Quelques diagnostics locaux pour le modèle de Fay-Herriot

Éric Lesage[1], Jean-François Beaumont[2] & Cynthia Bocci[3]

[1] *INSEE, eric.lesage@insee.fr*
[2] *Statistique Canada, jean-francois.beaumont@canada.ca*
[3] *Statistique Canada, cynthia.bocci@canada.ca*

Résumé.
Les besoins croissants d'information conduit à utiliser les enquêtes pour produire des estimateurs statistiques sur des domaines (par exemple des régions) de tailles très variables. Ainsi, dans le cadre d'une même enquête, la taille des échantillons dans les domaines peut aller de quelques unités à plusieurs centaines. Dans la mesure où les estimateurs directs (Horvitz-Thompson ou calage) manquent de précision sur les petits domaines, il est préférable de recourir à des estimateurs "petits domaines" tel que l'estimateur composite de Fay-Herriot (Fay et Herriot, 1979).

Il existe des instruments statistiques destinés à évaluer si l'estimation composite est globalement meilleure que l'estimation directe. On peut ainsi utiliser des tests statistiques standard pour juger de la validité du modèle de Fay-Herriot. Toutefois, il n'existe pas à notre connaissance de diagnostics locaux qui permettent de déterminer, domaine par domaine, quel estimateur est préférable. L'erreur quadratique moyenne, basée sur le modèle, pourrait être considérée, d'une certaine façon, comme un diagnostic local à la condition forte d'admettre que le modèle est valable pour tous les domaines.

Or, certains utilisateurs des statistiques regardent un domaine spécifique et ne sont pas intéressés par un critère global de qualité pour juger de la précision de l'estimateur composite qui porte sur leur unique domaine. Ces utilisateurs sont davantage séduits par une inférence conditionnelle aux paramètres d'intérêt de leur domaine, telle que l'inférence basée sur le plan de sondage, par opposition à l'inférence sous le modèle.

Nous avons utilisé cette approche conditionnelle pour produire des diagnostics locaux, pour chaque domaine, qui indiquent si l'estimateur composite est susceptible d'être plus précis que l'estimateur direct. Nous avons trouvé que, dépendamment de la taille de l'échantillon et de l'importance du résidu standardisé du modèle, il est possible de détecter si l'estimateur composite est susceptible d'avoir une erreur quadratique moyenne sous le plan plus petite que l'estimateur direct.

Abstract. The increasing need of information leads to the production of survey estimates for domains (e.g., regions) of various sizes. Thus, within the same survey, domain sample

sizes can range from a couple of units to more than thousand units. As direct estimators (Horvitz-Thompson or calibration estimators) suffer from a lack of precision in small domains, it may be desirable to use SAE estimators such as the composite estimator proposed by Fay and Herriot (1979). There exist tools to assess if the composite estimator is globally better than the direct es- timator. For instance, standard model diagnostics can be used to verify the validity of the underlying Fay-Herriot model. However, we are not aware of local diagnostics that could be used to determine, domain by domain, which estimator is preferable. The model-based mean square error can be viewed to some extent as a local diagnostic but it relies on the assumption that a model holds for all the domains. Yet, some users are concerned about a specific domain and are not really interested in a global criterion to assess the quality of the composite estimator for their single domain. Those users would be eager for inferences made conditionally on the domain parameters of interest, such as design-based inferences, as opposed to model-based inferences. We have adopted this conditional approach to produce local diagnostics, for each area, that give an indication of whether or not the composite estimator is likely to be more precise than the direct estimator. We have found that, depending on the domain sample size and the magnitude of the standardized model residual, it is possible to detect when the compos- ite estimator is expected to have a smaller design mean square error than the direct estimator. We will present the results of a simulation study, based on the Canadian Labor Force Survey data, that illustrate the effectiveness of our diagnostics.
Mots-clés : estimation petits domaines, inférence sous le plan, Fay-Herriot.

1 Le modèle Fay-Herriot

On considère une population U et un échantillon s tiré dans U selon un plan de sondage $p(s)$. U est partitionné en m domaines (domaines non planifiés) U_i ; $i = 1, \ldots, m$. On s'interesse aux paramètres d'intérêt θ_i associés aux domaines $i = 1, \ldots, m$. On dispose d'information auxiliaire sous la forme de vecteurs $\mathbf{z}_i$ de caractéristiques disponibles pour tous les domaines $i = 1, \ldots, m$.

On suppose qu'on dispose d'un **modèle de liaison** qui nous permet de décomposer les paramètres d'intérêt θ_i de la façon suivante :

$$\theta_i = \boldsymbol{\beta}^\top \mathbf{z}_i + v_i,$$

où :

(i) $\boldsymbol{\beta}^\top \mathbf{z}_i$ est l'effet connu ou expliqué par le modèle de θ_i,

(ii) v_i est l'effet inconnu ou l'effet local de θ_i (qui ne dépend pas de $\mathbf{z}_i$).

v_i est la réalisation d'une variable aléatoire qui suit une loi normale $\mathcal{N}(0, \sigma_v^2)$. Dans notre article, on fait l'hypothèse que $\boldsymbol{\beta}$ and σ_v^2 sont connus.

L'**estimateur direct** du domaine i est noté $\hat{\theta}_i$. Il peut s'agir d'un estimateur Horvitz-Thompson ou un estimateur par calage. On note e_i l'erreur d'échantillonnage :

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

On note ψ_i la variance de $\hat{\theta}_i$ sous le plan de sondage.

Le nombre d'unités sélectionnées dans le domaine i, n_i, peut être très faible (voir nul), ce qui peut conduire à une très faible précision de l'estimateur direct $\hat{\theta}_i$. Ce problème est à l'origine du domaine de recherche sur les petits domaines.

Fay et Herriot ont proposé un **estimateur composite** dans le but d'obtenir un estimateur plus précis que l'estimateur direct :

$$\hat{\theta}_i^* \ = \ \gamma_i \, \hat{\theta}_i \ + \ (1 - \gamma_i) \, \boldsymbol{\beta}^\top \mathbf{z}_i,$$

où $\gamma_i = \dfrac{\sigma_v^2}{\sigma_v^2 + \tilde{\psi}_i}$ et $\tilde{\psi}_i \ = \ \mathbb{E}(\psi_i \mid \mathbf{z}_i)$.

γ_i peut être vu comme une mesure de la force relative de l'échantillonnage par rapport au modèle de liaison.

(i) $\gamma_i \to 1$, lorsque **la taille de l'échantillon est importante** où lorsque le modèle est très peu puissant.

(ii) $\gamma_i \to 0$, lorsque la taille de l'échantillon est faible où lorsque **le modèle est très puissant** (à la limite on pourrait se dispenser d'échantillonner le domaine)

(iii) $\gamma_i \approx 0,5$, lorsque l'erreur d'échantillonnage est du même ordre de grandeur que l'erreur du modèle de liaison.

L'inférence sous le plan de sondage et le modèle de liaison est appelée inférence sous le **modèle combiné**. On peut écrire ce modèle sous la forme :

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

2 Inférence sous le modèle combiné ou sous le plan

Afin d'apprécier la précision de l'estimateur composite, on a deux possibilités : soit prendre l'erreur quadratique moyenne sous le plan de sondage, soit prendre l'erreur quadratique

moyenne sous le modèle combiné.
L'approche sous le modèle combiné, fournit les EQM suivantes pour les estimateurs direct et indirect du domaine i : (i) $mse(\hat{\theta}_i \mid \mathbf{z}_i) = \tilde{\psi}_i$ et (ii) $mse(\hat{\theta}_i^* \mid \mathbf{z}_i) = \gamma_i\ \tilde{\psi}_i$.

L'estimateur composite est donc toujours plus précis que l'estimateur direct avec une inférence sous le modèle combiné. Ce résultat résulte de la construction même de l'estimateur composite.
Par contre, et cela est une question légitime, on peut se demander si l'estimateur composite est toujours plus précis que l'estimateur direct avec une inférence sous le plan de sondage ? On montre que ça dépend de l'amplitude de l'effet local v_i et éventuellement de la valeur du coefficient γ_i. On donnera deux diagnostics locaux conçus pour évaluer si l'estimateur composite est susceptible d'être plus précis que l'estimateur direct. On entend par diagnostic local un diagnostic qui porte sur un domaine spécifique.

Les EQM sous le plan (on utilse la lettre D en indice) des estimateurs direct et indirect pour le domaine i valent : (i) $mse_D\left(\hat{\theta}_i\right) = \psi_i$ et (ii) $mse_D\left(\hat{\theta}_i^*\right) = \gamma_i^2\ \psi_i + (1-\gamma_i)^2\ v_i^2$.

Pour un domaine i donné et une valeur v_i fixée, on analyse les valeurs de $mse_D\left(\hat{\theta}_i\right)$ et $mse_D\left(\hat{\theta}_i^*\right)$ en fonction des valeurs de γ_i. Sous l'hypothèse simplificatrice $\psi_i \approx \tilde{\psi}_i$, on obtient : (i) $mse_D\left(\hat{\theta}_i\right) \approx \dfrac{1-\gamma_i}{\gamma_i}\ \sigma_v^2$ et (ii) $mse_D\left(\hat{\theta}_i^*\right) \approx (1-\gamma_i)\ \sigma_v^2 + (1-\gamma_i)^2\ (v_i^2 - \sigma_v^2)$ ou $mse_D\left(\hat{\theta}_i^*\right) \approx (1-\gamma_i)\gamma_i\ \sigma_v^2 + (1-\gamma_i)^2\ v_i^2$.

On en déduit que lorsque $|v_i| < \sqrt{2}\ \sigma_v$, $mse_D\left(\hat{\theta}_i^*\right)$ est plus petit que $mse_D\left(\hat{\theta}_i\right)\ \forall\ \gamma_i$. A l'inverse, lorsque $|v_i| > \sqrt{2}\ \sigma_v$, il existe une valeur de γ_i au delà de laquelle $mse_D\left(\hat{\theta}_i^*\right) > mse_D\left(\hat{\theta}_i\right)$.
L'effet local v_i doit donc être d'ampleur limitée si on espère obtenir que $mse_D\left(\hat{\theta}_i^*\right) \leq mse_D\left(\hat{\theta}_i\right)$.
On raisonne maintenant à γ_i fixé pour un domaine i donné.

Définition 1 (Limites de v_i)**.** *$v_i(-)$ et $v_i(+)$ sont les deux valeurs limites telles que :*

$$mse_D\left(\hat{\theta}_i^*\right) \leq mse_D\left(\hat{\theta}_i\right) \iff v_i \in [v_i(-),\ v_i(+)]$$

Proposition 1 (Limites de v_i)**.** *(i)* $v_i(+) = -\ v_i(-) \approx \sigma_v \sqrt{\dfrac{1+\gamma_i}{\gamma_i}}$ *, sous l'hypothèse* $\psi_i \approx \tilde{\psi}_i$.

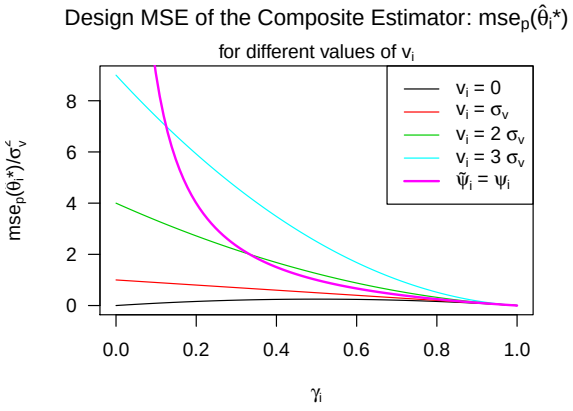

Figure 1: $mse_D\left(\hat{\theta}_i^*\right)$ peut être supérieure à $mse_D\left(\hat{\theta}_i\right)$

(ii) Pour des grandes valeurs de γ_i (i.e. γ_i proche de 1), le paramètre $|v_i|$ doit être inférieur à $\sqrt{2}\ \sigma_v$ pour avoir $mse_D\left(\hat{\theta}_i^*\right) \leq mse_D\left(\hat{\theta}_i\right)$.

(iii) Pour de petites valeurs de γ_i, le paramètre $v_i(+)$ peut prendre de très grandes valeurs ! (voir Figure (2))

En conséquence, l'estimateur direct peut devenir plus intéressant que l'estimateur indirect pour les domaines où l'effet local est important. Mais comment savoir si l'effet local est important où non pour un domaine i donné ? C'est l'objet de la section suivante où nous présentons deux diagnostics.

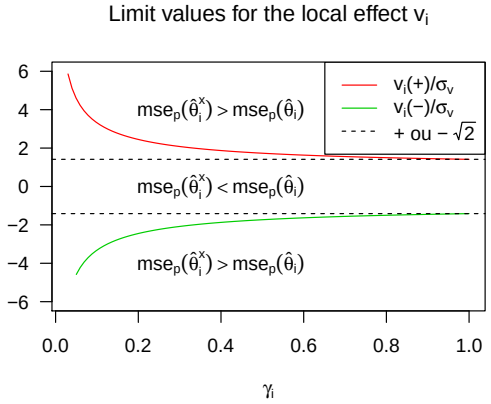

Figure 2: Valeurs limites de l'effet local en fonction de γ_i

3 Deux diagnostics pour choisir entre l'estimateur direct et l'estimateur indirect

Pour un domaine i donné, le choix entre l'estimateur direct et l'estimateur indirect dépend de :

(i) La valeur du coefficient γ_i qui est connue (ou estimée) ;

(ii) La valeur de v_i, qui est certes **inconnue** mais qui est captée indirectement à travers la quantité observée : $\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i \ = \ e_i \ + \ v_i$.

On utilise deux approches différentes pour dériver les deux diagnostics destinés à décider quel estimateur est le plus précis sous le plan de sondage :

(i) La première approche s'appuie sur des tests statistiques d'hypothèses qui concernent le paramètre v_i dans le cadre d'une inférence sous le plan.

(ii) La seconde approche dite " Empirical Bayes " (EB) s'appuie sur une distribution conditionnelle de v_i.

3.1 Première approche : test d'hypothèses sous le plan

On se place dans un domaine i donné. v_i est considéré comme un **paramètre** inconnu. Sous le plan de sondage, la variable aléatoire $\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i$ suit la loi :

$$\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i \mid v_i,\ \mathbf{z}_i \ \sim \ \mathcal{N}(v_i,\ \psi_i),$$

et nous avons une observation de cette variable aléatoire.

On utilise deux tests pour évaluer si $|v_i|$ est plus petit que $v_i(+)$

(i) **Deux jeux d'hypothèses** :
$\mathbf{H_0}(+) : v_i \ = \ v_i(+)$ $\qquad$ $\mathbf{H_1}(+) : v_i \ > v_i(+)$
$\mathbf{H_0}(-) : v_i \ = \ v_i(-)$ $\qquad$ $\mathbf{H_1}(-) : v_i \ < v_i(-)$

(ii) **Statistiques de test**:

$$S_i(+) = \frac{\varepsilon_i \ - \ \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \times \sqrt{\frac{\tilde{\psi}_i}{\psi_i}} \ \sim \ \mathcal{N}(0,\ 1)$$

$$S_i(-) = \frac{\varepsilon_i \ + \ \sqrt{1+\gamma_i}}{\sqrt{1-\gamma_i}} \times \sqrt{\frac{\tilde{\psi}_i}{\psi_i}} \ \sim \ \mathcal{N}(0,\ 1)$$

où $\varepsilon_i = \dfrac{\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i}{\sqrt{\sigma_v^2 + \tilde{\psi}_i}}$ est l'erreur standardisée du domaine i.

(iii) **P-values**:
$Pvalue(+) = \Phi\left(-S_i(+)\right)$
$Pvalue(-) = \Phi\left(S_i(-)\right)$
où $\Phi(\cdot)$ est la fonction de répartition de la loi normale.

Définition 2 (Premier diagnostic)**.** *Le premier diagnostic est défini comme le minimum des deux P-values* $(Pvalue(-),\ Pvalue(+))$ *:*

$$Diag(D)_i = \Phi\left(\frac{\sqrt{1+\gamma_i}\ -\ |\varepsilon_i|}{\sqrt{1-\gamma_i}}\right)$$

où $\varepsilon_i = \frac{\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i}{\sqrt{\sigma_v^2 + \tilde{\psi}_i}}$ *est l'erreur standardisée du domaine* i*.*

Le premier diagnostic s'interprète de la façon suivante. Quand $Diag(D)_i$ est faible, on rejète H_0, ce qui signifie qu'on "garde" l'estimateur direct $\hat{\theta}_i$. A l'inverse, lorsque $Diag(D)_i$ est grand, on accepte H_0, ce qui signifie qu'on "garde" l'estimateur composite $\hat{\theta}_i^*$.
Pour obtenir une règle de décision il est nécessaire de choisir un seuil. On peut par exemple s'inspirer des valeurs utilisées habituellement comme niveaux pour les tests ($\alpha = 5\%$ ou 10%). Avec ces valeurs, compte tenu du sens du test, on favorise l'estimateur composite. Une autre idée est d'adopter une approche empirique et de regarder la distribution des valeurs du diagnostic pour les m domaines $\{Diag(D)_i,\ i \in (1,\ \ldots,\ m)\}$ et de repérer une rupture.

3.2 Seconde approche : approche Empirical Bayes

Dans cette approche, v_i est vue comme la réalisation d'une variable aléatoire qui suit une loi normale $\mathcal{N}(0,\ \sigma_v^2)$. Selon l'approche "Empirical Bayes" (Rao et Molina, 2015, chap 9, pages 271-272), nous avons la loi conditionnelle de v_i :

$$v_i \mid \hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i\ \sim\ \mathcal{N}\left(\gamma_i(\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i),\ (1-\gamma_i)\sigma_v^2\right).$$

Le second diagnostic est défini comme la probabilité conditionnelle d'avoir $mse_D\left(\hat{\theta}_i^*\right) \leq mse_D\left(\hat{\theta}_i\right)$, c'est-à-dire:

$$\begin{aligned}
Diag(EB)_i &= Prob\left(mse_D\left(\hat{\theta}_i^*\right) \leq mse_D\left(\hat{\theta}_i\right) \mid \hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i\right) \\
&= Prob\left(v_i(-) \leq v_i \leq v_i(+) \mid \hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i\right) \\
&= \Phi\left(\frac{v_i(+) - \gamma_i(\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i)}{\sqrt{1-\gamma_i}\,\sigma_v}\right) - \Phi\left(\frac{v_i(-) - \gamma_i(\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i)}{\sqrt{1-\gamma_i}\,\sigma_v}\right)
\end{aligned}$$

Définition 3 (Second diagnostic)**.**

$$Diag(EB)_i \;=\; \Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(|\varepsilon_i|+\frac{\sqrt{1+\gamma_i}}{\gamma_i}\right)\right\}-1+\Phi\left\{\sqrt{\frac{\gamma_i}{1-\gamma_i}}\left(\frac{\sqrt{1+\gamma_i}}{\gamma_i}-|\varepsilon_i|\right)\right\},$$

où $\varepsilon_i = \dfrac{\hat{\theta}_i - \boldsymbol{\beta}^\top \mathbf{z}_i}{\sqrt{\sigma_v^2+\tilde{\psi}_i}}$ *est l'erreur standardisée du domaine* i*.*

4 Etude par simulation

On réalise l'étude par simulation à partir d'un jeu de données réelles de *l'enquête emploi canadienne de mai 2011* :

(i) $m = 140$ aires métropolitaines de recensement et aires de recensement

(ii) $\mathbf{z}_i$ est le ratio du nombre d'allocataires de l'assurance chômage dans l'aire i sur le nombre de personnes de plus de 15 dans l'aire i.

(iii) n_i est la taille de l'échantillon dans l'aire i

Ensuite, à partir de ces données réelles, on simule les paramètres θ_i pour les m domaines. Les paramètres d'intérêt θ_i peuvent être interprétés comme des taux de chômage des m domaines. On réalise trois jeux de simulations pour les θ_i : un jeu qu'on appellera jeu initial, puis un second jeu appelé "premier jeu modifié" et enfin un troisième jeu qualifié de "second jeu modifié". Dans le **jeu initial**, θ_i suit une loi béta $Beta(\boldsymbol{\beta}^\top \mathbf{z}_i,\ \sigma_v^2)$, où $\sigma_v = 0.87\%$ et $\boldsymbol{\beta}^\top = (0.0484,\ 1.95)$. Dans le **premier jeu modifié** $(5\sigma_v)$: on a modifié manuellement les valeurs de θ_i pour 4 aires de manière à avoir $v_i = 5 \times \sigma_v$. Dans le **second jeu modifié** $(15\sigma_v)$: on a modifié manuellement les valeurs de θ_i des mêmes 4 aires de manière à avoir $v_i = 15 \times \sigma_v$.

Le plan de sondage est un tirage aléatoire simple stratifié par aire avec remise. ψ_i and $\tilde{\psi}_i$ sont calculés une fois unique pour chaque $i = 1,\ \ldots,\ m$.

On répète $K = 10\ 000$ fois l'échantillonnage. Pour chaque itération k, $k = 1,\ \ldots,\ K$, on sélectionne un échantillon et on calcule l'estimateur direct $\hat{\theta}_i(k)$ et l'estimateur composite $\hat{\theta}_i^*(k)$: $\hat{\theta}_i^*(k) \;=\; \hat{\gamma}_i(k)\ \hat{\theta}_i(k) \;+\; \{1-\hat{\gamma}_i(k)\}\ \widehat{\boldsymbol{\beta}}(k)^\top \mathbf{z}_i$, où $\hat{\gamma}_i(k) = \dfrac{\hat{\sigma}_v^2(k)}{\hat{\sigma}_v^2(k)+\tilde{\psi}_i}$ et , $\widehat{\boldsymbol{\beta}}(k)$ et $\hat{\sigma}_v(k)$ sont estimés (moindres carrées pondérés pour $\widehat{\boldsymbol{\beta}}$ et maximum de vraisemblance réduite (REML) pour $\hat{\sigma}_v$).

On calcule également les quantités Monte Carlo $\hat{v}_i(+)(k)$, $\hat{v}_i(-)(k)$, $\hat{\varepsilon}_i(k)$, $\widehat{Diag(EB)}_i(k)$ et $\widehat{Diag(D)}_i(k)$. $\overline{\widehat{Diag(EB)}}_i$ et $\overline{\widehat{Diag(D)}}_i$ sont les moyennes Monte Carlo des diagnostics calculés à l'occasion des $K = 10\ 000$ itérations.

La mesure Monte Carlo de l'erreur quadratique sous le plan de sondage de l'estimateur $\hat{\theta}_i^*$ vaut : $mse_{MC}(\hat{\theta}_i^*) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_i^*(k) - \theta_i\right)^2$.

On calcule la différence relative entre la mesure Monte Carlo de l'EQM sous le plan de l'estimateur composite et l'EQM sous le plan de l'estimateur direct $\hat{\theta}_i$:

$$\frac{mse_{MC}(\hat{\theta}_i^*) \; - \; \psi_i}{\psi_i}.$$

Quand ce ratio est positif, on a un estimateur indirect moins précis que l'estimateur direct. On va comparer la valeur moyenne des deux diagnostics à ce ratio. Le diagnostic sera donc performant si les valeurs qu'il prend pour chacun des domaines sont corrélées aux valeurs du ratio pour ces mêmes domaines.

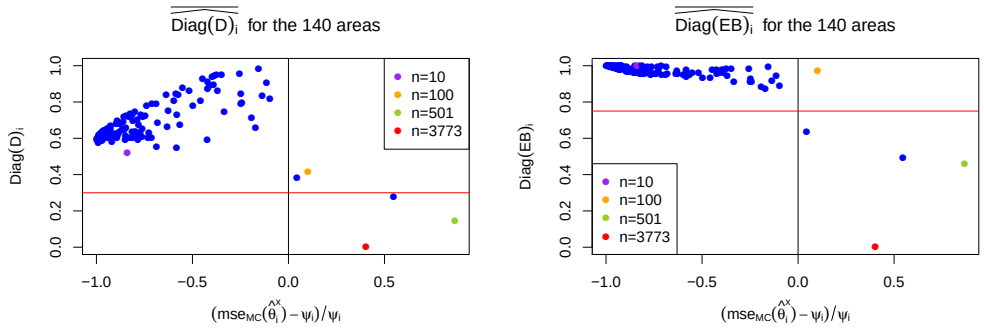

Figure 3: diagnostic D et EB Monte Carlo avec le premier jeu modifié (5 σ_v)

La Figure (3) présente les performances des diagnostics pour le premier jeu de données modifié. Les quatre aires dons les valeurs de v_i ont été modifiées sont colorés en *violet*, *jaune*, *rouge* et *vert*. Dans la légende, on a indiqué la taille de l'échantillon dans ces aires. L'aire de couleur *violette* est ainsi un "petit domaine" où la taille de l'échantillon est de 10 unités. À l'inverse l'aire en rouge est un "gros domaine".
On observe d'abord, sur la Figure (3), que l'estimateur composite est plus précis pour le petit domaine *violet* en dépit du fort effet local. Par contre, pour les trois autres aires à fort effet local, l'estimateur direct est plus efficace.
Le diagnostic D semble assez efficace en moyenne. Le diagnostic EB est un peu moins performant, puisqu'il n'arrive pas à repérer le domaine *jaune*.
La Figure (4) présente les performances des diagnostics pour le second jeu de données modifié. Cette fois, c'est le domaine *jaune* où l'écart relatif de performance entre l'estimateur direct et l'estimateur composite est le plus fort. Le diagnostic EB est plus performant car, à la différence du diagnostic D, il n'isole pas à tord le domaine *violet*.
La Figure (5) donne les diagnostics moyens avec le jeu de données initial. Ces résultats ont pour but de servir d'élément de comparaison pour les deux jeux de données modifiés<.

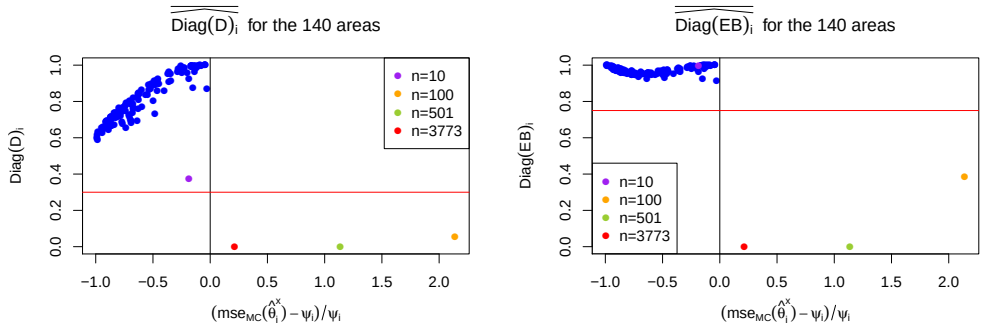

Figure 4: diagnostic D et EB Monte Carlo avec le premier jeu modifié ($15\ \sigma_v$)

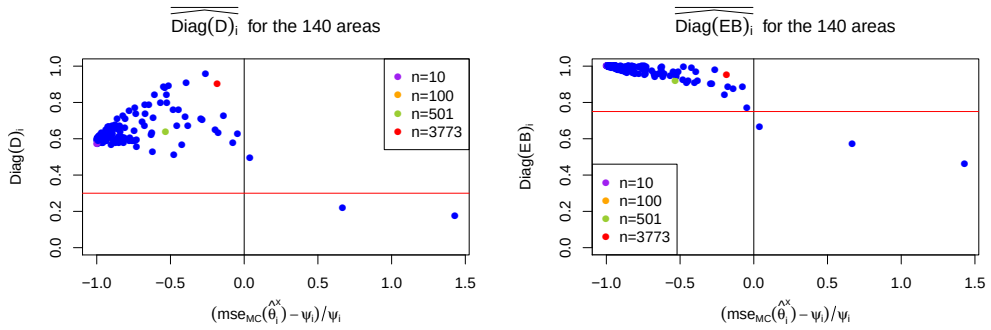

Figure 5: diagnostic D Monte Carlo avec le jeu initial

5 Conclusion

Les deux diagnostics proposés donnent des résultats prometteurs qui permettent d'envisager la mise au point d'un outil utile pour choisir entre l'estimateur direct et l'estimateur composite dans le cadre d'une approche sous le plan de sondage.

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.

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