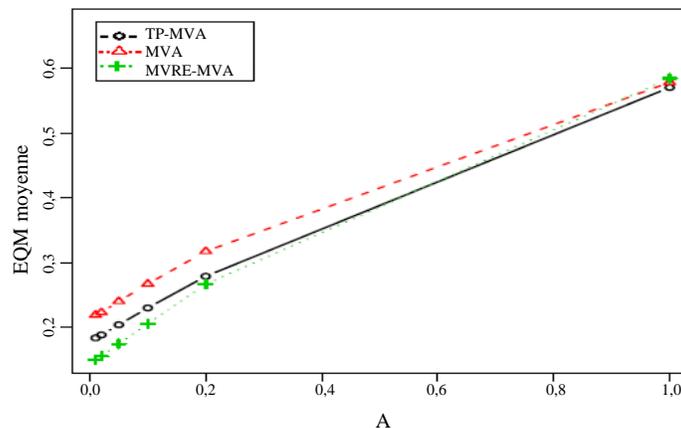
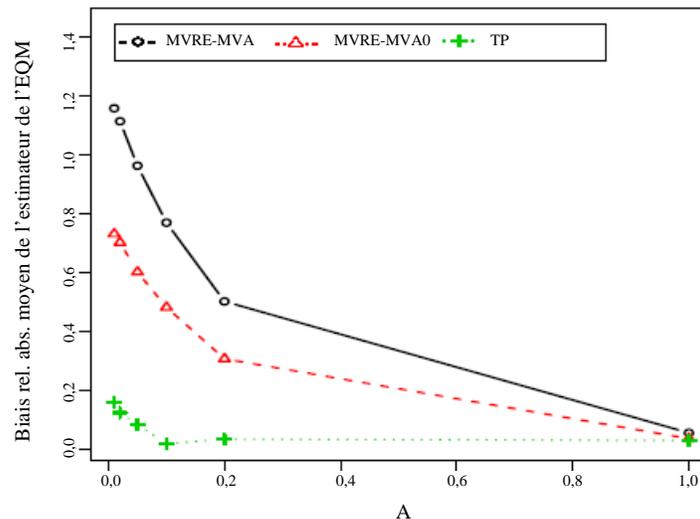


Figure 7.7 Moyenne sur les domaines des EQM pour l'estimateur TP-MVA avec  $\alpha = 0,2$ , l'EBLUP basé sur le MVA et l'estimateur MVRE-MVA en fonction de  $A$ .

Pour l'estimateur combiné MVRE-MVA, la figure 7.8 montre que l'estimateur de l'EQM basé sur le test préliminaire TP,  $eqm_{TP}(\hat{\theta}_{REMVA,i})$  qui utilise seulement  $g_{2i}$  quand  $\hat{A}_{RE} = 0$  ou que l'hypothèse nulle n'est pas rejetée, présente un biais relatif absolu moyen inférieur à 10 % pour  $A \geq 0,1$  et est plus faible que les valeurs correspondantes pour  $eqm(\hat{\theta}_{REMVA,i})$  et  $eqm_0(\hat{\theta}_{REMVA,i})$ , spécialement pour  $A \leq 0,4$ .



**Figure 7.8** Moyenne sur les domaines des biais relatifs absolus des estimateurs de l'EQM  $eqm(\hat{\theta}_{REMVA,i})$ ,  $eqm_0(\hat{\theta}_{REMVA,i})$  et  $eqm_{TP}(\hat{\theta}_{REMVA,i})$ , étiquetés respectivement MVRE-MVA, MVRE-MVA0 et TP, en fonction de A.

Enfin, nous analysons la moyenne sur les domaines des taux de couverture et des longueurs moyennes des intervalles de prédiction fondés sur l'hypothèse de normalité pour la moyenne de petit domaine  $\theta_i$  en utilisant l'EBLUP basé sur le MVRE comme estimation ponctuelle et les trois estimateurs différents de l'EQM de l'EBLUP, à savoir  $eqm(\hat{\theta}_{RE,i})$ ,  $eqm_0(\hat{\theta}_{RE,i})$  et  $eqm_{TP}(\hat{\theta}_{RE,i})$ . La figure 7.9 représente les taux de couverture des trois types d'intervalles, où les estimateurs de l'EQM basés sur la procédure TP ont été obtenus en prenant  $\alpha = 0,2; 0,3$ . Il semble que les bonnes propriétés de biais relatif de l'estimateur de l'EQM basé sur la procédure TP,  $eqm_{TP}(\hat{\theta}_{RE,i})$ , pour une valeur faible de A ne peuvent pas être extrapolées à la couverture basée sur les intervalles de prédiction normaux, et présentent une sous-couverture surtout pour  $A = 0,2$ . Dans ce cas, choisir un seuil de signification plus élevé,  $\alpha = 0,3$ , réduit un peu la couverture insuffisante des intervalles de prédiction obtenus en utilisant  $eqm_{TP}(\hat{\theta}_{RE,i})$ . Néanmoins, les taux de couverture de  $eqm_0(\hat{\theta}_{RE,i})$  sont meilleurs pour toutes les valeurs de A. Comme prévu, l'estimateur usuel de l'EQM  $eqm(\hat{\theta}_{RE,i})$  donne une surcouverture pour les petites valeurs de A, laquelle résulte de la forte surestimation de l'EQM. Par ailleurs, les intervalles pour lesquels on observe une sous-couverture entraînent aussi des intervalles de prédiction plus courts, comme le montre la figure 7.10.

Il est utile de mentionner que la construction des intervalles de prédiction pour  $\theta_i$  basés sur le modèle de Fay-Herriot avec des taux de couverture exacts n'est pas une tâche évidente. Plusieurs articles traitant de ce problème ont été publiés. Par exemple, Chatterjee, Lahiri et Li (2008) ont proposé des intervalles de

prédiction avec taux de couverture corrects jusqu'à l'ordre deux en utilisant uniquement le terme  $g_{li}$  comme estimation de l'EQM et en appliquant une procédure bootstrap pour trouver les quantiles calés. Diao, Smith, Datta, Maiti et Opsomer (2014) ont obtenu récemment des intervalles de prédiction avec taux de couverture corrects jusqu'à l'ordre deux en évitant d'utiliser des procédures de rééchantillonnage et en utilisant l'estimateur complet de l'EQM. L'obtention d'intervalles de prédiction dont la couverture est exacte en utilisant d'autres estimations de l'EQM pose encore des difficultés et dépasse le cadre du présent article.

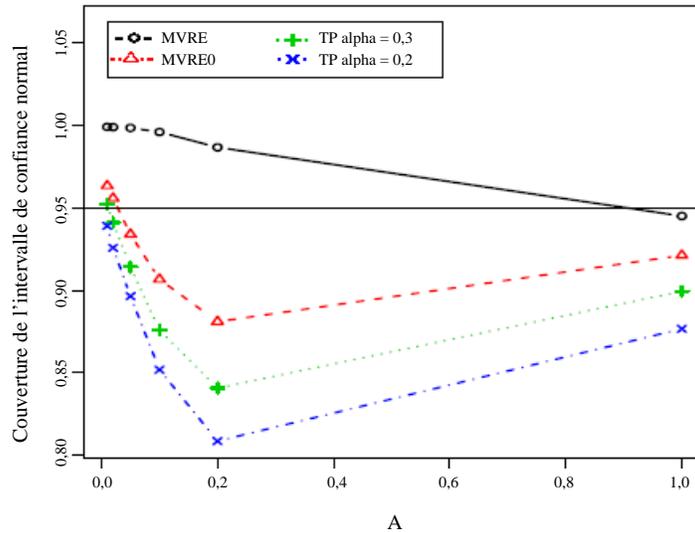


Figure 7.9 Moyenne sur les domaines des taux de couverture des intervalles de prédiction fondés sur la normalité pour  $\theta_i$  en utilisant les estimateurs de l'EQM  $eqm(\hat{\theta}_{RE,i})$ ,  $eqm_0(\hat{\theta}_{RE,i})$  et  $eqm_{TP}(\hat{\theta}_{RE,i})$  avec  $\alpha = 0,2; 0,3$ , étiquetés respectivement MVRE, MVRE0 et TP, en fonction de A.

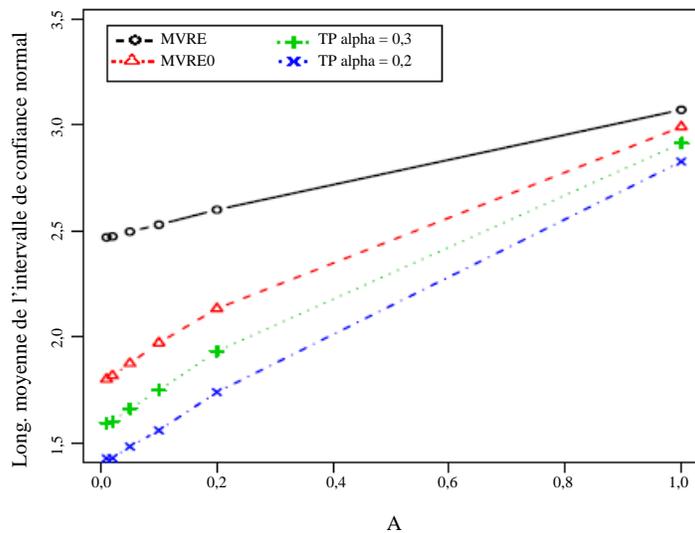


Figure 7.10 Moyenne sur les domaines des longueurs moyennes des intervalles basés sur l'hypothèse de normalité pour  $\theta_i$  en utilisant les estimateurs de l'EQM  $eqm(\hat{\theta}_{RE,i})$ ,  $eqm_0(\hat{\theta}_{RE,i})$  et  $eqm_{TP}(\hat{\theta}_{RE,i})$  avec  $\alpha = 0,2; 0,3$ , étiquetés respectivement MVRE, MVRE0 et TP, en fonction de A.

L'étude par simulation dont la description précède a été répétée pour plusieurs profils de variances d'échantillonnage inégales  $D_i$ . Les résultats ne sont pas présentés ici, mais les conclusions sont très semblables à condition que le profil de variance ne soit pas extrêmement irrégulier.

## 8 Conclusion

Les principales conclusions qui se dégagent des résultats de notre étude par simulation de l'estimation des moyennes de petit domaine basée sur le modèle de Fay-Herriot au niveau du domaine quand le nombre de domaines est modéré (disons  $m = 15$ ) sont les suivantes : 1) Sous le modèle de Fay-Herriot avec une valeur de la variance des effets aléatoires,  $A$ , clairement différente de zéro, l'estimateur ETP ne semble pas améliorer appréciablement l'efficacité relative de l'EBLUP usuel, à moins que le seuil de signification soit faible ( $\alpha \leq 0,1$  dans notre étude par simulation). 2) Nos résultats de simulation indiquent que l'utilisation de la procédure TP avec une valeur modérée de  $\alpha$ , en particulier  $\alpha = 0,2$ , pour estimer l'EQM de l'EBLUP usuel donne lieu à une réduction du biais comparativement à l'estimateur de l'EQM usuel. D'où, nous recommandons d'utiliser  $eqm_{TP}(\hat{\theta}_{RE,i})$ , donné par (4.1), pour estimer l'EQM de l'EBLUP. 3) Parmi les estimateurs qui appliquent un poids strictement positif à l'estimateur direct pour tous les domaines, nous recommandons l'estimateur combiné MVRE-MVA donné par (6.2), parce que son efficacité est un peu plus élevée que celle de l'EBLUP basé sur le MVA et le TP-MVA donné par (6.1). 4) Pour estimer l'EQM de l'estimateur MVRE-MVA recommandé, l'estimateur  $eqm_{TP}(\hat{\theta}_{REMVA,i})$  donné par (6.5) produit de meilleurs résultats que les estimateurs de rechange. 5) Nos résultats concernant les intervalles de prédiction, basés sur la théorie de la normalité, indiquent que la bonne performance des estimateurs proposés de l'EQM pourrait ne pas se traduire en bonnes propriétés de couverture de ces intervalles. La construction d'intervalles de prédiction donnant une couverture exacte en utilisant les estimations proposées de l'EQM semble être une tâche difficile.

Des options lisses des estimations avec test préliminaire dans le cas des paramètres de position ont été proposées dans la littérature en utilisant des moyennes pondérées des estimations obtenues sous les hypothèses nulle et alternative, avec des poids dépendant de la statistique de test, voir par exemple, Saleh (2006). Les estimations de l'erreur quadratique moyenne de ce type n'ont pas été étudiées et nous réservons ce sujet pour de futurs travaux de recherche.

## Remerciements

Nous tenons à remercier le rédacteur de ses suggestions très constructives. Les travaux de recherche de Gauri S. Datta ont été financés en partie par la subvention H98230-11-1-0208 de la *National Security Agency*, les travaux de recherche d'Isabel Molina, par les subventions MTM2009-09473, MTM2012-37077-C02-01 et SEJ2007-64500 du *Ministerio de Educación y Ciencia* de l'Espagne, et les travaux de recherche de J.N.K. Rao, par le Conseil de recherches en sciences naturelles et en génie du Canada.

## Bibliographie

- Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics*, 15, 190-204.
- Chatterjee, S., Lahiri, P. et Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, 36, 1221-1245.
- Das, K., Jiang, J. et Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.
- Datta, G.S., et Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.
- Datta, S., Hall, P. et Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106, 362-374.
- Diao, L., Smith, D.D., Datta, G.S., Maiti, T. et Opsomer, J.D. (2014). Accurate confidence interval estimation of small area parameters under the Fay-Herriot model. *Scandinavian Journal of Statistics*, à paraître.
- Fay, R.E., et Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Han, C.-P., et Bancroft, T.A. (1968). On pooling means when variance is unknown. *Journal of the American Statistical Association*, 63, 1333-1342.
- Li, H., et Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.
- 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*. Hoboken, NJ : Wiley.
- Rubin-Bleuer, S., et Yu, Y. (2013). A positive variance estimator for the Fay-Herriot small area model. SRID2-12-001E, Statistique Canada.
- Saleh, A.K. Md. E. (2006). *Theory of Preliminary Test and Stein-type Estimation with Applications*. New York : John Wiley & Sons, Inc.