Statistics Canada
www.statcan.gc.ca

Un survol de quelques méthodes d'estimation pour les petites régions 81e Congrès de l'Acfas

M.A. Hidiroglou
Statistique Canada
6 mai 2013

Plan de la présentation

1. Introduction
2. Quelques méthodes de base
3. Bâtir un système
4. Exemples
5. Conclusions

1. Introduction

Définition

- Une sous-population (ou domaine) : petit nombre ou aucune unité échantillonnée
  - estimations directes ne peuvent pas être produites avec une bonne précision
  - emprunt de données à travers les petites régions par l'intermédiaire de la modélisation

2. Quelques méthodes de base

- **Échantillon** s
  - sélectionné à partir d'un mécanisme probabiliste
  - Petits domaines : U_i, $i=1,\ldots, M$
  - se divise en M petits domaines s_i $(i=1,\ldots, M)$
  - Certains sont vides $i=1,\ldots m$
  - Les autres ont des unités: $n_i > 0$
  - Union peut être tout l'échantillon
- **Poids**
  - Sondage : w_j pour $j \in s$
  - Final : $\tilde{w}_j$ pour $j \in s$ tient compte de la non-réponse, ainsi que des données auxiliaires ($\boldsymbol{x}$)

2. Quelques méthodes de base (suite)

- **Paramètres d'intérêt**

  Totaux Y_i

  Moyennes $\overline{Y}_i$

  Proportions P_i

- **Plus général (Rao, 2003)**

  $\theta_i =$ une fonction des paramètres

  Moyenne: $\theta_i = \overline{Y}_{1i}$; Total : $\theta_i = Y_{1i}$

  Rapport: $\theta_i = \overline{Y}_{1i} / \overline{Y}_{2i}$; Logarithme : $\theta_i = \log\left(\overline{Y}_{1i}\right)$

2. Quelques méthodes de base (suite)

Estimation des paramètres

Directe: seulement les poids finaux et les données auxiliaires	$\hat{\theta}_i^{DIR}$
Indirecte: emprunte « de la Force » à l'ensemble des petits domaines: les données auxiliaires $\boldsymbol{z}$ ne correspondent pas nécessairement à $\boldsymbol{x}$	$\hat{\theta}_i^{SYN} = \sum_{U_i} \boldsymbol{z}_j^T \hat{\boldsymbol{\beta}} = \boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}$
Combiner directe et indirecte	$\hat{\theta}_i^{EPD} = \gamma_i \ \hat{\theta}_i^{DIR} + (1-\gamma_i) \ \hat{\theta}_i^{SYN}$ avec $0 \le \gamma_i \le 1$

2. Quelques méthodes de base (suite)

Questions

1. Où sont les données auxiliaires?
2. À quel niveau modéliser? Domaines, données individuelles.
3. Comment choisir les variables auxiliaires les plus fortement liées à l'estimation que l'on veut produire?
4. Devons-nous modéliser chaque variable d'intérêt? Implique beaucoup de travail.
5. Devons-nous tenir compte du plan de sondage (poids / variance)?

2. Quelques méthodes de base (suite)

Modèle au niveau des domaines (avec des poids finaux)

Modèle de l'échantillon: $\hat{\theta}_i^{DIR} = \theta_i + e_i;\ E_p(e_i) = 0,\ Var_p(e_i) = \psi_i$ (connu)

e_i - erreurs d'échantillonnage

Modèle de liaison: $\theta_i = \underbrace{\boldsymbol{Z}_i^T \boldsymbol{\beta}}_{\substack{effets \\ fixes}} + \underbrace{b_i v_i}_{\substack{effets \\ aléatoire}},\ v_i \overset{iid}{\sim} (0, \sigma_v^2)$

b_i - constantes - Erreur structurel plus souple

Modèle combiné: $\hat{\theta}_i^{DIR} = \boldsymbol{Z}_i^T \boldsymbol{\beta} + b_i v_i + e_i$

Modèle de sondage lié au modèle de sondage

Tient compte du plan de sondage

2. Quelques méthodes de base (suite)

Modèle au niveau des domaines (avec des poids finaux)

Meilleur prédicteur linéaire sans biais (BLUP) de θ_i :

$$\hat{\theta}_i^{FH} = \boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}^{FH} + \gamma_i \left(\hat{\theta}_i^{DIR} - \boldsymbol{Z}_i^T \hat{\boldsymbol{\beta}}^{FH} \right)$$

$$\hat{\boldsymbol{\beta}}^{FH} = \left(\sum_{i=1}^{m} \frac{\boldsymbol{Z}_i \boldsymbol{Z}_i^T}{\psi_i + \sigma_v^2} \right)^{-1} \sum_{i=1}^{m} \frac{\boldsymbol{Z}_i \hat{\theta}_i^{DIR}}{\sigma \psi_i + \sigma_v^2}$$

$$\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$$

Meilleur prédicteur empirique linéaire sans biais (EBLUP)

Remplacer σ_v^2 par $\hat{\sigma}_v^2 \rightarrow \tilde{\theta}_i^{FH}$

2. Quelques méthodes de base (suite)

$$\hat{\gamma}_i = b_i^2 \hat{\sigma}_v^2 / (\psi_i + b_i^2 \hat{\sigma}_v^2)$$

b_i : sert à atténuer l'impact des données aberrantes

: on le met plus grand que les autres

$\hat{\gamma}_i$ près de 0 : estimateur synthétique

$\hat{\gamma}_i$ près de 1 : estimateur directe

2. Quelques méthodes de base (suite)

Modèle au niveau des unités (sans poids finaux)

Battese, Harter & Fuller (1988)

$$y_{ij} = \boldsymbol{z}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \quad v_i \overset{iid}{\sim} (0, \sigma_v^2); \quad e_{ij} \overset{iid}{\sim} (0, \sigma_e^2)$$

Paramètre d'intérêt: $\bar{Y}_i \approx \bar{\boldsymbol{Z}}_i^T \boldsymbol{\beta} + v_i$

BLUP: $\hat{\theta}_i^{BHF} = \bar{\boldsymbol{Z}}_i^T \hat{\boldsymbol{\beta}}^{BHF} + \gamma_i \left(\hat{\bar{Y}}_{i,DIR} - \hat{\bar{\boldsymbol{Z}}}_i^T \hat{\boldsymbol{\beta}}^{BHF} \right)$

$$\hat{\boldsymbol{\beta}}^{BHF} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\boldsymbol{z}_{ij} \boldsymbol{z}_{ij}^T - \hat{\gamma}_i \, \hat{\bar{\boldsymbol{Z}}}_i \hat{\bar{\boldsymbol{Z}}}_i^T \right) \right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\boldsymbol{z}_{ij} y_{ij} - \hat{\gamma}_i \, \hat{\bar{\boldsymbol{Z}}}_i \hat{\bar{Y}}_i \right)$$

Ne tient pas compte du plan de sondage

2. Quelques méthodes de base (suite)

Modèle au niveau des unités (avec poids finaux)

- Prasad-Rao (1999): poids finaux $\tilde{w}_j$

  - Pseudo BLUP de $\bar{Y}_i$:

$$\hat{\bar{Y}}_i^{PR} = \bar{\boldsymbol{Z}}_i^T \hat{\boldsymbol{\beta}}^{PR} + \gamma_{iw}\left(\bar{y}_{iw} - \bar{\boldsymbol{z}}_{iw}^T \hat{\boldsymbol{\beta}}^{PR}\right)$$

$$\bar{y}_{iw} = \sum_{j\in s_i} \tilde{w}_j\, y_j \Big/ \sum_{j\in s_i} \tilde{w}_j$$

$$\gamma_{iw} = \sigma_v^2 \Big/ \left(\sigma_v^2 + \sigma_e^2 \left(\sum_{j\in s_i} \tilde{w}_j^2\right) \Big/ \left(\sum_{j\in s_i} \tilde{w}_j\right)^2\right)$$

$\hat{\boldsymbol{\beta}}^{PR}$ = reg. pondérée de $\bar{y}_{iw}$ sur $\bar{\boldsymbol{z}}_{iw}$

2. Quelques méthodes de base (suite)

Modèle au niveau des unités (avec poids finaux)

- You- Rao (2002): étalonnage automatique
- Pseudo-BLUP de $\bar{Y}_i$:

$$\hat{\bar{Y}}_i^{YR} = \bar{\boldsymbol{Z}}_i^T \hat{\boldsymbol{\beta}}^{YR} + \gamma_{iw}\left(\bar{y}_{iw} - \bar{\boldsymbol{z}}_{iw}^T \hat{\boldsymbol{\beta}}^{YR}\right)$$

$$\bar{y}_{iw} = \sum_{j\in s_i} \tilde{w}_j\, y_j \Big/ \sum_{j\in s_i} \tilde{w}_j$$

$$\hat{\boldsymbol{\beta}}^{PR} = \text{ reg. pondérée de } y_j \text{ sur } \boldsymbol{z}_j - \gamma_{iw}\bar{\boldsymbol{z}}_{iw}$$

Étalonnage automatique: $\sum_{i=1}^{m} N_i \hat{\bar{Y}}_i^{YR} = \hat{Y}_w$

3. Bâtir un système

Pourquoi?

- Méthodes d'estimation pour petits domaines (EPD)
  - Pas simples
  - Pas facile à traduire dans les logiciels
- Logiciels disponibles dans les bureaux nationaux de statistique (ex: projet EURAREA)
  - Pas faciles à transférer
  - Ne représentent pas les progrès récents en estimation
  - Souvent plus en mode de développement, non supportés

3. Bâtir un système (suite)

- Développer un nouveau système à STC bénéfique
  - Pour la production
  - Outil d'apprentissage pour les méthodologistes
  - Estevao, You, et Hidiroglou (2013)
- Ou en sommes nous?
  - Modélisation de domaines basée sur la méthode de Fay et Herriot (1979)
  - Modélisation individuelle avec pondération, au niveau des données et du sondage

3. Bâtir un système (suite)

- Méthodes hiérarchiques de Bayes

  Modèle de sondage: $\hat{\theta}_i^{DIR} = \theta_i + e_i$

  Modèle de liaison: $h\left(\theta_i\right) = \boldsymbol{Z}_i^T \boldsymbol{\beta} + v_i$

  Modèle de liaison non lié au modèle de sondage

  Exemple: $h\left(\theta_i\right) = p_i$

  Assure que $\hat{p}_i > 0$

- Diagnostiques associées à chaque méthode

3. Bâtir un système (suite)

Estimation au niveau du domaine

- Calcule les moyennes, les totaux ou d'autres statistiques pour lesquels ψ_i est "connu"
- Deux problèmes avec l'estimation des variances ψ_i et σ_v^2

a. Théorie suppose que σ_i^2 est connue: Estimé par $\hat{\psi}_i$

b. L'estimateur de σ_v^2 peut être négatif

3. Bâtir un système (suite)

Estimation au niveau du domaine

Estimation de ψ_i

- Se servir de données historiques afin d'estimer ψ_i
- Lisser $\hat{\psi}_i$ à l'aide de la fonction de variance généralisée
- Exemple: Régresser le log de $\hat{\psi}_i$ sur des données auxiliaires, et prévoir $\hat{\psi}_i$
- Utilisez les données brutes $\hat{\psi}_i$ tel que suggéré par Wang et Fuller (2003), Rivest et Vandal (2003)

3. Bâtir un système (suite)

- Algorithme récursif pour estimer σ_v^2
- Maximum de vraisemblance restreint (REML): Bartlett (1937)

  - Cependant $\hat{\sigma}_v^2$ peut converger vers une valeur négative
  - Evitez ce problème $\tilde{\sigma}_v^2 = \max\left(0, \hat{\sigma}_v^2\right)$
  - Ce qui veut dire que $\tilde{\gamma}_i = 0$
  - Et on obtient $\hat{\theta}_i^{FH} = \overline{\mathbf{Z}}_i^T \hat{\boldsymbol{\beta}}^{FH}$

3. Bâtir un système (suite)

- Autres méthodes pour contourner $\hat{\sigma}_v^2$ négatif
  - Méthode des moment de Fay-Herriot (1979) qui contraint $\sigma_v^2 \geq 0$
  - Maximisation de la densité ajustée: (MDA) Li et Lahiri (2010) qui contraint $\sigma_v^2 > 0$
  - Règle max de Wang and Fuller (2003)

$$\tilde{\sigma}_v^2 = \max\left(0.5\sqrt{\hat{V}\left(\hat{\sigma}_v^2\right)}, \hat{\sigma}_v^2\right)$$

  - Applicable à toutes les autres méthodes

3. Bâtir un système (suite)

- **Étalonnage des totaux ou moyennes**

  Désirons que $\hat{\theta}_i^{EPD}$ s'ajoutent directement

  $\hat{Y}^{DIR} = \sum_{i=1}^{M} \hat{\theta}_i^{EPD} = \sum_{i=1}^{M} \hat{Y}_i^{DIR}$

  On suppose que $m \leq M$ petits domaines

  ont des unités échantillonnées

- **Au niveau du domaine:** You et Rao (2003)
- **Au niveau de l'unité:** Wang, Fuller et Qu (2008) ,
  You, Rao et Hidiroglou (2013)

3. Bâtir un système (suite)

- Estimation de l'EQM

$$mse\left(\hat{\theta}_i^{EPD}\right) = g_{1,i} + g_{2,i} + 2g_{3,i}$$

$$g_{1,i} = \text{Terme le plus important}$$

$$g_{2,i} = \text{Tient compte de l'estimation de } \beta$$

$$2g_{3,i} = \text{Tient compte de l'estimation } \sigma_v^2$$

- Terme de correction du biais: Fay-Herriot et MDA
  - ajouter un autre terme g_{0i}
- Estimation de l'EQM de Wang and Fuller (2003)
  - ajouter le terme 4 g_{4i} reflétant que ψ_i est estimé

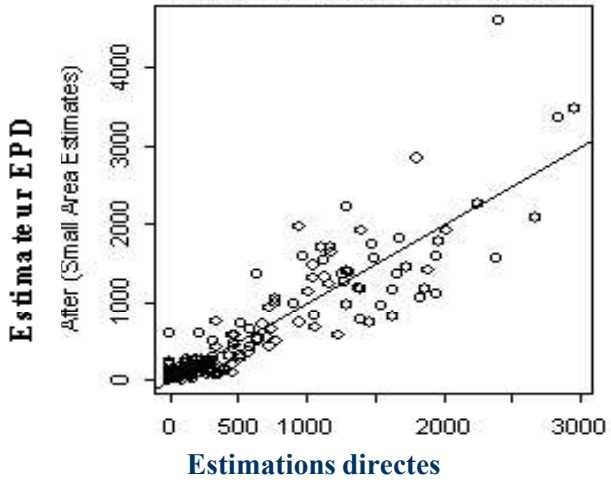
3. Bâtir un système (suite)

Diagnostics / Vérification

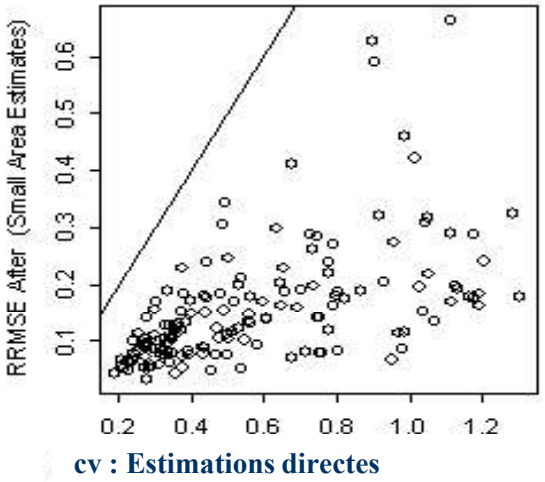
- Résumés de fiabilité des estimations (graphiques, tableaux)
- Identification de domaines aberrants par l'intermédiaire de graphiques de résidus standardisés
- Crédibilité des estimations vis-à-vis du bon sens
- Améliorations des estimations: meilleures données auxiliaires et modèles
- Validation externe

3. Bâtir un système (suite)

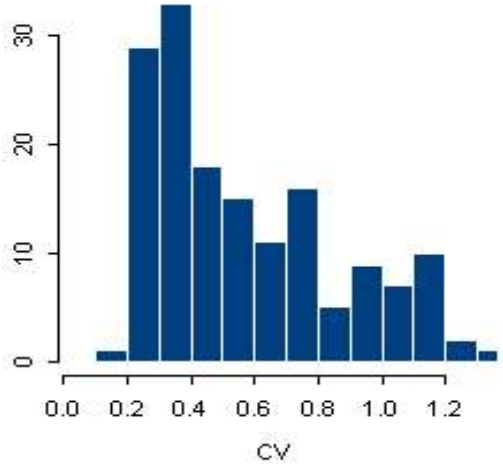
Graphique : Estimateur direct et EPD



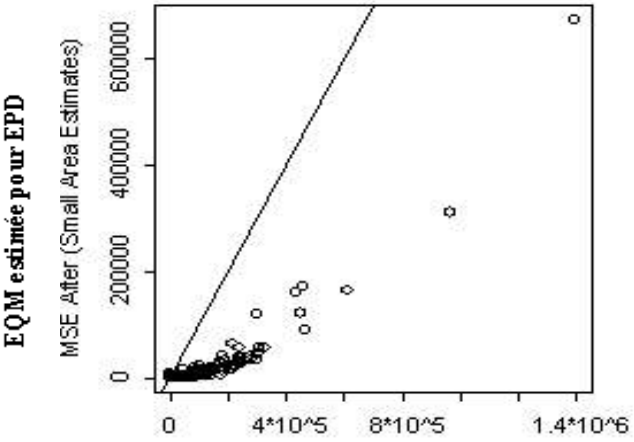
Réduction du « cv »



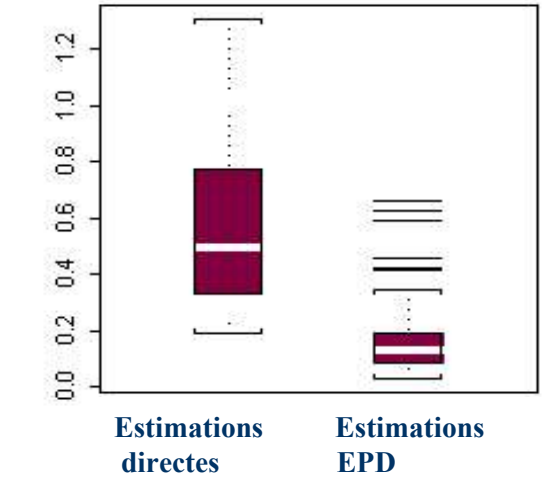
Histogramme: cv des estimations directes



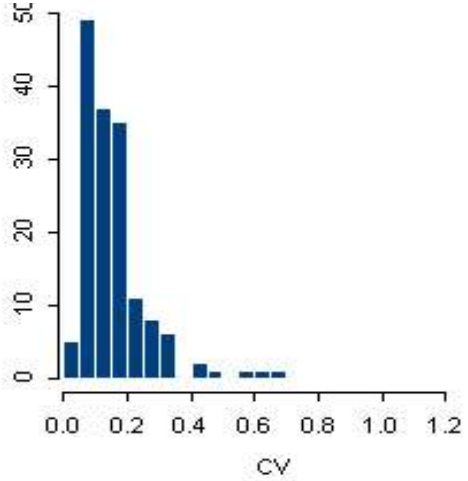
Graphique : var. des estimations directes EQM estimée des EPD



Box plot: cv des estimations directes et EPD

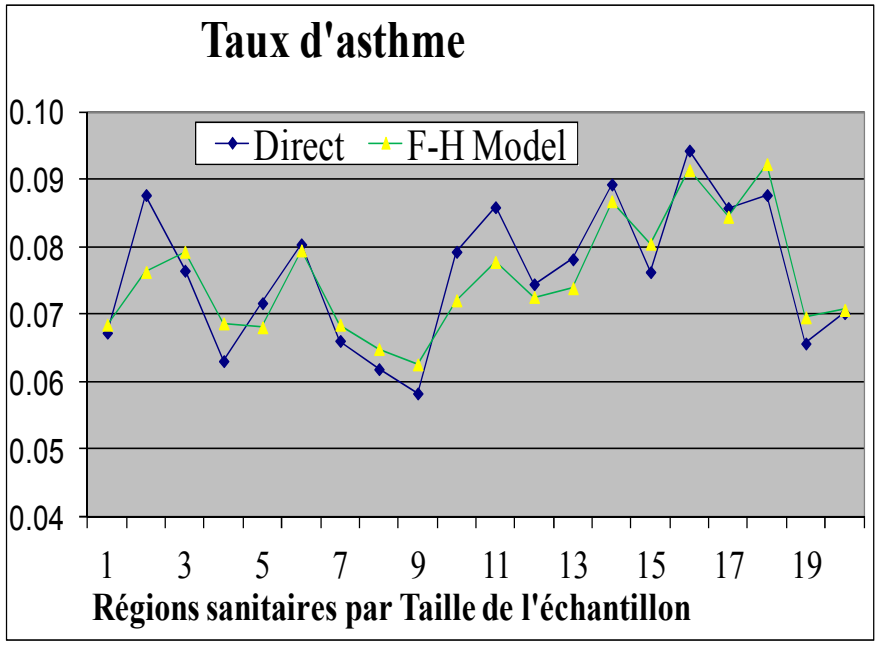


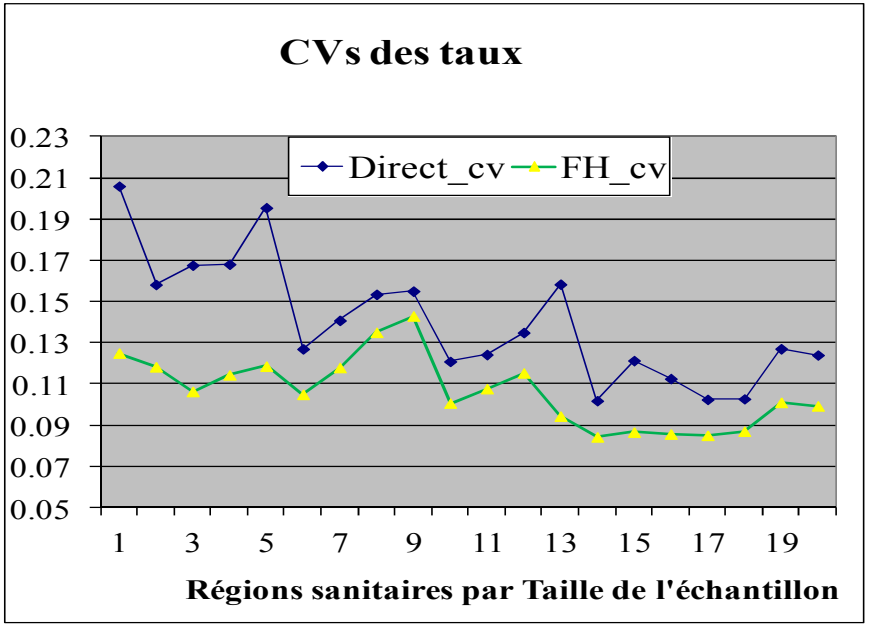
Histogramme: EQM estimée pour EPD



4. Exemple 1

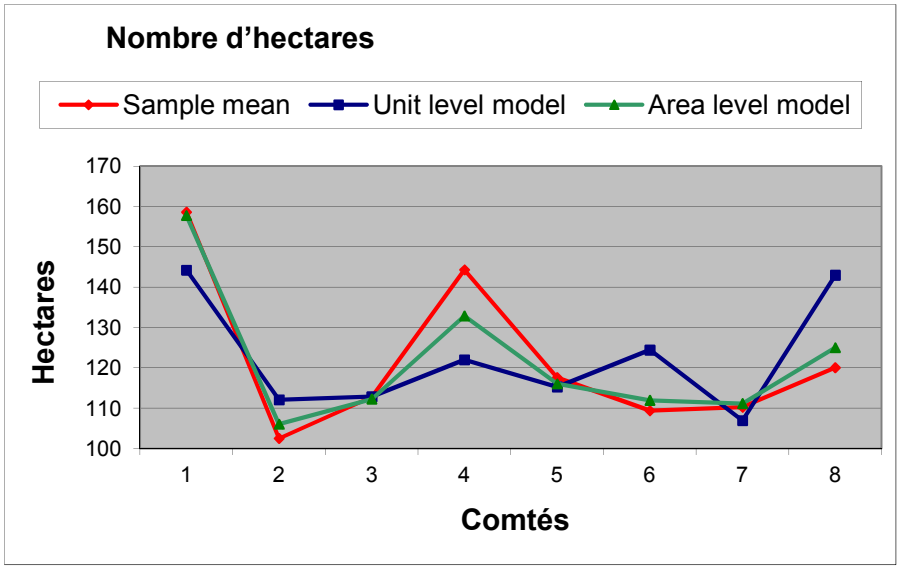
- **Enquête sur la santé dans les collectivités canadiennes: modèle au niveau région**
  - Estimations directes et basée sur un modèle: 20 régions de santé en Colombie-Britannique énumérés par la taille de la population (petites et grandes)
  - Estimation de taux d'asthme et de sa fiabilité

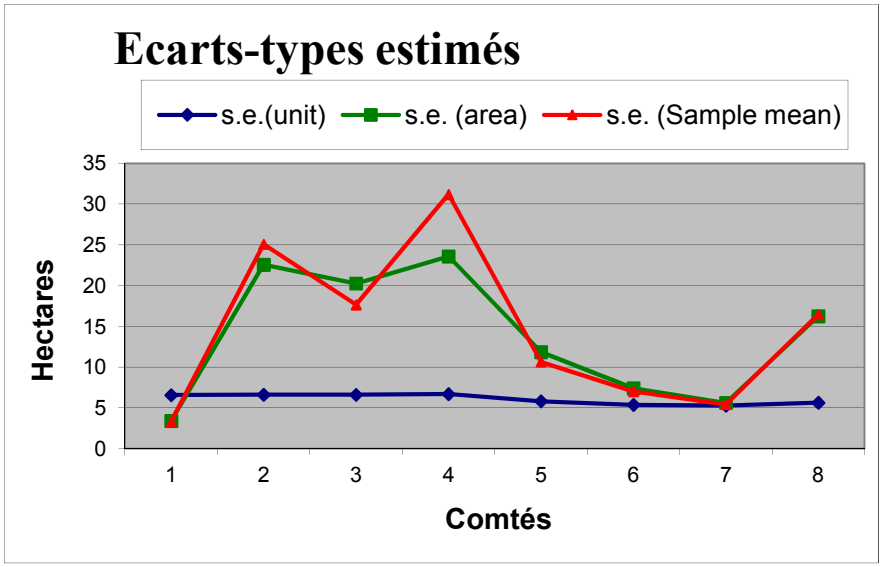




4. Exemple 2

- **Battese, Harter et Fuller (1988): données de maïs**
  - Superficie estimée sous maïs pour chacune des $m = 12$ comtés de l'Iowa Centre-Nord en utilisant des données d'entrevue agricoles et les données LANDSAT
  - Comparaison des estimations ((moyenne/segment) au niveau des comtés et des segments (unité)





5. Conclusions

- La demande de statistiques pour des petits domaines s'accroît
- Ces méthodes sont connues par un groupe relativement restreint de personnes
- Comment pouvons-nous combler cet écart?
  - Avec le développement de logiciels accessibles
  - En développant de la formation sur ces techniques
- Le mse mesure la fiabilité.
  - Si le modèle est bon, on peut s'en servir pour établir des intervalles de confiances raisonnable
  - Et ceci en se basant sur une approche basée uniquement sur les modèles

5. Conclusions

- A quel niveau modéliser?
  - Ma préférence: données individuelles pondérées
    - plus robuste à un mauvais modèle
    - protège contre un plan de sondage informatif
- Quelle est la méthode avec le moins de problèmes?
  - Méthode Hiérarchique de Bayes inclue au niveau du domaine
    - Evite le problème de la méthode de Fay–Herriot: $\hat{\sigma}_v^2 < 0$
    - Proportions estimées tout le temps > 0
    - Peut être appliqué à des modèles de liaison non lié au modèle de sondage
  - Nous devons ajouter cette méthode au niveau de l'unité