Mesure de l'incertitude liée à l'estimation sur petits domaines basée sur un modèle : une évaluation

J. N. K. Rao

Université Carleton, Ottawa, Canada

Communication sollicitée présentée lors du 9e Colloque francophone sur les sondages, du 11 au 14 octobre 2016 à Gatineau, Québec, Canada

Mesures d'incertitude pour les estimateurs directs

- Échantillon s, poids de sondage $d_k, k \in s$
  Estimateur du total de la population Y basé sur le plan:

$$\hat{Y} = \sum_{k \in s} d_k y_k = \hat{Y}(y)$$

- Estimateur de variance : $v(\hat{Y}) = v(y) = s^2(\hat{Y})$

- $\mathrm{CV}(\hat{Y}) = s(\hat{Y}) / \hat{Y}$

- Intervalles de confiance s'appuyant sur la loi normale :

$$[\hat{Y} - z_{\alpha/2} s(\hat{Y}), \hat{Y} + z_{\alpha/2} s(\hat{Y})]$$

- Neyman (1934) : « Si les méthodes d'échantillonnage et d'estimation nous permettent d'attribuer à chacun des échantillons possibles s un intervalle de confiance tel que la fréquence des erreurs de couverture ne dépasse pas la valeur $1-\alpha$ prévue, *quelles que soient les propriétés inconnues de la population,* je qualifierais la méthode d'échantillonnage de *représentative* et la méthode d'estimation, de *convergente*. »

L'estimateur de variance n'est pas unique :

Estimateur **GREG** : $\hat{Y}_G = \sum_{k \in s} (d_k g_k) y_k = \hat{Y} + (X - \hat{X})^T \hat{B}$

Estimateurs de variance : $v(\hat{Y}_G) = v(e)$ et $v_1(\hat{Y}_G) = v(ge)$

Résidus : $e_k = y_k - x_k' \hat{B} = y_k - \hat{y}_k$

Poids GREG : $w_k = g_k d_k \ , k \in s$

Exemple : Estimateur par le ratio $\hat{Y}_R = (\hat{Y} / \hat{X}) X = \hat{R} X$

$g_k = X / \hat{X}$ and $e_k = y_k - \hat{R} x_k$

Quel estimateur de variance choisir ?

L'estimateur par le ratio converge sous le plan et est sans biais par rapport au modèle en vertu du modèle du ratio

$$E_m(y_k) = \beta x_k,\ V_m(y_k) = \sigma^2 x_k$$

- Contrairement à $v(\hat{Y}_R)$, $v_1(\hat{Y}_R)$ est convergent sous le plan et également asymptotiquement sans biais pour la variance sous le modèle de $\hat{Y}_R$.

Qu'en est-il de la couverture de l'intervalle de confiance ? Neyman a-t-il raison ?

- Pivot $(\hat{Y}_R - Y) / s(\hat{Y}_R) \approx \hat{Y}(e) / s(e)$

- Si les écarts par rapport au modèle du ratio ne sont pas importants, alors l'asymétrie des résidus e_k sera faible même si y et x présentent une forte asymétrie et les IC normaux fonctionnent bien. Par conséquent, la structure de la population revêt une importance dans les inférences fondées sur le plan, contrairement à la croyance populaire.

Exemple (Dorfman, 1994; Rao et coll., 2003)

- EAS en deux phases, estimateur par la régression linéaire $\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x})$

- Vrai modèle : $y_i = 8x_i^2 + \varepsilon_i$ et asymétrie des résidus de régression linéaire 6,40. Pour la taille de l'échantillon de première phase $n' = 80$ et la taille de l'échantillon de deuxième phase $n = 40$, la couverture de l'IC est de 61% .

- Estimateur assisté par un modèle de régression quadratique : la couverture de l'IC est de 91%.

Estimation directe par domaine

$s(i)$ = échantillon appartenant au domaine i, a_{ik} étant l'indicatrice d'appartenance au domaine

Estimateur sans biais : $\hat{Y}_i = \sum_{k \in s(i)} d_k y_k$

Estimateur GREG :

$$\hat{Y}_{Ri} = \sum_{k \in s(i)} w_k y_k = (X / \hat{X})\hat{Y}_i$$

$\hat{Y}_{Ri}$ n'apporte aucun gain d'efficacité par rapport à $\hat{Y}_i$.

Approche modèle pour l'EPD : modèle de Fay-Herriot au niveau des domaines

- Les paramètres d'intérêt sont les moyennes des domaines $\bar{Y}_i$ et les covariables associées au niveau du domaine z_i sont disponibles.
- Les estimateurs directs $\hat{\bar{Y}}_i$ ne sont pas fiables si les tailles d'échantillon des domaines sont petites. Il faut aller chercher de la stabilité en mobilisant tous les domaines par le biais de modèles de liaison qui utilisent les z_i.
- Modèle de liaison :

$$\theta_i = g(\bar{Y}_i) = z_i'\beta + v_i,\ v_i \sim_{iid} N(0, \sigma_v^2)$$

- Exemples de modèles de liaison : $g(P_i) = \sin^{-1}(\sqrt{P_i})$ pour les proportions $\bar{Y}_i = P_i$ (Casas-Cordero et coll., 2016) et $g(\bar{Y}_i) = \log(\bar{Y}_i)$ pour le revenu dans des petites localités (Fay et Herriot, 1987).

- Modèle d'échantillonnage « correspondant » au modèle de liaison :
  $\hat{\theta}_i = \theta_i + e_i,\ e_i \mid \theta_i \sim_{ind} N(0, \psi_i)$, où ψ_i est connu
- Modèle combiné : $\hat{\theta}_i = z_i'\beta + v_i + e_i$ (modèle linéaire mixte)
- Le modèle « sans correspondance » utilise $\hat{\bar{Y}}_i = \bar{Y}_i + f_i$ avec $E(f_i) = 0$.

- Données : $\{(\hat{\theta}_i, z_i), i = 1, \ldots, m\}$. Estimateurs des paramètres du modèle (β, σ_v^2) : Méthode de moments de FH, MV, MV réduit.

- Meilleur estimateur empirique (EB) de $\theta_i = \bar{Y}_i$:

$$\hat{\theta}_i^{EB} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) z_i' \hat{\beta} \ , \ \hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \psi_i)$$

- Modèles sans correspondance : Estimateur hiérarchique bayésien (HB) de la moyenne du domaine $\bar{Y}_i$ (You et Rao, 2002).

Estimation de l'erreur quadratique moyenne (EQM) inconditionnelle :

- Estimateur de l'EQM sans biais d'ordre deux pour MLR :

$$\text{mse}(\hat{\theta}_i^{EB}) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2)$$

- Le premier terme $g_{1i}(\hat{\sigma}_v^2) = \hat{\gamma}_i \psi_i$. Le second terme dû à l'estimation de β et le dernier terme dû à l'estimation de σ_v^2 sont d'ordre inférieur : $O(m^{-1})$.

Bootstrap paramétrique

- Étape 1 : Générez θ_i^* à partir de $N(z_i'\hat{\beta}, \hat{\sigma}_v^2)$.

  Étape 2 : Générez $\hat{\theta}_i^*$ à partir de $N(\theta_i^*, \psi_i)$.

  Étape 3 : Appliquez EB aux données de bootstrap $(\hat{\theta}_i^*, z_i), i = 1, \ldots, m$ pour obtenir les estimations d'EB $\hat{\theta}_i^{EB*}$.

  Étape 4 : Répétez les étapes 1 à 3 B fois.

- Estimateur de l'EQM bootstrap :

$$mse_B(\hat{\theta}_i^{EB}) = B^{-1}\sum\nolimits_b [\hat{\theta}_i^{EB*}(b) - \theta_i^*(b)]^2$$

- $\mathrm{mse}_B(\hat{\theta}_i^{EB}) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2)$

- Il est possible d'obtenir un estimateur ajusté pour le biais de $mse(\hat{\theta}_i^{EB})$, mais cela exige de mettre en œuvre un double bootstrap.

Propriétés sous le plan de $\text{mse}(\hat{\theta}_i^{EB})$

- On suppose que tous les paramètres sont connus :

$$\text{mse}(\hat{\theta}_i^{EB}) = g_{1i}(\sigma_v^2)$$

- L'espérance sous le plan de la moyenne des estimateurs de l'EQM convergent en probabilité sous le modèle vers la moyenne des EQM sous le plan des estimateurs EB quand $m \to \infty$.

Estimateur sans biais sous le plan de l'EQM

- Ecrivons $\hat{\theta}_i^{EB}$ sous la forme $\hat{\theta}_i + h_i(\theta)$ où $\theta = (\theta_1, \ldots \theta_m)'$.

- $\text{mse}_d(\hat{\theta}_i^{EB}) = \psi_i + 2\partial h_i(\hat{\theta}) / \partial \hat{\theta}_i + h_i^2(\theta)$

  est sans biais par rapport au modèle d'échantillonnage. Il peut prendre des valeurs **négatives**.

- Dans le cas particulier où les paramètres du modèle sont connus, nous avons

$$mse_d(\hat{\theta}_i^B) = \gamma_i \psi_i + (1-\gamma_i)[(\hat{\theta}_i - z_i^T \beta)^2 - (\sigma_v^2 + \psi_i)]$$

- La CV de l'estimateur de l'EQM sous le plan peut être très grand, particulièrement pour les domaines qui comportent de grandes variances d'échantillonnage (Datta et coll., 2011). Notons que pour de grandes valeurs de ψ_i nous avons $\gamma_i \approx 0$ et

$$mse_d(\hat{\theta}_i^B) \approx (\hat{\theta}_i - z_i^T \beta)^2 - (\psi_i + \sigma_v^2)$$

ce qui est instable et peut prendre des valeurs négatives.

Estimateur composite de l'EQM

- $mse_c(\hat{\theta}_i^{EB}) = \hat{\gamma}_i mse_d(\hat{\theta}_i^{EB}) + (1 - \hat{\gamma}_i) mse(\hat{\theta}_i^{EB})$

- On accorde moins de poids à l'estimateur de l'EQM sous le plan lorsque la variance de l'échantillonnage est grande et cela contrôle son CV.
- L'estimateur composite de l'EQM a un biais sous le plan plus petit que l'estimateur de l'EQM basé sur le modèle.

Résultats de la simulation

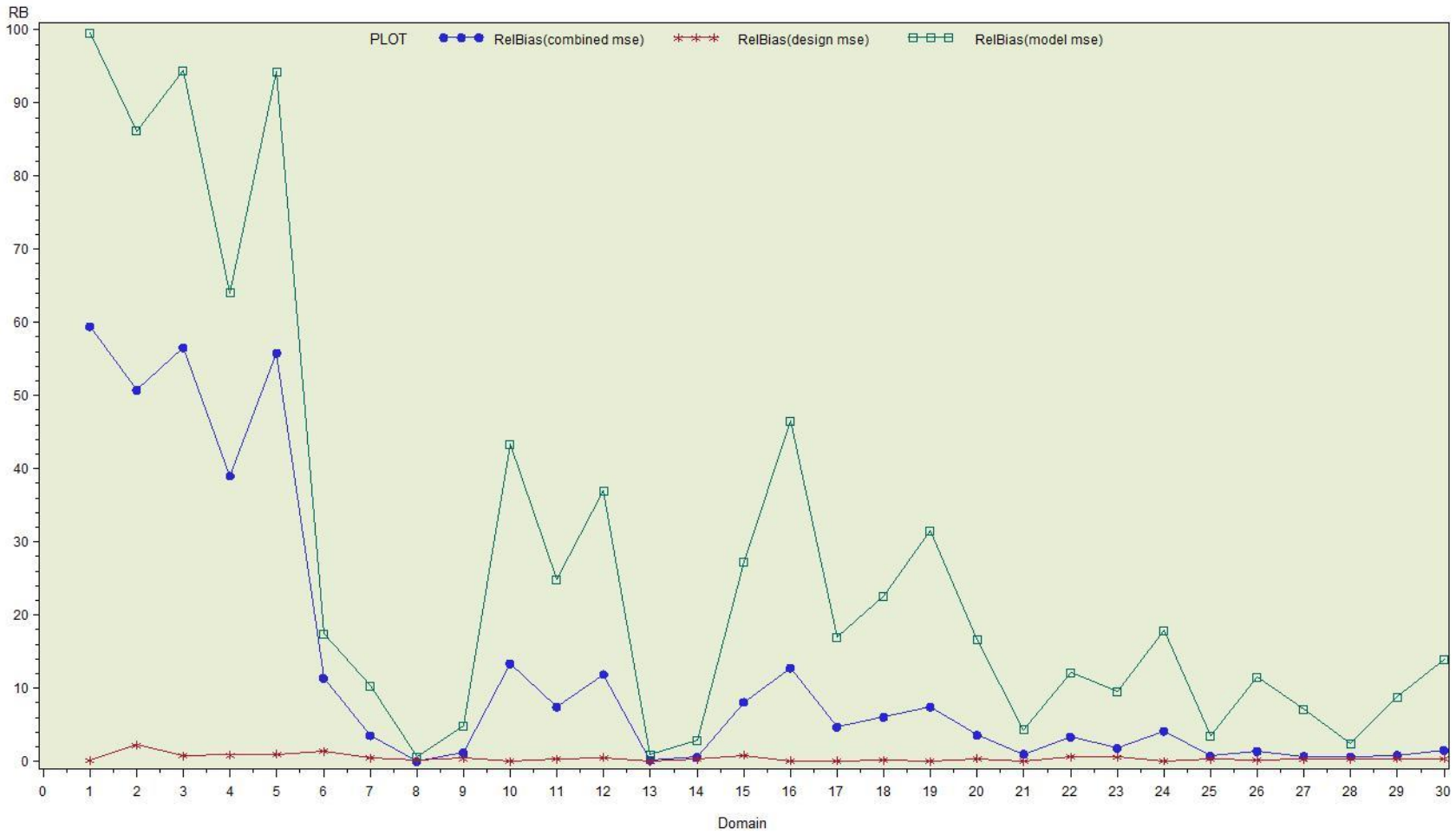
- $m = 30$ domaines avec des valeurs de ψ_i (2; 0,6; 0,5; 0,4; 0,2) données à six domaines chacune.
  On génère les $v_i \sim N(0,1)$, qui sont gardées fixes ;
  On calcule ensuite $\theta_i = z_i^T \beta + v_i$.

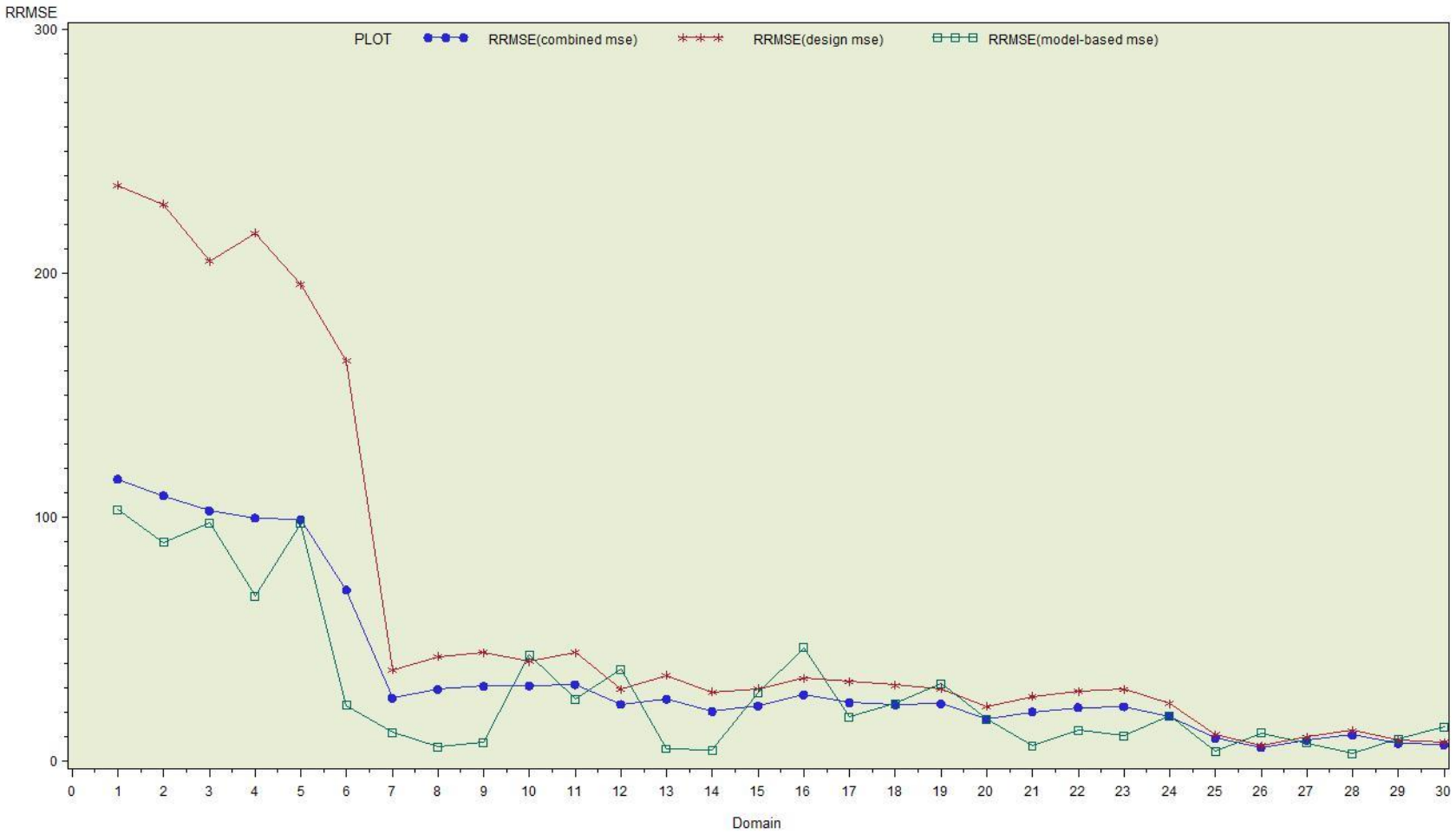
- On génère $\{\hat{\theta}_i^{(r)}, i = 1,...,30; \mathrm{r} = 1,...,\mathrm{R}\}$ à partir du modèle d'échantillonnage $\hat{\theta}_i = \theta_i + e_i$ avec $e_i \sim N(0, \psi_i)$.

- Pour chaque simulation, on estime les paramètres du modèle et on calcule les estimateurs EB et les estimateurs de leur EQM ($R = 30,000$).

  L'EQM réelle est calculée à l'aide de $R = 100,000.$

- Le biais relatif (BR) des estimateurs de l'EQM et la racine carrée de l'erreur quadratique moyenne relative (REQMR) sont calculés pour chacun des domaines triés par ordre de taille d'échantillon croissant.


% Relative Bias of MSE estimators
Pattern b, based on 30K samples
RB
PLOT
RelBias(combined mse)
RelBias(design mse)
RelBias(model mse)
100
90
80
70
60
50
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Domain


%RRMSE of MSE estimators
Pattern b, based on 30K samples
RRMSE
PLOT
RRMSE(combined mse)
RRMSE(design mse)
RRMSE(model-based mse)
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Domain

EQM conditionnelle à $\hat{\theta}_i$

- $MSE_c(\hat{\theta}_i^{EB}) = E[(\hat{\theta}_i^{EB} - \theta_i)^2 \mid \hat{\theta}_i]$

- Datta et coll. (2011) ont obtenu un estimateur sans biais d'ordre deux de l'EQM conditionnelle. Le terme principal demeure toutefois inchangé : $g_{1i}(\hat{\sigma}_v^2)$. Seul le terme g_{3i} est touché; il dépend de $\hat{\theta}_i$.

- L'estimateur de l'EQM conditionnelle est similaire à l'estimateur de l'EQM inconditionnelle en termes de BR et de CV.

Intervalles de confiance

Pivot : $t_i = (\hat{\theta}_i^{EB} - \theta_i) / \{g_{1i}(\hat{\sigma}_v^2)\}^{1/2}$

Pivot bootstrap : $t_i^* = (\hat{\theta}_i^{*EB} - \hat{\theta}_i^{EB}) / \{g_{1i}(\hat{\sigma}_v^{*2})\}^{1/2}$

Procédure : Générer B pivots bootstrap $t_i^*(1),...,t_i^*(B)$ et déterminer les points inférieur et supérieur q_1 and q_2 de sorte que l'aire entre les points inférieur et supérieur de la distribution bootstrap empirique soit égale au niveau spécifique $1-\alpha$.

- Intervalle calibré bootstrap de θ_i obtenu à partir de $q_1 \leq t_i \leq q_2$ sous la forme $c_{1i} \leq \theta_i \leq c_{2i}$.

- Cet intervalle est bon à l'ordre deux, à la condition que l'hypothèse de normalité tienne (Chatterjee et coll., 2008). L'intervalle pour une relation bijective $h(\theta_i) = g^{-1}(\theta_i)$ est obtenu sous la forme $h(c_{1i}) \leq h(\theta_i) \leq h(c_{2i})$.

- Casas-Cordero et coll. (2014) ont utilisé les intervalles bootstrap pour obtenir les taux de pauvreté dans les communes du Chili. Dans leur cas

$$\bar{Y}_i = P_i = h(\theta_i) = \sin^2(\theta_i).$$

Modèle au niveau de l'unité statistique

- Les données au niveau de l'unité $\{(y_{ij}, x_{ij}), j = 1,\ldots,n_i; i = 1,\ldots,m\}$ et les moyennes de la région de la population $\bar{X}_i$ sont disponibles.
- Modèle de régression linéaire à erreurs emboitées de la population :

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij},\ j = 1,\ldots,N_i; i = 1,\ldots,m$$

- Le modèle de la population tient pour l'échantillon : aucun biais lié à la sélection de l'échantillon.

- On dispose du meilleur estimateur empirique (EB) $\hat{\bar{Y}}_i^{EB}$ et de l'estimateur de son EQM basé sur le modèle (Rao et Molina, 20015, chapitre 7).

- L'estimateur sans biais sous le plan de l'EQM ($MSE_d(\hat{\bar{Y}}_i^{EB})$) a été calculé dans le cas d'un échantillonnage aléatoire simple à l'intérieur des domaines (en présumant que les paramètres du modèle sont connus).

- L'estimateur composite connexe de l'EQM est obtenu sous forme de moyenne pondérée des deux estimateurs de l'EQM.