Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot

par Éric Lesage, Jean-François Beaumont et Cynthia Bocci

Date de diffusion : le 6 janvier 2022

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

Courriel à infostats@statcan.gc.ca

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

Programme des services de dépôt

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

Normes de service à la clientèle

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

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.

Techniques d'enquête, décembre 2021
Vol. 47, nº2, pp. 303-323
Statistique Canada, nº 12-001-X au catalogue

Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu du modèle de Fay-Herriot

Éric Lesage, Jean-François Beaumont et Cynthia Bocci[1]

Résumé

Le modèle de Fay-Herriot est souvent utilisé pour obtenir des estimations sur petits domaines. Ces estimations sont généralement plus efficaces que les estimations directes classiques. Afin d'évaluer les gains d'efficacité obtenus par les méthodes d'estimation sur petits domaines, on produit généralement des estimations de l'erreur quadratique moyenne fondée sur le modèle. Cependant, ces estimations ne permettent pas de tenir compte de toute la spécificité d'un domaine en particulier car elles font disparaître l'effet local en prenant une espérance par rapport au modèle. Une alternative consiste à estimer l'erreur quadratique moyenne fondée sur le plan de sondage des estimateurs sur petits domaines. Cette dernière est souvent plus attrayante du point de vue des utilisateurs. Il est cependant connu que les estimateurs de l'erreur quadratique moyenne fondée sur le plan de sondage peuvent être très instables, particulièrement pour les domaines qui contiennent peu d'unités échantillonnées. Dans cet article, nous proposons deux diagnostics locaux qui ont pour objectif de faire un choix entre le meilleur prédicteur empirique et l'estimateur direct pour un domaine en particulier. Nous trouvons d'abord un intervalle de valeurs de l'effet local tel que le meilleur prédicteur est plus efficace sous le plan que l'estimateur direct. Ensuite, nous considérons deux approches différentes pour évaluer s'il est plausible que l'effet local se trouve dans cet intervalle. Nous examinons nos diagnostics au moyen d'une étude par simulation. Nos résultats préliminaires semblent prometteurs quant à l'utilité de ces diagnostics pour choisir entre le meilleur prédicteur empirique et l'estimateur direct.

Mots-clés : Meilleur prédicteur empirique; erreur quadratique moyenne fondée sur le plan; erreur quadratique moyenne fondée sur le modèle; diagnostic local; effet local; modèle de Fay-Herriot.

1. Introduction

Les gouvernements ont besoin d'information socio-économique à des niveaux de détail de plus en plus fins. Les instituts de statistique nationaux sont donc amenés à produire des statistiques pour des sous-populations qui n'avaient pas été identifiées au moment de la définition des objectifs de précision de l'enquête ou du moins qui ne pouvaient pas être pris en compte. Il en résulte que, pour ces sous-populations, le nombre d'unités enquêtées peut être trop faible pour garantir une bonne précision des estimateurs directs classiques de la théorie des sondages comme l'estimateur de Horvitz-Thompson ou l'estimateur par calage. Ce type de sous-population, où l'échantillon est de taille insuffisante, est appelé un petit domaine. Pour pallier le manque de précision des estimateurs directs sur les petits domaines, on peut avoir recours à des estimateurs indirects, ou estimateurs sur petits domaines, qui s'appuient sur des modèles, comme par exemple le modèle de Fay-Herriot (Fay et Herriot, 1979). Le meilleur prédicteur empirique, aussi appelé l'estimateur EB (pour *Empirical Best* ou *Empirical Bayes*), est un estimateur sur petits domaines fréquemment utilisé en pratique.

Les méthodes d'estimation sur petits domaines utilisent des modèles statistiques pour tirer le meilleur parti de l'information provenant de l'enquête et de l'information provenant de sources de données auxiliaires. Le modèle de Fay-Herriot est un modèle linéaire qui permet de décomposer le paramètre

1. Éric Lesage, Insee, Direction régionale, 35, place du Colombier - CS 94439 - 35044 Rennes Cedex. Courriel : eric.lesage@insee.fr; Jean-François Beaumont, Statistique Canada, 100 Pré Tunney, Immeuble R.H. Coats, Ottawa, Ontario, Canada, K1A 0T6. Courriel : jean-francois.beaumont@statcan.gc.ca; Cynthia Bocci, Statistique Canada, 100 Pré Tunney, Immeuble R.H. Coats, Ottawa, Ontario, Canada, K1A 0T6. Courriel : cynthia.bocci@statcan.gc.ca.

d'intérêt d'un domaine en deux termes : le premier terme est l'effet expliqué par le modèle et le second terme est l'erreur du modèle qu'on peut interpréter comme un effet local inexpliqué et inconnu.

On peut utiliser des outils statistiques classiques pour juger de la validité du modèle de Fay-Herriot tels que, par exemple, des graphiques de résidus du modèle. Ces outils donnent toutefois peu d'indications sur l'efficacité d'une estimation indirecte pour un domaine en particulier. L'erreur quadratique moyenne (EQM) fondée sur le modèle de Fay-Herriot d'un estimateur indirect peut être considérée comme un indicateur local de qualité puisqu'elle varie d'un domaine à l'autre. L'EQM fondée sur le modèle prend en compte l'effet expliqué par le modèle mais fait disparaître l'effet local inexpliqué en prenant une espérance par rapport au modèle.

L'EQM fondée sur le plan de sondage est une alternative à l'EQM fondée sur le modèle qui ne fait pas disparaître l'effet local inexpliqué. Cependant, les estimateurs sans biais de l'EQM fondée sur le plan d'estimateurs sur petits domaines ont tendance à être très instables, particulièrement pour les domaines qui contiennent peu d'unités échantillonnées (Rivest et Belmonte, 2000; Rao, Rubin-Bleuer et Estevao, 2018; et Pfeffermann et Ben-Hur, 2019). Pour contourner ce problème, on a suggéré dans la littérature de prendre une moyenne sur plusieurs domaines de l'EQM fondée sur le plan (Rao et Molina, 2015; et Pfeffermann et Ben-Hur, 2019) comme mesure de qualité. Cependant, de nombreux utilisateurs des statistiques publiques ne s'intéressent qu'à leur seul domaine et ne sont pas preneurs d'un critère global de qualité pour juger de l'efficacité des estimations pour leur domaine d'intérêt. C'est d'autant plus le cas lorsqu'ils sont convaincus que leur domaine est très spécifique et que cette spécificité ne se retrouve pas dans le terme explicatif du modèle mais plutôt dans le terme d'erreur, i.e. l'effet local inexpliqué.

Pour faire face au problème de l'instabilité des estimateurs sans biais de l'EQM fondée sur le plan, Rao, Rubin-Bleuer et Estevao (2018) ont proposé un estimateur composite qu'ils évaluent au moyen d'une étude par simulations. Leur estimateur composite consiste à prendre une moyenne pondérée d'un estimateur de l'EQM fondée sur le modèle et d'un estimateur de l'EQM fondée sur le plan. Ils obtiennent ainsi plus de stabilité au prix d'une augmentation du biais. Pfeffermann et Ben-Hur (2019) proposent également une méthode pour estimer l'EQM fondée sur le plan d'un estimateur sur petits domaines. La méthode est plutôt complexe et repose principalement sur le choix d'un modèle approprié. Elle n'est donc pas entièrement fondée sur le plan de sondage. Outre ces tentatives d'estimer l'EQM fondée sur le plan, il n'existe pas, à notre connaissance, de diagnostics locaux qui permettent de déterminer si l'estimation sur petits domaines est préférable à l'estimation directe pour un domaine en particulier.

Dans cet article, on propose une approche différente qui a pour objectif de comparer l'efficacité sous le plan des estimateurs EB et direct. On procède en deux temps. Dans un premier temps on détermine l'intervalle des valeurs de l'effet local inexpliqué qui garantissent que l'EQM fondée sur le plan du meilleur prédicteur, ou estimateur B (pour *Best* ou *Bayes*), soit inférieure à l'EQM fondée sur le plan de l'estimateur direct. Dans un second temps, on cherche à évaluer s'il est plausible que l'effet local inexpliqué se situe dans cet intervalle. A cette fin, on propose deux diagnostics : un premier qui s'appuie sur la distribution conditionnelle aux estimations directes de l'effet local inexpliqué et un second qui s'appuie sur un test d'hypothèse sur l'effet local inexpliqué, réalisé par rapport au plan de sondage. Nous

avons trouvé que, selon l'importance du résidu normalisé du modèle et d'un facteur associé à la précision de l'estimation directe, il est possible de détecter si les estimateurs B ou EB sont susceptibles d'avoir une EQM fondée sur le plan plus petite que celle de l'estimateur direct.

À la section 2, on présente le modèle de Fay-Herriot et on rappelle comment est construit le meilleur prédicteur (estimateur B) du paramètre d'intérêt de la population. À la section 3, on dérive les EQM fondées sur le modèle et fondées sur le plan de l'estimateur direct et du meilleur prédicteur. À la section 4, on décrit les deux diagnostics proposés. À la section 5, on explique comment estimer les paramètres du modèle et ainsi obtenir le meilleur prédicteur empirique (estimateur EB) et les estimateurs des diagnostics. À la section 6, on présente les résultats d'une étude par simulation à partir de données auxiliaires réelles. On termine par une brève conclusion à la section 7.

2. Le modèle de Fay-Herriot et le meilleur prédicteur

On considère une population finie U de taille N et un échantillon s de taille n tiré dans U selon un plan de sondage $p(s)$. La population U est partitionnée en m domaines qui ne se chevauchent pas. Les domaines sont repérés par un indice i qui prend les valeurs de là m. La population du domaine i, de taille N_i, est notée U_i. L'échantillon du domaine i est noté s_i et sa taille est n_i. On s'intéresse à estimer m paramètres de la population finie, $\theta_i, i = 1, \ldots, m,$ associés aux m domaines. Le paramètre θ_i est habituellement un total, une moyenne ou un ratio pour le domaine i. On dispose d'information auxiliaire sous la forme de vecteurs, $\mathbf{z}_i$, disponibles pour tous les domaines $i = 1, \ldots, m$. On note $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1,\ldots,m}$, l'ensemble contenant les m vecteurs auxiliaires. On note également Ω, l'ensemble de toutes les variables utilisées pour faire les inférences excluant les variables indicatrices d'appartenance à l'échantillon s; Ω inclut entre autres $\mathcal{Z}$ et $\theta_i, i = 1, \ldots, m$. Ainsi, on notera l'espérance par rapport au plan de sondage d'une variable aléatoire, disons A, par $\mathbb{E}(A \mid \Omega)$.

On pose un modèle de liaison qui nous permet de décomposer les paramètres d'intérêt θ_i de la façon suivante :

$$\theta_i = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i + b_i v_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

où $\boldsymbol{\beta}$ est un vecteur de paramètres du modèle de même dimension que $\mathbf{z}_i$, les b_i sont des facteurs fixes qui permettent de tenir compte de l'hétérosédasticité dans le modèle et les v_i sont des termes d'erreur qui suivent la loi normale : $v_i \mid \mathcal{Z} \sim \mathcal{N}(0, \sigma_v^2)$, où σ_v^2 est un paramètre du modèle. En pratique, on pose souvent $b_i = 1$ mais il est plus naturel de poser $b_i = N_i$ quand θ_i est un total. Le terme $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{z}_i$ est l'effet connu ou expliqué par le modèle du paramètre de population finie θ_i alors que $b_i v_i$ est l'effet inconnu ou inexpliqué que l'on appelle effet local inexpliqué de θ_i ou encore simplement effet local de θ_i.

L'estimateur direct de θ_i est noté $\hat{\theta}_i$. Il est habituellement obtenu en pondérant chaque unité de l'échantillon s_i par un poids de sondage. Le poids de sondage d'une unité peut tout simplement être l'inverse de sa probabilité de sélection dans l'échantillon s ou un poids de calage. On note e_i l'erreur d'échantillonnage :

$$e_i = \hat{\theta}_i - \theta_i. \tag{2.2}$$

On considèrera par la suite que l'estimateur direct est sans biais sous le plan de sondage, c'est-à-dire que $\mathbb{E}(\hat{\theta}_i \mid \Omega) = \theta_i$ ou encore que $\mathbb{E}(e_i \mid \Omega) = 0.$ Cette hypothèse n'est pas toujours satisfaite en pratique, par exemple lorsqu'on utilise des poids de calage, mais on fera l'hypothèse habituelle que le biais reste négligeable. On suppose également que l'estimateur direct $\hat{\theta}_i,$ et ainsi l'erreur $e_i,$ suit une loi normale. Tel que discuté dans Rao et Molina (2015, page 77), l'hypothèse de normalité des erreurs e_i est possiblement moins forte que celle de la normalité des erreurs v_i en raison de l'effet du théorème central limite sur les $\hat{\theta}_i$. Bien sûr, cet effet est moins important pour les plus petits domaines. Sous ces hypothèses, on a : $e_i \mid \Omega \sim \mathcal{N}(0, \psi_i),$ où $\psi_i = \mathbb{V}(\hat{\theta}_i \mid \Omega)$ est la variance sous le plan de sondage de $\hat{\theta}_i$. La taille d'échantillon n_i peut être très faible, ce qui peut conduire à une mauvaise précision de l'estimateur direct $\hat{\theta}_i$. Ce problème est à l'origine de la recherche sur l'estimation sur petits domaines.

En combinant le modèle (2.1) et l'expression (2.2), on obtient le modèle combiné, encore appelé modèle de Fay-Herriot :

$$\hat{\theta}_i = \boldsymbol{\beta}^\top \mathbf{z}_i + b_i\, v_i + e_i. \tag{2.3}$$

En notant que v_i est fixe sous le plan de sondage, on peut facilement montrer que $\mathbb{V}\left(b_i v_i + e_i \mid \mathcal{Z}\right) = b_i^2 \sigma_v^2 + \tilde{\psi}_i,$ où $\tilde{\psi}_i = \mathbb{E}\left(\psi_i \mid \mathcal{Z}\right)$ est la variance lissée (voir la remarque à la fin de cette section). On note ε_i l'erreur normalisée du modèle combiné :

$$\varepsilon_i = \frac{\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i}{\sqrt{b_i^2 \sigma_v^2 + \tilde{\psi}_i}}. \tag{2.4}$$

L'observation des estimations directes $\hat{\theta}_i$ fournit de l'information sur les θ_i. On peut obtenir les distributions conditionnelles des θ_i (Rao et Molina, 2015, chapitre 9, pages 271-272) :

$$\theta_i \mid \mathcal{Z}, \hat{\theta}_i \sim \mathcal{N}\left\{\boldsymbol{\beta}^\top \mathbf{z}_i + \gamma_i(\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i), (1-\gamma_i)\, b_i^2 \sigma_v^2\right\}, \tag{2.5}$$

où $\gamma_i = \frac{b_i^2 \sigma_v^2}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}$.

Le meilleur prédicteur de θ_i conditionnellement à l'observation de $\hat{\theta}_i$ (Rao et Molina, 2015) est alors donné par :

$$\hat{\theta}_i^B = \mathbb{E}(\theta_i \mid \mathcal{Z}, \hat{\theta}_i) = \gamma_i\, \hat{\theta}_i + (1-\gamma_i)\, \boldsymbol{\beta}^\top \mathbf{z}_i. \tag{2.6}$$

On utilisera dans la suite de cet article le nom d'estimateur B pour le meilleur prédicteur $\hat{\theta}_i^B$.

Aux sections 3 et 4, on développe la théorie en supposant que $\boldsymbol{\beta}$, σ_v^2 et $\tilde{\psi}_i$ sont connus. À la section 5, on discute de l'estimation de ces trois quantités ce qui permet d'obtenir une version empirique du meilleur prédicteur et de nos diagnostics.

Remarque : Dans la littérature sur l'estimation sur petits domaines, on développe habituellement la théorie en supposant que ψ_i est fixe. Par conséquent, on fait implicitement l'hypothèse que $\tilde{\psi}_i = \psi_i$. Lorsqu'on fait l'inférence sous le modèle de Fay-Herriot, on ne peut s'attendre à ce que ψ_i soit fixe. Par exemple, considérons le cas où θ_i est une proportion dans le domaine i et où on utilise un plan stratifié aléatoire simple avec remise dont les strates coïncident avec les domaines. L'estimateur direct $\hat{\theta}_i$ est simplement la

proportion échantillonnale dans le domaine i et il est bien connu que sa variance est donnée par $\psi_i = n_i^{-1}\theta_i(1-\theta_i)$. Dans ce cas, il est évident que ψ_i est aléatoire puisqu'il dépend de θ_i. On peut également facilement montrer que $\tilde{\psi}_i = n_i^{-1}\left(\boldsymbol{\beta}^\top \mathbf{z}_i(1-\boldsymbol{\beta}^\top \mathbf{z}_i) - b_i^2\sigma_v^2\right) \neq \psi_i$ sauf si $v_i = \sigma_v = 0$. Dans le reste de cet article, on développe toute la théorie en faisant l'hypothèse habituelle que $\tilde{\psi}_i = \psi_i$. En pratique, ces deux variances sont inconnues et doivent être estimées. À la section 5, on discute de l'estimation de $\tilde{\psi}_i$ au moyen d'un modèle de lissage. On peut facilement montrer que si on dispose d'un estimateur, $\hat{\tilde{\psi}}_i$, sans biais sous le modèle pour $\tilde{\psi}_i$, c'est-à-dire que $\mathbb{E}(\hat{\tilde{\psi}}_i \mid \mathcal{Z}) = \tilde{\psi}_i$, alors cet estimateur est également sans biais sous le modèle pour ψ_i, c'est-à-dire que $\mathbb{E}(\hat{\tilde{\psi}}_i - \psi_i \mid \mathcal{Z}) = 0$. L'inverse est aussi vrai : un estimateur sans biais sous le modèle pour ψ_i sera également sans biais sous le modèle pour $\tilde{\psi}_i$. Par conséquent, même si $\tilde{\psi}_i \neq \psi_i$, on peut estimer ces deux variances par le même estimateur. Cela suggère que l'hypothèse $\tilde{\psi}_i = \psi_i$ n'est peut-être pas si critique en pratique.

3. Les erreurs quadratiques moyennes de l'estimateur direct et de l'estimateur B

Afin d'apprécier l'efficacité de l'estimateur B donné à l'équation (2.6), on se tourne habituellement vers un calcul d'erreur quadratique moyenne. On peut envisager deux possibilités naturelles : soit prendre l'EQM par rapport au plan de sondage, soit prendre l'EQM par rapport au modèle combiné (2.3).

L'EQM fondée sur le modèle de l'estimateur direct $\hat{\theta}_i$ vaut :

$$\mathrm{EQM}_m(\hat{\theta}_i) = \mathbb{E}\left\{(\hat{\theta}_i - \theta_i)^2 \mid \mathcal{Z}\right\} = \tilde{\psi}_i$$

et l'EQM fondée sur le modèle de l'estimateur B vaut :

$$\mathrm{EQM}_m(\hat{\theta}_i^B) = \mathbb{E}\left\{(\hat{\theta}_i^B - \theta_i)^2 \mid \mathcal{Z}\right\} = \gamma_i\tilde{\psi}_i.$$

L'estimateur B est donc toujours plus efficace que l'estimateur direct avec une inférence fondée sur le modèle. Cette propriété résulte de la construction même de l'estimateur B. Par contre, et cela est une question légitime, on peut se demander si l'estimateur B est toujours plus efficace que l'estimateur direct avec une inférence fondée sur le plan de sondage.

Les erreurs quadratiques moyennes fondées sur le plan de sondage des estimateurs direct et B pour le domaine i valent :

$$\begin{aligned}\mathrm{EQM}_p(\hat{\theta}_i) = \mathbb{E}\left\{(\hat{\theta}_i - \theta_i)^2 \mid \Omega\right\} &= \psi_i \\ &= \tilde{\psi}_i \end{aligned} \tag{3.1}$$

et

$$\begin{aligned}\mathrm{EQM}_p(\hat{\theta}_i^B) = \mathbb{E}\left\{(\hat{\theta}_i^B - \theta_i)^2 \mid \Omega\right\} &= \gamma_i^2\psi_i + (1-\gamma_i)^2\, b_i^2 v_i^2 \\ &= \gamma_i\tilde{\psi}_i + (1-\gamma_i)^2 b_i^2(v_i^2 - \sigma_v^2). \end{aligned} \tag{3.2}$$

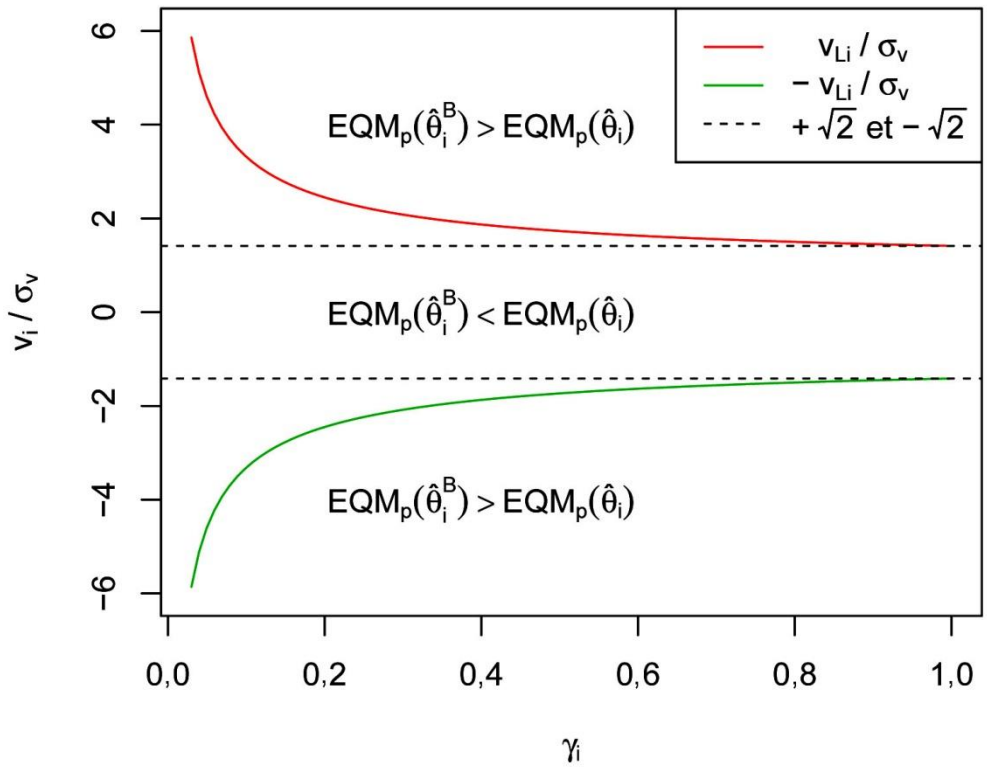
Notons que la deuxième égalité de (3.1) et (3.2) vient de l'hypothèse $\tilde{\psi}_i = \psi_i$. On observe que $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right)$ peut être très différente de $\mathrm{EQM}_m\left(\hat{\theta}_i^B\right)$ quand la valeur inconnue v_i^2 est loin de σ_v^2. Par conséquent, pour un domaine avec une grande valeur de v_i^2, $\mathrm{EQM}_m\left(\hat{\theta}_i^B\right)$ pourrait être significativement plus petite que $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right)$ et mener à une conclusion erronée sur l'efficacité relative de l'estimateur B par rapport à l'estimateur direct.

En remarquant que $\gamma_i \tilde{\psi}_i = (1-\gamma_i) b_i^2 \sigma_v^2$, on peut montrer que $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right) \leq \mathrm{EQM}_p\left(\hat{\theta}_i\right)$ si et seulement si

$$v_i \in \left[-v_{L,i}; v_{L,i}\right],$$

où $v_{L,i} = \sigma_v \sqrt{\frac{1+\gamma_i}{\gamma_i}}$. La figure 3.1 donne les valeurs limites $v_{L,i}/\sigma_v$ et $-v_{L,i}/\sigma_v$ en fonction de γ_i. On note que lorsque $|v_i| \leq \sigma_v \sqrt{2}$, $\mathrm{EQM}_p\left(\hat{\theta}_i^B\right) \leq \mathrm{EQM}_p\left(\hat{\theta}_i\right)$ quelque soit la valeur de γ_i. On note également que l'estimateur direct peut devenir plus intéressant que l'estimateur B pour les domaines où l'effet local est important, particulièrement lorsque γ_i n'est pas petit. Mais comment savoir si l'effet local est important ou non pour un domaine i donné ? C'est l'objet de la section suivante où nous présentons deux diagnostics.

Figure 3.1 Valeurs limites de l'effet local normalisé par σ_v en fonction de γ_i.



4. Deux diagnostics pour évaluer la performance locale de l'estimateur B

4.1 Une approche conditionnelle à l'observation de $\hat{\theta}_i$

A partir de l'expression (2.5) de la section 2 et en notant que $\gamma_i\left(\hat{\theta}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i\right) = b_i\sigma_v\sqrt{\gamma_i}\varepsilon_i,$ on obtient la distribution conditionnelle de v_i :

$$v_i \,|\, \mathcal{Z}, \hat{\theta}_i \sim \mathcal{N}\left(\sigma_v\sqrt{\gamma_i}\varepsilon_i, (1-\gamma_i)\sigma_v^2\right).$$

Le conditionnement sur $\hat{\theta}_i$ permet d'avoir une meilleure idée des valeurs possibles de v_i. En particulier, lorsque γ_i prend une valeur strictement supérieure à 0, la distribution conditionnelle de v_i peut s'écarter nettement de sa distribution non conditionnelle : $v_i \,|\, \mathcal{Z} \sim \mathcal{N}(0, \sigma_v^2)$.

Le premier diagnostic est défini comme la probabilité conditionnelle :

$$\begin{aligned} D_{1i} &= \text{Prob}\left(\text{EQM}_p(\hat{\theta}_i^B) \le \text{EQM}_p(\hat{\theta}_i) \,\middle|\, \mathcal{Z}, \hat{\theta}_i\right) \\ &= \text{Prob}\left(-v_{L,i} \le v_i \le v_{L,i} \,|\, \mathcal{Z}, \hat{\theta}_i\right). \end{aligned} \tag{4.1}$$

Ce diagnostic peut s'écrire en fonction de γ_i et de l'erreur normalisée (2.4) :

$$\begin{aligned} D_{1i} = D_{1i}\left(\gamma_i, |\,\varepsilon_i\,|\right) &= \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\,\varepsilon_i\,| + \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\} \\ &\quad - \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\,\varepsilon_i\,| - \frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}, \end{aligned} \tag{4.2}$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale centrée réduite. La preuve du résultat (4.2) est donnée à l'annexe A.

Quand ce diagnostic prend des valeurs proches de 0, on peut conclure que $|v_i|$ est vraisemblablement plus grand que $v_{L,i}$ et que l'estimateur direct est préférable à l'estimateur B. Pour obtenir une règle de décision associée à ce diagnostic, il est nécessaire de choisir un seuil en dessous duquel on décide de prendre l'estimateur direct et au dessus duquel on prend l'estimateur B. Un seuil à 50 % semble assez naturel. Une autre idée est d'adopter une approche empirique et de repérer une rupture dans la distribution des valeurs du diagnostic D_{1i} pour les m domaines.

Ce diagnostic n'est pas entièrement fondé sur le plan de sondage car il fait intervenir la distribution conditionnelle $v_i \,|\, \mathcal{Z}, \hat{\theta}_i$. Il est donc nécessaire de valider au mieux possible le modèle de Fay-Herriot avant de l'utiliser. Il n'est malheureusement pas possible de valider séparément les hypothèses sur v_i et e_i parce que les valeurs des paramètres $\theta_i, i = 1, \ldots, m,$ ne sont pas observées. On peut toutefois valider le modèle combiné de Fay-Herriot (2.3) au moyen des résidus du modèle (voir, par exemple, Hidiroglou, Beaumont et Yung, 2019). Ces résidus sont obtenus en remplaçant les quantités inconnues dans l'erreur

normalisée (2.4) par leurs estimations (voir section 5). Un graphique des résidus en fonction des valeurs prédites du modèle est souvent suggéré pour valider l'hypothèse de linéarité du modèle. L'hypothèse de normalité de l'erreur $b_i v_i + e_i$ peut être vérifiée par un graphique Q-Q des résidus ou des tests de normalité tels que le test de Shapiro-Wilk. Pour être conservateur, au cas où le modèle ne serait pas complètement satisfaisant, un seuil à 75 % peut être approprié.

Le diagnostic de la section suivante est fondé entièrement sur le plan de sondage. Il ne dépend donc pas de la validité du modèle de liaison. En ce sens, il est considéré plus robuste que le diagnostic (4.2). Il repose toutefois sur les hypothèses sur les erreurs d'échantillonnage e_i, discutées à la section 2, notamment l'hypothèse de normalité des e_i.

4.2 Utilisation d'un test d'hypothèse fondé sur le plan de sondage portant sur le paramètre v_i

Dans le cadre d'une inférence fondée sur le plan de sondage, v_i est fixe et l'erreur normalisée (2.4) suit la loi :

$$\varepsilon_i \,|\Omega \sim \mathcal{N}\left(v_i \frac{\sqrt{\gamma_i}}{\sigma_v}, (1-\gamma_i) \right). \tag{4.3}$$

Nous avons une observation unique de cette variable aléatoire. Nous l'utilisons pour tester si $|v_i|$ est plus grand que $v_{L,i}$. On pose le test :

$$\mathbf{H_0} : |v_i| = v_{L,i} \quad \text{contre} \quad \mathbf{H_1} : |v_i| > v_{L,i}.$$

On choisit d'utiliser $|\varepsilon_i|$ comme statistique de test. On s'attend à ce que $|\varepsilon_i|$ prenne des valeurs plus faibles sous H_0 que sous H_1. Soient $\varepsilon_{\text{obs},i}$, la valeur observée de la statistique ε_i et $P_i(v_i) = \text{Prob}\left(|\varepsilon_i| > |\varepsilon_{\text{obs},i}| \,|\, \Omega; v_i \right)$. La valeur- p du test est définie comme étant la probabillité que la statistique $|\varepsilon_i|$ soit plus grande que la valeur observée $|\varepsilon_{\text{obs},i}|$ sous l'hypothèse nulle. On montre à l'annexe B que la valeur- p vaut :

$$P_i\left(v_{L,i}\right) = P_i\left(-v_{L,i}\right) = \Phi\left(-\tau_i\right) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \right),$$

où

$$\tau_i = \frac{|\varepsilon_{\text{obs},i}| - \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}.$$

Dans la mesure où le second terme est souvent négligeable par rapport au premier terme, en particulier quand $\tau_i > 0$ ou γ_i est grand, on prend comme second diagnostic :

$$D_{2i} = D_{2i}\left(\gamma_i, \left|\varepsilon_{\text{obs},i}\right|\right) = \Phi\left(\frac{\sqrt{1+\gamma_i} - \left|\varepsilon_{\text{obs},i}\right|}{\sqrt{1-\gamma_i}}\right). \tag{4.4}$$

Le second diagnostic s'interprète de la façon suivante. Quand D_{2i} est faible, on peut supposer que $|v_i|$ est vraisemblablement plus grand que $v_{L,i}$ et on préfère alors l'estimateur direct à l'estimateur B. Pour le choix du seuil de décision, on peut s'inspirer des valeurs utilisées habituellement comme niveaux pour les tests (par exemple, 5 % ou 10 %). Avec ces valeurs, compte tenu du sens du test, on favorise l'estimateur B. Comme pour le diagnostic précédent, on peut déterminer la valeur du seuil en repérant une rupture dans la distribution des valeurs du diagnostic D_{2i} pour les m domaines.

4.3 Quelques propriétés des diagnostics 1 et 2

Dans cette section, nous étudions le comportement des fonctions $D_{1i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ et $D_{2i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ pour des cas limites de γ_i et $\left|\varepsilon_i\right|$ et nous notons leurs similarités et différences.

Cas 1 : $0 < \gamma_i < 1$ est fixe et $\left|\varepsilon_i\right| \to \infty$.

À partir des équations (4.2) et (4.4), on peut montrer que pour $\left|\varepsilon_i\right| > 0$, les deux fonctions $D_{1i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ et $D_{2i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ décroissent à mesure que $\left|\varepsilon_i\right|$ croît. Autrement dit, la dérivée de ces fonctions par rapport à $\left|\varepsilon_i\right|$ est négative. De plus, la limite quand $\left|\varepsilon_i\right| \to \infty$ de ces deux fonctions tend vers 0. Pour une valeur de $\left|\varepsilon_i\right|$ suffisamment élevée, les deux diagnostics vont donc privilégier l'estimateur direct.

Cas 2 : $0 < \gamma_i < 1$ est fixe et $\left|\varepsilon_i\right| = 0$.

À partir de l'équation (4.2), on observe que

$$D_{1i}\left(\gamma_i, 0\right) = \Phi\left(\sqrt{\frac{1+\gamma_i}{\gamma_i\left(1-\gamma_i\right)}}\right) - \Phi\left(-\sqrt{\frac{1+\gamma_i}{\gamma_i\left(1-\gamma_i\right)}}\right).$$

On peut montrer que $D_{1i}\left(\gamma_i, 0\right)$ est minimisé quand $\gamma_i = -1+\sqrt{2}$. Par conséquent, $D_{1i}\left(\gamma_i, 0\right) \geq D_{1i}\left(-1+\sqrt{2}, 0\right) =$ 0,98. Puisque cette valeur est près de 1, le diagnostic 1 conduit à choisir l'estimateur B dans ce cas si on prend un seuil à 0,50 ou même à 0,75.

À partir de l'équation (4.4), on obtient :

$$D_{2i}\left(\gamma_i, 0\right) = \Phi\left(\sqrt{\frac{1+\gamma_i}{1-\gamma_i}}\right).$$

On peut montrer que, pour $0 \leq \gamma_i < 1$, la fonction $D_{2i}\left(\gamma_i, 0\right)$ est minimisée quand $\gamma_i = 0$. Ainsi, $D_{2i}\left(\gamma_i, 0\right) \geq D_{2i}\left(0, 0\right) =$ 0,84. Avec un seuil plus petit que 0,50, le diagnostic 2 conduit à la même décision que le diagnostic 1 dans ce cas, c'est-à-dire celle de choisir l'estimateur B.

Cas 3 : $\left|\varepsilon_i\right| < \sqrt{2}$ est fixe et $\gamma_i \to 1$.

Les deux fonctions $D_{1i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ et $D_{2i}\left(\gamma_i, \left|\varepsilon_i\right|\right)$ tendent vers 1 dans ce cas. Les diagnostics 1 et 2 conduisent donc à choisir l'estimateur B.

Cas 4 : $|\varepsilon_i| > \sqrt{2}$ est fixe et $\gamma_i \to 1$.

Les deux fonctions $D_{1i}(\gamma_i, |\varepsilon_i|)$ et $D_{2i}(\gamma_i, |\varepsilon_i|)$ tendent vers 0 dans ce cas. Les diagnostics 1 et 2 conduisent ici à choisir l'estimateur direct.

Cas 5 : $|\varepsilon_i|$ est fixe et $\gamma_i \to 0$.

La fonction $D_{1i}(\gamma_i, |\varepsilon_i|)$ tend vers 1 pour n'importe quelle valeur fixe de $|\varepsilon_i|$. Le diagnostic 1 privilégie donc l'estimateur B pour de petites valeurs de γ_i.

On note que $D_{2i}(0, |\varepsilon_i|) = \Phi(1 - |\varepsilon_i|)$. Par conséquent, contrairement au diagnostic 1, le diagnostic 2 conduira à choisir l'estimateur direct si $|\varepsilon_i|$ est suffisamment grand même quand γ_i est infiniment près de 0. Par exemple, avec un seuil de décision à 0,05 et $\gamma_i = 0$, le diagnostic 2 favorise l'estimateur direct quand $|\varepsilon_i| > 1 - \Phi^{-1}(0{,}05) = 2{,}64$.

Dans les quatre premiers cas ci-dessus, les deux diagnostics conduisent à la même décision. On observe une différence seulement dans le cas 5 où $\gamma_i \to 0$. On s'attend donc à ce que le diagnostic 2 choisisse plus souvent l'estimateur direct que le diagnostic 1 pour de petites valeurs de γ_i. Considérons, par exemple, un seuil à 0,5 pour le diagnostic 1 et à 0,05 pour le diagnostic 2. Pour un seuil à 0,5, on peut montrer que le diagnostic 1 conduit à choisir l'estimateur direct dès que $|\varepsilon_i|$ est supérieur à une valeur approximativement égale à $\frac{\sqrt{1+\gamma_i}}{\gamma_i}$, c'est-à-dire dès que $|\varepsilon_i| \gtrsim \frac{\sqrt{1+\gamma_i}}{\gamma_i}$. Quant au diagnostic 2, pour un seuil à 0,05, il conduit à choisir l'estimateur direct dès que $|\varepsilon_i| > \sqrt{1+\gamma_i} - \sqrt{1-\gamma_i}\,\Phi^{-1}(0{,}05)$. Pour γ_i = 0,01, le diagnostic 1 choisit donc l'estimateur direct quand $|\varepsilon_i| \gtrsim 100{,}5$ alors que le diagnostic 2 choisit l'estimateur direct dès que $|\varepsilon_i| > 2{,}64$. L'écart se rétrécit à mesure que γ_i augmente. Par exemple, pour γ_i = 0,2, le diagnostic 1 choisit l'estimateur direct quand $|\varepsilon_i| \gtrsim 5{,}48$ et le diagnostic 2 choisit l'estimateur direct quand $|\varepsilon_i| > 2{,}57$. La discussion ci-dessus semble suggérer que le diagnostic 2 choisit plus souvent l'estimateur direct que le diagnostic 1. Il existe cependant des cas où le diagnostic 1 choisit l'estimateur direct contrairement au diagnostic 2. Ces cas surviennent généralement pour des valeurs assez grandes de γ_i. Par exemple, pour γ_i = 0,8, le diagnostic 1 choisit l'estimateur direct quand $|\varepsilon_i| \gtrsim 1{,}68$ alors que le diagnostic 2 choisit l'estimateur direct seulement quand $|\varepsilon_i| > 2{,}08$.

5. Version empirique de l'estimateur B et des diagnostics

On a développé la théorie en supposant les paramètres $\boldsymbol{\beta}$, σ_v^2 et $\tilde{\psi}_i$ connus. En pratique, ces quantités sont inconnues et on ne peut utiliser le meilleur prédicteur $\hat{\theta}_i^B$. On peut les remplacer par des estimateurs $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_v^2$ et $\hat{\tilde{\psi}}_i$ pour ainsi obtenir le meilleur prédicteur empirique (estimateur EB) :

$$\hat{\theta}_i^{\mathrm{EB}} = \hat{\gamma}_i\, \hat{\theta}_i + (1 - \hat{\gamma}_i)\, \hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{z}_i,$$

où $\hat{\gamma}_i = \frac{b_i^2 \hat{\sigma}_v^2}{b_i^2 \hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}$.

Dans ce qui suit, on commence d'abord par discuter de l'estimation de $\boldsymbol{\beta}$ en supposant σ_v^2 et $\tilde{\boldsymbol{\psi}} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_m)^{\mathrm{T}}$ connus. On obtient ainsi l'estimateur $\tilde{\boldsymbol{\beta}}(\sigma_v^2, \tilde{\boldsymbol{\psi}})$ de $\boldsymbol{\beta}$. Ensuite, on discute de l'estimation de σ_v^2 en supposant que $\tilde{\boldsymbol{\psi}}$ est connu et on obtient l'estimateur $\tilde{\sigma}_v^2(\tilde{\boldsymbol{\psi}})$ de σ_v^2. Finalement, on discute de l'estimation des variances lissées $\tilde{\psi}_i, i = 1, \ldots, m$. On note $\hat{\tilde{\psi}}_i, i = 1, \ldots, m$, les estimateurs

obtenus et $\hat{\tilde{\boldsymbol{\psi}}} = \left(\hat{\tilde{\psi}}_1, \ldots, \hat{\tilde{\psi}}_m\right)^\top$. En pratique, il faut d'abord commencer par estimer les variances lissées et ensuite on calcule successivement $\hat{\sigma}_v^2 = \tilde{\sigma}_v^2\left(\hat{\tilde{\boldsymbol{\psi}}}\right)$ et $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2, \hat{\tilde{\boldsymbol{\psi}}}\right)$, les estimations de σ_v^2 et $\boldsymbol{\beta}$.

En supposant σ_v^2 et $\tilde{\boldsymbol{\psi}} = \left(\tilde{\psi}_1, \ldots, \tilde{\psi}_m\right)^\top$ connus, l'estimation de $\boldsymbol{\beta}$ peut se faire en utilisant la méthode des moindres carrés généralisés qui est équivalente à la méthode du maximum de vraisemblance sous l'hypothèse d'indépendance et de normalité des erreurs $b_i v_i + e_i$. On obtient :

$$\tilde{\boldsymbol{\beta}}\left(\sigma_v^2, \tilde{\boldsymbol{\psi}}\right) = \left(\sum_{i=1}^m \frac{\mathbf{z}_i \mathbf{z}_i^\top}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}\right)^{-1} \sum_{i=1}^m \frac{\mathbf{z}_i \hat{\theta}_i}{b_i^2 \sigma_v^2 + \tilde{\psi}_i}.$$

Différentes méthodes existent pour estimer σ_v^2. Par exemple, on peut utiliser la méthode des moments de Fay et Herriot (1979) ou encore la méthode du maximum de vraisemblance ou du maximum de vraisemblance restreint. Cette dernière est la plus fréquemment utilisée en pratique. Toutes ces méthodes consistent à résoudre itérativement une équation d'estimation de la forme $g\left(\sigma_v^2, \tilde{\boldsymbol{\psi}}\right) = 0$, où la fonction g dépend de la méthode. On note l'estimateur qui en résulte par $\tilde{\sigma}_v^2\left(\tilde{\boldsymbol{\psi}}\right)$. Rao et Molina (2015, Chapitres 5 et 6) fournissent plus de détails sur l'estimation de $\boldsymbol{\beta}$ et σ_v^2 et sur les propriétés des estimateurs telles que la convergence sous le modèle.

Avant d'estimer σ_v^2 et $\boldsymbol{\beta}$ par $\hat{\sigma}_v^2 = \tilde{\sigma}_v^2\left(\hat{\tilde{\boldsymbol{\psi}}}\right)$ et $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}\left(\hat{\sigma}_v^2, \hat{\tilde{\boldsymbol{\psi}}}\right)$, il faut d'abord estimer la variance lissée $\tilde{\psi}_i = \mathbb{E}\left(\psi_i \,|\, \mathcal{Z}\right), i = 1, \ldots, m.$ On suppose qu'on dispose d'un estimateur sans biais sous le plan, $\hat{\psi}_i$, c'est-à-dire que $\mathbb{E}\left(\hat{\psi}_i \,|\, \Omega\right) = \psi_i$. Sous cette hypothèse, on observe que $\mathbb{E}\left(\hat{\psi}_i \,|\, \mathcal{Z}\right) = \tilde{\psi}_i$. L'estimateur $\hat{\psi}_i$ est donc sans biais pour la variance lissée $\tilde{\psi}_i$ mais peut être très instable quand n_i est petite. En général, on préfère plutôt modéliser $\hat{\psi}_i$ en fonction de $\mathbf{z}_i$ pour obtenir un estimateur moins variable. Le modèle de lissage suivant est fréquemment utilisé en pratique :

$$\log\left(\hat{\psi}_i\right) = \boldsymbol{\alpha}^\top \mathbf{x}_i + \eta_i,$$

où $\mathbf{x}_i$ est une fonction de $\mathbf{z}_i$, $\boldsymbol{\alpha}$ est un vecteur de paramètres du modèle et $\eta_i, i = 1, \ldots, m,$ sont des erreurs indépendantes et identiquement distribuées avec une moyenne égale à 0 et une variance égale à σ_η^2. On peut facilement montrer que

$$\tilde{\psi}_i = \mathbb{E}\left(\hat{\psi}_i \,|\, \mathcal{Z}\right) = \exp\left(\boldsymbol{\alpha}^\top \mathbf{x}_i\right) \Delta,$$

où $\Delta = \mathbb{E}\left\{\exp\left(\eta\right)\right\}$ et η est une variable aléatoire qui suit la même distribution que le terme d'erreur du modèle de lissage ci-dessus. On estime $\boldsymbol{\alpha}$ par un estimateur convergent sous le modèle, $\hat{\boldsymbol{\alpha}}$, en utilisant la méthode des moindres carrés. Hidiroglou, Beaumont et Yung (2019) suggèrent d'estimer Δ par un estimateur convergent sous le modèle, $\hat{\Delta}$, en utilisant une méthode des moments. L'estimateur de la variance lissée s'écrit :

$$\hat{\tilde{\psi}}_i = \exp\left(\hat{\boldsymbol{\alpha}}^\top \mathbf{x}_i\right) \hat{\Delta},$$

où

$$\hat{\Delta} = \frac{\sum_{i=1}^m \hat{\psi}_i}{\sum_{i=1}^m \exp\left(\hat{\boldsymbol{\alpha}}^\top \mathbf{x}_i\right)}.$$

On peut s'attendre à ce que l'EQM fondée sur le plan de l'estimateur EB,

$$\text{EQM}_p(\hat{\theta}_i^{\text{EB}}) = \mathbb{E}\left\{(\hat{\theta}_i^{\text{EB}} - \theta_i)^2 \,\middle|\, \Omega\right\},$$

soit supérieure à l'EQM fondée sur le plan de l'estimateur B donné à l'équation (3.2). Tel que mentionné ci-dessus, les estimateurs des paramètres $\boldsymbol{\beta}$, σ_v^2, $\boldsymbol{\alpha}$ et Δ sont convergents sous le modèle à mesure que m augmente pourvu que certaines conditions de régularité soient satisfaites. On note également que l'erreur quadratique moyenne fondée sur le plan de l'estimateur B (voir l'équation 3.2) ne dépend pas de m. Par conséquent, on peut s'attendre à ce que l'augmentation de l'erreur quadratique moyenne résultant de l'estimation de ces paramètres soit modeste quand le nombre de domaines est grand. Cela suggère que la dérivation de la borne $v_{L,i}$ sera peu affectée par l'estimation des paramètres $\boldsymbol{\beta}$, σ_v^2, $\boldsymbol{\alpha}$ et Δ si m est grand. Ainsi, nos deux diagnostics (4.2) et (4.4) devraient rester pertinents même si c'est l'estimateur EB qui est utilisé plutôt que l'estimateur B. Il faut cependant remplacer γ_i par $\hat{\gamma}_i$ et ε_i par

$$\hat{\varepsilon}_i = \frac{\hat{\theta}_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{z}_i}{\sqrt{b_i^2\hat{\sigma}_v^2 + \hat{\tilde{\psi}}_i}}$$

dans les expressions (4.2) et (4.4) pour être en mesure de calculer ces diagnostics avec des données réelles. On obtient ainsi $\hat{D}_{1i}$, l'estimateur de D_{1i}, et $\hat{D}_{2i}$, l'estimateur de D_{2i}.

6. Étude par simulation

On réalise une étude par simulation pour évaluer l'efficacité de $\hat{D}_{1i}$ et $\hat{D}_{2i}$ à détecter lequel des estimateurs direct et EB est préférable. On s'intéresse à $m = 140$ domaines qui représentent des villes canadiennes. Le vecteur de variables auxiliaires est : $\mathbf{z}_i^{\mathrm{T}} = (1, z_{1i})$. La variable auxiliaire z_{1i} est obtenue à partir de fichiers administratifs et est définie comme le ratio du nombre de bénéficiaires d'assurance-emploi dans la ville i sur le nombre de personnes de plus de 15 ans dans la ville i. La taille de l'échantillon dans la ville i, n_i, est obtenue au moyen de l'Enquête canadienne sur la population active (EPA). Parmi les 140 domaines, 2 ont une taille d'échantillon plus petite que 10, 10 ont une taille d'échantillon plus petite que 30, 40 ont une taille d'échantillon plus petite que 60 et 68 ont une taille d'échantillon plus petite que 100, soit tout près de 50 % des domaines. Pour ces 68 domaines, les coefficients de variation estimés du taux de chômage de l'EPA sont la plupart du temps trop élevés pour pouvoir publier les estimations directes du taux de chômage; des techniques d'estimation sur petits domaines sont donc requises pour ces domaines. À l'opposé, il y a également 17 domaines parmi les 140 qui ont une taille d'échantillon plus grande que 1 000 et pour lesquels l'estimation directe du taux de chômage est fiable.

À partir des valeurs réelles de n_i et $\mathbf{z}_i$, on simule le paramètre de population θ_i pour les m domaines. Le paramètre θ_i peut être interprété comme la proportion de personnes sans emploi dans le domaine i. On génère θ_i selon la loi béta de moyenne $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}_i$ et de variance σ_v^2, où $\sigma_v^2 = 7{,}58 \times 10^{-5}$ et $\boldsymbol{\beta}^{\mathrm{T}} = (0{,}0484;\ 0{,}95)$. Ces valeurs de $\boldsymbol{\beta}$ et σ_v^2 ont été choisies à partir de données réelles. On pose $b_i = 1, i = 1, \ldots, m$.

Ensuite, on change manuellement les valeurs de θ_i pour 4 domaines (villes) de manière à avoir un effet local v_i égal à $5\sigma_v$. On a choisi des domaines avec des tailles d'échantillon différentes : 10, 100, 501 et 3 773. Dans le reste de cette section, on identifiera la plus petite de ces 4 villes par la ville 1 $(n_i = 10)$, la deuxième plus petite par la ville 2 $(n_i = 100)$, la deuxième plus grande par la ville 3 $(n_i = 501)$ et la plus grande par la ville 4 (n_i = 3 773).

On considère un plan de sondage stratifié aléatoire simple avec remise dont les strates coïncident avec les domaines. L'estimateur direct $\hat{\theta}_i$ de θ_i est simplement la proportion de personnes échantillonnées dans le domaine i qui ont la caractéristique d'intérêt (par exemple, être sans emploi). Sous un tel plan simple, il est facile de constater que l'estimateur direct peut être généré comme suit : $\hat{\theta}_i = n_i^{-1}$ Binomiale (n_i, θ_i). Il n'est donc pas nécessaire de créer la population de personnes dans le domaine i pour générer $\hat{\theta}_i$. On a procédé de cette façon dans la simulation. La variance sous le plan de $\hat{\theta}_i$ est donnée par $\psi_i = n_i^{-1}\theta_i(1-\theta_i)$ et son estimateur par $\hat{\psi}_i = (n_i - 1)^{-1}\hat{\theta}_i(1-\hat{\theta}_i)$. On estime la variance lissée $\tilde{\psi}_i$ en utilisant le modèle de lissage de la section 5 avec $\mathbf{x}_i^{\top} = (1, \log(z_{1i}), \log(1-z_{1i}), \log(n_i))$.

Afin de simuler un scénario réaliste, nous n'avons pas respecté exactement toutes les hypothèses du modèle de Fay-Herriot. Entre autres, les erreurs v_i et e_i ne sont pas distribuées selon des lois normales. Nous avons opté pour une loi béta pour générer θ_i; l'hypothèse de normalité des v_i n'est donc pas satisfaite bien que la déviation de la loi normale ne soit pas drastique dans notre simulation. Les estimations $\hat{\theta}_i$ ont été générées selon une loi binomiale qui peut être approchée par une loi normale pour les domaines qui ont une grande valeur de n_i. La relation entre les estimations $\hat{\theta}_i$ simulées et les $\mathbf{z}_i$ ressemble à celle qu'on observe avec les estimations réelles de l'EPA. De plus, notre scénario de simulation est tel que l'hypothèse $\tilde{\psi}_i = \psi_i$ n'est pas respectée puisque, pour ce plan simple,

$$\tilde{\psi}_i = n_i^{-1}\left(\boldsymbol{\beta}^{\top}\mathbf{z}_i(1-\boldsymbol{\beta}^{\top}\mathbf{z}_i) - b_i^2\sigma_v^2\right)$$

(voir la remarque de la section 2). On note toutefois que le coefficient de corrélation entre $\tilde{\psi}_i$ et ψ_i est égal à 0,98 ce qui indique que la déviation de l'hypothèse $\tilde{\psi}_i = \psi_i$ est modeste. Tel que mentionné au paragraphe précédent, on utilise le modèle de lissage de la section 5 pour estimer $\tilde{\psi}_i$. Cela permet de rester dans un cadre réaliste où le modèle de lissage postulé est différent du vrai modèle qui a été utilisé pour générer les estimations $\hat{\psi}_i$.

On effectue une simulation fondée sur le plan de sondage, c'est-à-dire que les paramètres de population $\theta_i, i = 1, \ldots, m,$ ne sont générés qu'une seule fois. On répète K = 10 000 fois l'échantillonnage. Pour chaque itération $k, k = 1, \ldots, K,$ on génère une estimation directe $\hat{\theta}_i(k)$ et on calcule une estimation de la variance lissée $\hat{\tilde{\psi}}_i(k)$ tel que décrit ci-dessus. On calcule ensuite l'estimation EB :

$$\hat{\theta}_i^{\mathrm{EB}}(k) = \hat{\gamma}_i(k)\,\hat{\theta}_i(k) + \left(1-\hat{\gamma}_i(k)\right)\hat{\mathbf{\beta}}(k)^{\top}\mathbf{z}_i,$$

où $\hat{\gamma}_i(k) = \frac{\hat{\sigma}_v^2(k)}{\hat{\sigma}_v^2(k)+\hat{\tilde{\psi}}_i(k)}$ et $\hat{\boldsymbol{\beta}}(k)$ et $\hat{\sigma}_v^2(k)$ sont calculés tels que décrits à la section 5. On utilise la méthode des moindres carrés généralisés pour obtenir $\hat{\boldsymbol{\beta}}(k)$ et la méthode du maximum de vraisemblance restreint

pour $\hat{\sigma}_v^2(k)$. Les calculs des estimations ont été réalisés avec le système d'estimation sur petits domaines de Statistique Canada (Hidiroglou, Beaumont et Yung, 2019).

Pour chaque itération, on calcule également les résidus normalisés $\hat{\varepsilon}_i(k)$ et les diagnostics $\hat{D}_{1i}(k)$ et $\hat{D}_{2i}(k)$ pour les m domaines. On enregistre si on a préféré l'estimateur direct à l'estimateur EB pour chacun des deux diagnostics. On utilise pour ce faire des seuils de décision. En deçà des seuils, on prend l'estimateur direct. Pour le diagnostic 1, on a retenu des seuils de 50 % et 75 % et pour le diagnostic 2, des seuils de 5 % et 25 %.

À partir des quantités précédentes, calculées pour chacune des 10 000 itérations, on calcule les moyennes Monte Carlo des diagnostics 1 et 2 pour les m domaines : $\bar{\hat{D}}_{1i}$ et $\bar{\hat{D}}_{2i}$. On calcule aussi le taux de sélection de l'estimateur direct pour chacun des deux diagnostics, c'est-à-dire le pourcentage de fois où un diagnostic donné a conduit à choisir l'estimateur direct.

On calcule l'approximation Monte Carlo de $\mathrm{EQM}_p\left(\hat{\theta}_i^{\mathrm{EB}}\right)$:

$$\mathrm{EQM}_{\mathrm{MC}}(\hat{\theta}_i^{\mathrm{EB}}) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_i^{\mathrm{EB}}(k) - \theta_i\right)^2.$$

A partir de cette dernière quantité, on calcule l'efficacité relative de l'estimateur EB :

$$\frac{\mathrm{EQM}_{\mathrm{MC}}(\hat{\theta}_i^{\mathrm{EB}}) - \psi_i}{\psi_i}. \tag{6.1}$$

Ce ratio est positif lorsque l'estimateur EB est moins efficace que l'estimateur direct sous le plan de sondage. Un diagnostic est potentiellement performant s'il est corrélé négativement à ce ratio.

Les figures 6.1 et 6.2 présentent les moyennes Monte Carlo des diagnostics 1 et 2 respectivement en fonction de l'efficacité relative de l'estimateur EB définie à l'équation (6.1). Les quatre villes dont les valeurs de v_i ont été modifiées, les villes 1 à 4, sont colorées en *violet*, *orange*, *vert* et *rouge*. Dans la légende, on a indiqué la taille de l'échantillon dans ces villes. Les valeurs du paramètre γ_i pour les villes 1 à 4 valent respectivement : 0,01; 0,08; 0,35 et 0,81. Toutes les autres villes sont colorées en bleu.

On observe d'abord, sur les figures 6.1 et 6.2, que l'estimateur EB est plus efficace que l'estimateur direct pour la ville 1 (en violet) puisque cette ville se retrouve à gauche de la ligne verticale (efficacité relative négative) et cela en dépit du fort effet local. L'explication de ce phénomène se trouve à la figure 3.1. On y constate que l'étendue des valeurs de v_i pour lesquelles l'estimateur B est plus efficace que l'estimateur direct augmente à mesure que γ_i diminue. Puisque γ_i est petit pour la ville 1 $(\gamma_i = 0{,}01)$, il n'est pas surprenant d'observer une efficacité relative négative malgré un effet local prononcé. Pour la ville 2 (en orange), l'estimateur direct est légèrement plus efficace que l'estimateur EB. Par contre, pour les villes 3 (en vert) et 4 (en rouge), l'estimateur direct est nettement plus efficace. On note également qu'il y a 5 villes pour lesquelles l'estimateur direct est plus efficace que l'estimateur EB : les villes 2 à 4 de même que deux autres villes dont les valeurs de v_i ont été générées aléatoirement et n'ont pas été modifiées manuellement. Une de ces villes a la plus petite valeur de v_i et l'autre a la plus

grande valeur de v_i après celle des 4 villes modifiées manuellement. Ces deux villes ont également de grandes valeurs de γ_i (0,62 et 0,49).

Les figures 6.1 et 6.2 indiquent que nos deux diagnostics semblent assez performants à déceler les cas où l'estimateur direct est plus efficace que l'estimateur EB sauf pour la ville 2 $\left(n_i = 100\right)$ où la moyenne Monte Carlo du diagnostic 1 est très élevée à 0,97. C'est toutefois un domaine pour lequel choisir l'estimateur le moins efficace n'est pas dramatique puisqu'il y a très peu de différence entre les efficacités des deux estimateurs. Outre ce cas spécifique, le diagnostic 1 semble avoir un meilleur comportement que le diagnostic 2. La moyenne Monte Carlo du diagnostic 1 est tout près de 1 quand l'estimateur EB est significativement plus efficace que l'estimateur direct, diminue lentement quand les efficacités des deux estimateurs se rapprochent et devient petite quand l'estimateur direct est significativement plus efficace que l'estimateur EB. On n'observe pas tout-à-fait le même comportement pour le diagnostic 2. La moyenne Monte Carlo du diagnostic 2 est petite quand l'estimateur direct est significativement plus efficace que l'estimateur EB mais elle n'est pas près de 1 quand l'estimateur EB est nettement plus efficace que l'estimateur direct. Surtout, elle semble augmenter quand les efficacités des deux estimateurs se rapprochent ce qui est contre-intuitif.

Figure 6.1 Moyenne Monte Carlo des estimations du diagnostic 1 pour les 140 domaines en fonction de l'efficacité relative de l'estimateur EB.

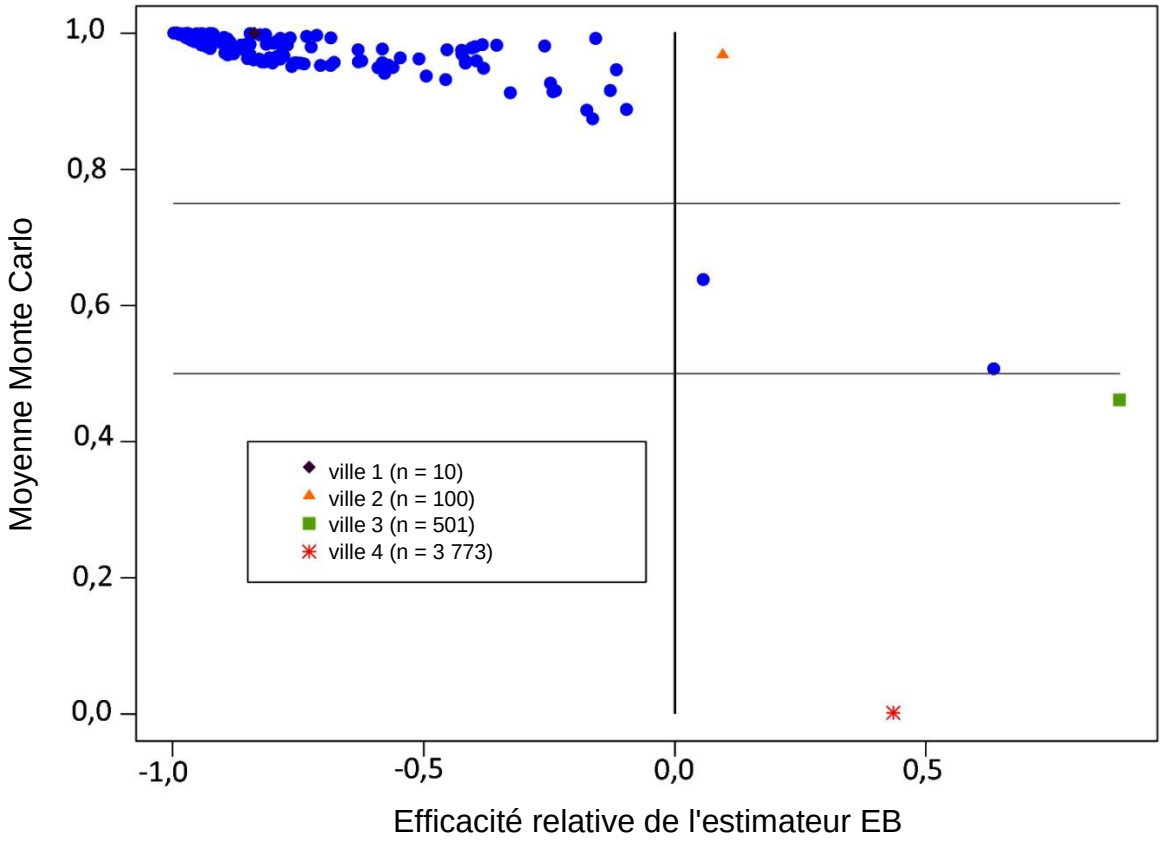
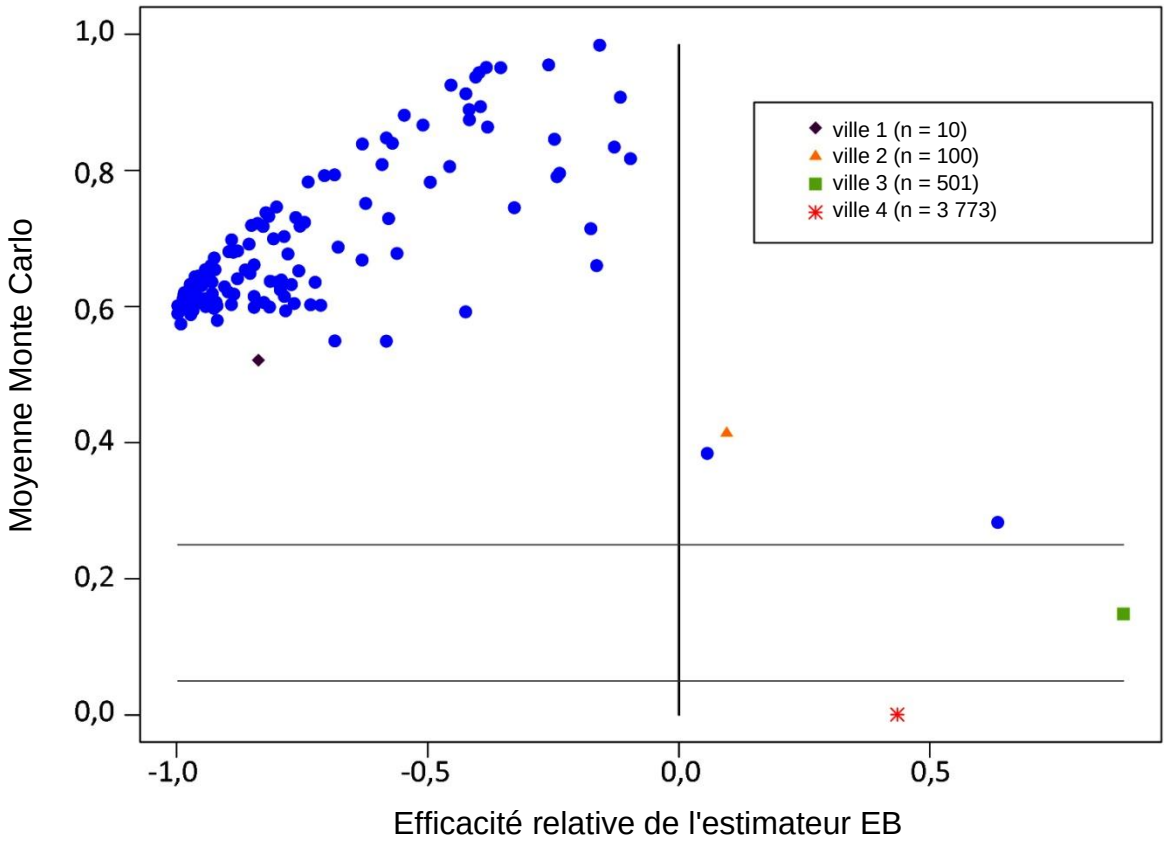


Figure 6.2 Moyenne Monte Carlo des estimations du diagnostic 2 pour les 140 domaines en fonction de l'efficacité relative de l'estimateur EB.



Les figures 6.3 et 6.4 montrent le taux de sélection de l'estimateur direct sur les 10 000 répétitions pour les diagnostics 1 et 2. On peut tirer des conclusions similaires à celles obtenues en analysant les figures 6.1 et 6.2. Comme on s'y attend, les seuils de 75 % pour le diagnostic 1 et de 25 % pour le diagnostic 2 permettent de mieux détecter les cas où l'estimateur direct est plus efficace que l'estimateur EB mais ces seuils conduisent également à choisir un peu trop souvent l'estimateur direct alors qu'il était moins efficace que l'estimateur EB. Cette constatation est particulièrement notable pour le diagnostic 2. On peut réduire cette erreur en réduisant les seuils mais on détectera alors moins souvent les situations où l'estimateur direct était plus efficace que l'estimateur EB. Tel que noté précédemment, le diagnostic 1 semble avoir un comportement plus désirable que le diagnostic 2, peu importe les seuils choisis, avec des taux de sélection de l'estimateur direct très petits quand l'estimateur direct est nettement moins efficace que l'estimateur EB. Cela semble montrer les limites d'une approche entièrement fondée sur le plan de sondage, telle que celle présentée à la section 4.2, pour faire face au défi posé par l'observation de petites tailles d'échantillon dans les domaines.

Figure 6.3 Taux de sélection de l'estimateur direct pour le diagnostic 1.

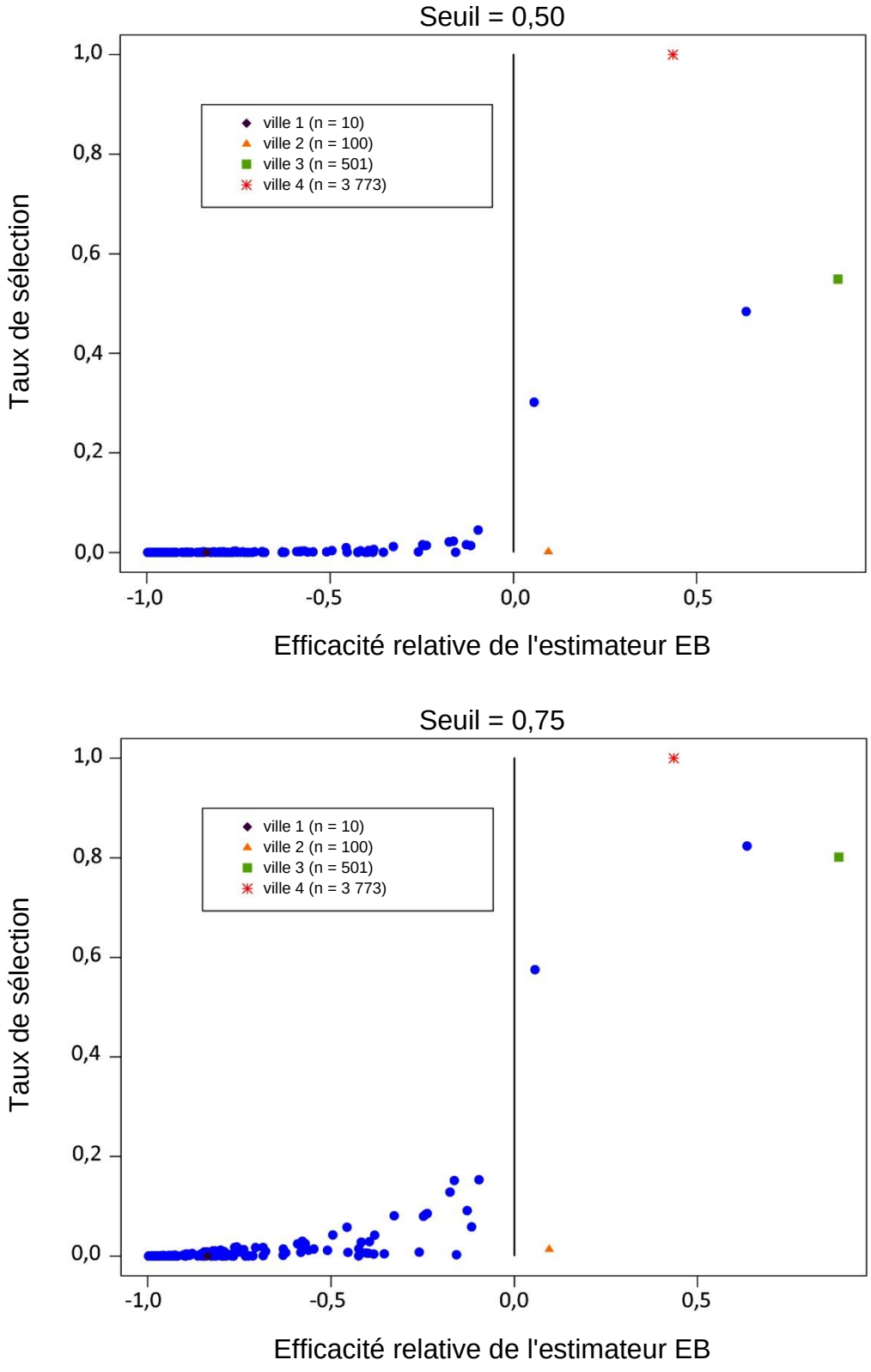
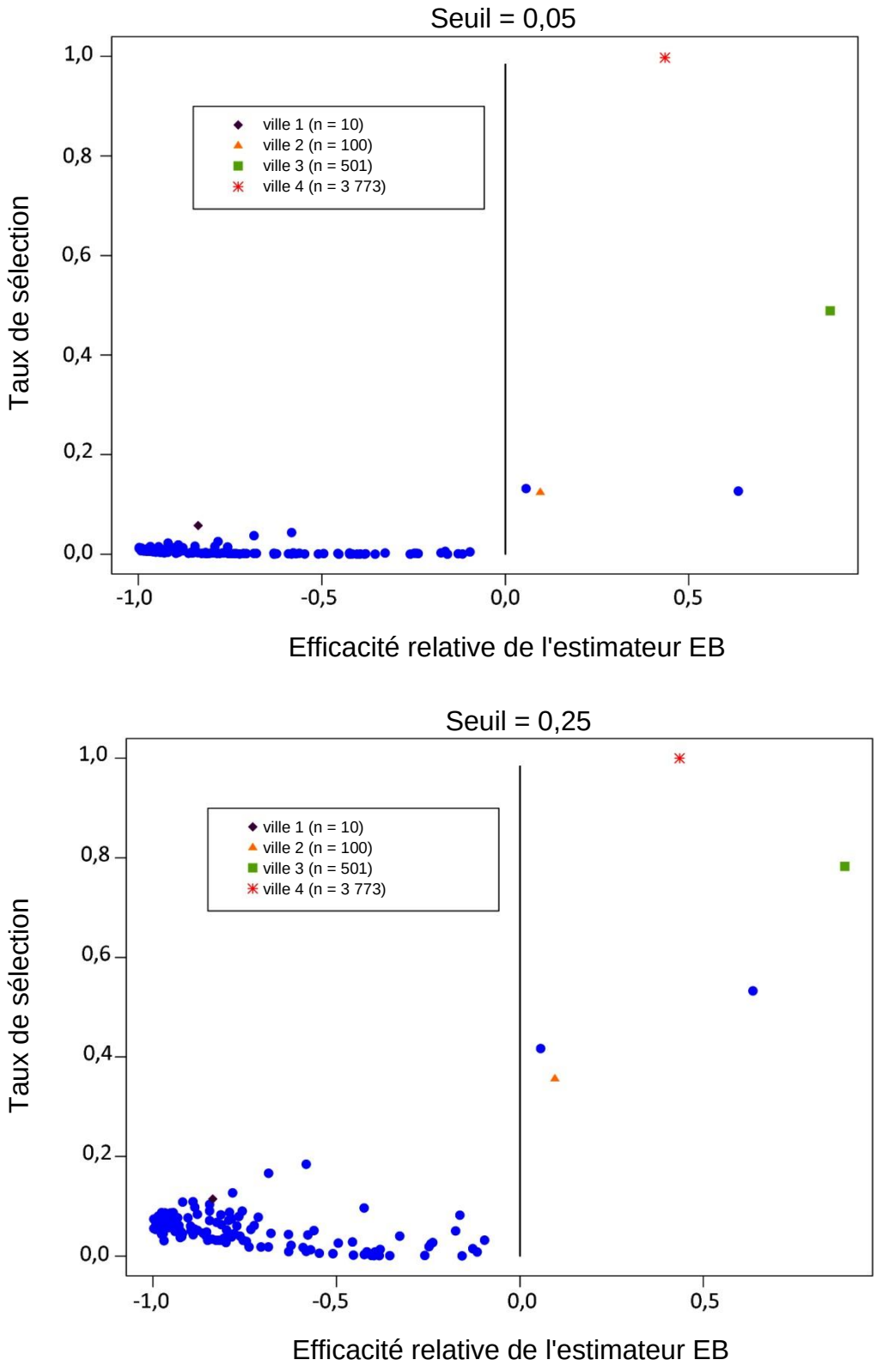


Figure 6.4 Taux de sélection de l'estimateur direct pour le diagnostic 2.



7. Conclusion

Les utilisateurs d'estimations sur petits domaines ne s'intéressent en général qu'à un unique domaine. Ils désirent donc avoir un indicateur de qualité qui s'applique à leur domaine et non un indicateur global. L'EQM fondée sur le plan de sondage des estimateurs sur petits domaines est un indicateur de qualité conceptuellement attrayant puisqu'il conditionne sur l'effet local inexpliqué du modèle. Il est cependant connu que les estimateurs sans biais de l'EQM fondée sur le plan de sondage sont généralement instables quand la taille de l'échantillon dans le domaine est petite. Pour contourner ce problème, nous avons proposé deux diagnostics qui ont pour but de déceler les domaines pour lesquels l'EQM fondée sur le plan de sondage de l'estimateur direct est plus petite que celle de l'estimateur EB. Nos résultats de simulation semblent prometteurs et permettent d'envisager la mise au point d'un indicateur utile pour choisir entre l'estimateur direct et l'estimateur EB pour un domaine en particulier. Dans la recherche future, il serait intéressant d'évaluer l'efficacité d'un estimateur hybride qui s'appuierait sur ces diagnostics.

Annexe

A. Preuve de l'équivalence entre les équations (4.1) et (4.2)

À partir de l'équation (4.1) et de la distribution conditionnelle de v_i donnée à la section 4.1, on a :

$$\begin{aligned} D_{1i} &= \text{Prob}\left(-v_{L,i} \le v_i \le v_{L,i} \middle| \mathcal{Z}, \hat{\theta}_i\right) \\ &= \text{Prob}\left(\frac{-v_{L,i} - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{v_i - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{v_{L,i} - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \middle| \mathcal{Z}, \hat{\theta}_i \right). \end{aligned}$$

En remplaçant $v_{L,i}$ par $\sigma_v\sqrt{\frac{1+\gamma_i}{\gamma_i}}$, on obtient :

$$\begin{aligned} D_{1i} &= \text{Prob}\left(\frac{-\sqrt{(1+\gamma_i)/\gamma_i} - \sqrt{\gamma_i}\varepsilon_i}{\sqrt{1-\gamma_i}} \le \frac{v_i - \sigma_v\sqrt{\gamma_i}\varepsilon_i}{\sigma_v\sqrt{1-\gamma_i}} \le \frac{\sqrt{(1+\gamma_i)/\gamma_i} - \sqrt{\gamma_i}\varepsilon_i}{\sqrt{1-\gamma_i}} \middle| \mathcal{Z}, \hat{\theta}_i \right) \\ &= \Phi\left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(-\varepsilon_i + \sqrt{\frac{1+\gamma_i}{\gamma_i}} \right) \right\} - \Phi\left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(-\varepsilon_i - \sqrt{\frac{1+\gamma_i}{\gamma_i}} \right) \right\}. \end{aligned}$$

Puisque pour n'importe quelle valeur t, on a $\Phi(t) = 1 - \Phi(-t)$ alors

$$D_{1i} = \Phi\left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(\varepsilon_i + \sqrt{\frac{1+\gamma_i}{\gamma_i}} \right) \right\} - \Phi\left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(\varepsilon_i - \sqrt{\frac{1+\gamma_i}{\gamma_i}} \right) \right\}.$$

On remarque que D_{1i} est une fonction symétrique de ε_i autour de 0, c'est-à-dire que $D_{1i}(\varepsilon_i) = D_{1i}(-\varepsilon_i)$. Par conséquent, on peut ré-écrire D_{1i} comme à l'équation (4.2) :

$$D_{1i} = \Phi\left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(|\varepsilon_i| + \sqrt{\frac{1+\gamma_i}{\gamma_i}} \right) \right\} - \Phi\left\{ \sqrt{\frac{\gamma_i}{1-\gamma_i}} \left(|\varepsilon_i| - \sqrt{\frac{1+\gamma_i}{\gamma_i}} \right) \right\}.$$

B. Valeur-p associée à la statistique de test $\left|\varepsilon_i\right|$

Rappelons d'abord que $P_i(v_i) = \text{Prob}\left(|\varepsilon_i| > |\varepsilon_{\text{obs},i}| \,\middle|\, \Omega; v_i\right)$. On définit la valeur-$p$ comme le maximum de $P_i(v_{L,i})$ et $P_i(-v_{L,i})$. Puisque que $\tau_i = \frac{|\varepsilon_{\text{obs},i}| - \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}$, on peut écrire :

$$\begin{aligned} P_i(v_i) &= \text{Prob}\left(|\varepsilon_i| > \sqrt{1+\gamma_i} + \sqrt{1-\gamma_i}\,\tau_i \,\middle|\, \Omega; v_i\right) \\ &= \text{Prob}\left(\varepsilon_i > \sqrt{1+\gamma_i} + \sqrt{1-\gamma_i}\,\tau_i \,\middle|\, \Omega; v_i\right) \\ &\quad + \text{Prob}\left(\varepsilon_i < -\sqrt{1+\gamma_i} - \sqrt{1-\gamma_i}\,\tau_i \,\middle|\, \Omega; v_i\right). \end{aligned}$$

En utilisant la loi du résidu normalisé (4.3), on obtient :

$$\begin{aligned} P_i(v_i) &= \text{Prob}\left(\frac{\varepsilon_i - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} > \frac{\sqrt{1+\gamma_i} - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} + \tau_i \,\middle|\, \Omega; v_i\right) \\ &\quad + \text{Prob}\left(\frac{\varepsilon_i - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} < \frac{-\sqrt{1+\gamma_i} - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i \,\middle|\, \Omega; v_i\right) \\ &= \Phi\left(\frac{-\sqrt{1+\gamma_i} + v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i\right) \\ &\quad + \Phi\left(\frac{-\sqrt{1+\gamma_i} - v_i\sqrt{\gamma_i}/\sigma_v}{\sqrt{1-\gamma_i}} - \tau_i\right). \end{aligned}$$

En utilisant l'expression $v_{L,i} = \sigma_v\sqrt{\frac{1+\gamma_i}{\gamma_i}}$, on a :

$$\begin{aligned} P_i(v_i) &= \Phi\left(-\tau_i + \frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\left[\frac{v_i}{v_{L,i}} - 1\right]\right) \\ &\quad + \Phi\left(-\tau_i + \frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\left[-\frac{v_i}{v_{L,i}} - 1\right]\right). \end{aligned}$$

Sous l'hypothèse nulle H_0, $v_i = v_{L,i}$ ou $v_i = -v_{L,i}$ et dans les deux cas l'équation ci-dessus se réduit à :

$$P_i(v_{L,i}) = P_i(-v_{L,i}) = \Phi(-\tau_i) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\right).$$

On va maintenant montrer que si on rejette H_0 (avec un seuil plus petit que 0,5 tel que 0,1) alors on rejetterait encore plus fortement l'hypothèse nulle $H_0^*: |v_i| = v_i^*$ pour n'importe quelle valeur $0 \le v_i^* < v_{L,i}$. Premièrement, si $\tau_i \le 0$, c'est-à-dire que $|\varepsilon_{\text{obs},i}| \le \sqrt{(1+\gamma_i)}$, on observe que $P_i(v_{L,i}) = P_i(-v_{L,i}) \ge 0{,}5$ et on ne rejette jamais l'hypothèse H_0. Deuxièmement, si $\tau_i > 0$, on peut facilement montrer que la fonction $P_i(v_i)$ est croissante en v_i sur l'intervalle $\left[0, v_{L,i}\right]$. On note également que c'est une fonction de v_i qui est symétrique autour de l'abscisse $v_i = 0$ puisque $P_i(v_i) = P_i(-v_i)$. Par conséquent, $P_i(v_i)$ est

décroissante sur l'intervalle $\left[-v_{L,i}, 0\right]$, est minimale en $v_i = 0$ et maximale en $v_i = v_{L,i}$ et $v_i = -v_{L,i}$. Lorsque $|v_i| < v_{L,i}$, on a donc :

$$P_i(v_i) < P_i(v_{L,i}) = \Phi(-\tau_i) + \Phi\left(-\tau_i - 2\frac{\sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}}\right).$$

Bibliographie

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.

Hidiroglou, M.A., Beaumont, J.-F. et Yung, W. (2019). Élaboration d'un système d'estimation sur petits domaines à Statistique Canada. *Techniques d'enquête*, 45, 1, 107-133. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2019001/article/00009-fra.pdf.

Pfeffermann, D., et Ben-Hur, D. (2019). Estimation of randomisation mean square error in small area estimation. *Revue Internationale de Statistique*, 87, S1, S31-S49.

Rao, J.N.K., et Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Rao, J.N.K., Rubin-Bleuer, S. et Estevao, V.M. (2018). Mesure de l'incertitude associée aux estimateurs pour petits domaines basés sur un modèle. *Techniques d'enquête*, 44, 2, 163-180. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2018002/article/54958-fra.pdf.

Rivest, L.-P., et Belmonte, E. (2000). Une erreur quadratique moyenne conditionnelle des estimateurs régionaux. *Techniques d'enquête*, 26, 1, 79-90. Article accessible à l'adresse https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2000001/article/5179-fra.pdf.