Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

Une application du modèle de Fay-Herriot pour l'estimation du taux de chômage dans les villes canadiennes

Octobre 2016
Colloque francophone sur les sondages,
Gatineau

Jean-François Beaumont et Cynthia Bocci
Statistique Canada

Sommaire

- Contexte et notation
- Aperçu de notre méthodologie d'Estimation pour petits domaines (EPD)
- Résultats d'une étude empirique

Historique de L'EPD à Statistique Canada

- Les travaux de recherche sur l'EPD à Statistique Canada sont peu nombreux avant 2000 mais remontent au moins aux années 70 (M.P. Singh et coll.)
- **Rarement implémentés:** Réticence à l'utilisation d'estimateurs dépendants d'un modèle
  - **Exception:** Dick (1995)
- Une théorie unifiée existe maintenant: Rao and Molina (2015)
- Plus de recherche à Statistique Canada dans les dernières années
  - **Prototype:** Estevao, Hidiroglou and You (2015)

Un nouvel intérêt!

- Rapport du printemps 2014 du Vérificateur général du Canada
  - *"Statistique Canada devrait évaluer s'il est possible de mieux répondre aux besoins des utilisateurs en matière de données sur les petites régions et sous-populations."*
- On a donc initié un projet d'EPD en avril 2015:
  - **Objectif principal:** Démontrer l'efficacité des méthodes d'EPD à combler les besoins d'estimations fiables pour des petits domaines en utilisant les données de quatre enquêtes
  - Aujourd'hui: Enquête sur la population active (EPA)

Pourquoi utiliser des méthodes d'EPD?

- Les **estimations directes** pour un domaine sont typiquement fiables si
  - La taille d'échantillon dans le domaine est "grande"
  - Les erreurs non dues à l'échantillonnage (telles que les erreurs de mesure) sont relativement petites
- **Estimation directe:** N'utilise que des données collectées qui proviennent du domaine d'intérêt
- Pourquoi les estimations pour petits domaines peuvent être fiables même si la taille d'échantillon dans le domaine est petite?
  - Comble la petite quantité de données par de l'information additionnelle qui prend la forme d'un modèle **(hypothèses)**

L'inconvénient principal de l'EPD

- On prend plus de risques mais ça peut en valoir la peine
- **Comment gérer les risques?**
  - Les risques peuvent être grandement réduits en vérifiant et validant soigneusement le modèle
  - **Exactement comme lorsqu'on vérifie et valide les données collectées pour s'assurer que les erreurs de mesure sont négligeables**
- Les estimations pour petits domaines reposent sur la validité du modèle
  - Particulièrement pour les plus petits domaines
  - Normalement pas très loin des estimations directes pour les grands domaines (qui ont une grande taille d'échantillon)

Contexte de l'Enquête sur la population active

- Enquête mensuelle avec un plan de sondage stratifié à deux degrés
  - **Estimateur direct:** Estimateur composite par la régression
  - **Estimateur direct de variance:** bootstrap de Rao-Wu
- Le problème d'EPD
  - Estimation des taux de chômage pour 146 domaines (villes) au Canada: 34 RMR (plus grandes villes en terme de taille de population) and 112 AR
  - L'enquête n'est pas conçue pour produire des estimations directes précises pour toutes les villes à tous les mois (particulièrement pour les AR)
  - **Les méthodes d'EPD peuvent-elles résoudre ce problème?**

Notation

- Taux de chômage dans le domaine i, $i = 1, \ldots, m$:

$$\theta_i = \frac{\#\text{ de personnes en chômage dans le domaine } i}{\#\text{ de personnes dans la population active dans le domaine } i}$$

- Variable auxiliaire: Taux de bénéficiaires calculé à partir de sources externes

$$z_{1i} = \frac{\#\text{ de bénéficiaires d'assurance emploi dans le domaine } i}{\text{taille de la population (15+) dans le domaine } i}$$

- Estimateur direct: $\hat{\theta}_i$
- Variance de l'estimateur direct: $\psi_i = \text{var}_p(\hat{\theta}_i)$

Modèle de Fay-Herriot (au niveau des domaines)

Modèle à deux composantes

- **Modèle d'échantillonnage**:

$$\hat{\theta}_i = \theta_i + e_i \ , \ e_i \rightarrow N(0, \psi_i)$$

  - Hypothèse clé: $E_p(e_i) = E_p(\hat{\theta}_i - \theta_i) = 0$

- **Modèle de liaison**:

$$\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i \ , \ v_i \rightarrow N(0, \sigma_v^2)$$

$$\mathbf{z}_i' = (1 \ , \ z_{1i})$$

Modèle de Fay-Herriot

- **Modèle combiné**:

$$\hat{\theta}_i = \mathbf{z}_i'\boldsymbol{\beta} + a_i \quad , \quad a_i = \left(v_i + e_i\right)$$

  - $E_m\left(a_i \middle| \mathbf{z}_i\right) = 0$
  - $\text{var}_m\left(a_i \middle| \mathbf{z}_i\right) = \sigma_v^2 + \tilde{\psi}_i$
  - $\tilde{\psi}_i = E_m\left(\psi_i \middle| \mathbf{z}_i\right)$ est une **variance lisse inconnue**
  - **Est-ce que les deux variances sont égales?** Seulement si la variance ψ_i peut être traitée comme étant fixe étant donné les variables auxiliaires

Modèle de Fay-Herriot

- Exemple:
  - Supposons que θ_i est une proportion dans le domaine *i*
  - Supposons que: strates = domaines et Éch. Strat. A. S. A. R.
  - Supposons que $\hat{\theta}_i$ est la proportion échantillonnale
  - La variance sous le plan est:

$$\psi_i = \theta_i(1-\theta_i)/n_i$$

  - Elle est aléatoire parce que θ_i est aléatoire

$$\psi_i \neq \tilde{\psi}_i = E_m\left(\psi_i \middle| \mathbf{z}_i\right)$$

  - Laquelle devrait être estimée?

Modèle de Fay-Herriot

- **Prédicteur BLUP de θ_i sous le modèle de Fay-Herriot:**

$$\hat{\theta}_i^{BLUP} = \gamma_i \hat{\theta}_i + (1-\gamma_i)\mathbf{z}_i'\hat{\boldsymbol{\beta}}$$

$$\gamma_i = \sigma_v^2 \Big/ \left(\sigma_v^2 + \tilde{\psi}_i\right)$$

- Le prédicteur BLUP de θ_i dépend de $\tilde{\psi}_i$ et pas de ψ_i
  ⇨ Ce qu'il faut vraiment estimer est donc: $\tilde{\psi}_i$
- Estimation de $\tilde{\psi}_i$ … Prochaines diapositives
- Estimation de σ_v^2 : REML ⇨ EBLUP de θ_i

Modèle de Fay-Herriot

- **Estimation de $\tilde{\psi}_i$:**
  - Estimateur **sans biais sous le plan** de ψ_i : $\hat{\psi}_i$
  - $\tilde{\psi}_i = E_m\left(\psi_i \middle| \mathbf{z}_i\right) = E_m\left(\hat{\psi}_i \middle| \mathbf{z}_i\right)$
  - Estimateur sans biais de $\tilde{\psi}_i$: $\hat{\psi}_i$ ⇨ instable
  - Une autre approche consiste à modéliser $\hat{\psi}_i$
  - Nous avons considéré le modèle de lissage suivant (semblable à celui de Mohadjer et al., 2012):

$$\log\left(\hat{\psi}_i\right) = \mathbf{x}_i'\boldsymbol{\alpha} + \varepsilon_i \quad , \quad \varepsilon_i \rightarrow \left(0 \, , \, \sigma_\varepsilon^2\right)$$

$$\mathbf{x}_i' = \left(1 \, , \, \log(z_{1i}) \, , \log(1 - z_{1i}) , \log(N_i^{15+})\right)$$

Modèle de Fay-Herriot

- De ce modèle log-linéaire, on obtient:

$$\tilde{\psi}_i = E_m\left(\hat{\psi}_i \middle| \mathbf{z}_i\right) = \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)\Delta$$

$$\Delta = E_m\left(\exp\left(\varepsilon_i\right)\middle|\mathbf{z}_i\right)$$

- Estimation de Δ par la méthode des moments:

$$\hat{\Delta}(\boldsymbol{\alpha}) = \frac{\sum_k \hat{\psi}_k}{\sum_k \exp\left(\mathbf{x}_k'\boldsymbol{\alpha}\right)}$$

  - Ne suppose pas la normalité des ε_i
  - Suppose que les ε_i suivent la même distribution

- **Estimateur lisse de la variance** : $\hat{\tilde{\psi}}_i = \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\alpha}}\right)\hat{\Delta}(\hat{\boldsymbol{\alpha}})$

Coefficient de détermination (R^2)

- En réalité, on est intéressé au coefficient de détermination associé au modèle de liaison
- Coefficient de détermination idéal :

$$R^2_{idéal} = 1 - \frac{\sum_i \left(\theta_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_*\right)^2 \Big/ (m-p)}{\sum_i \left(\theta_i - \bar{\theta}\right)^2 \Big/ (m-1)}$$

- Coefficient de détermination naïf : $R^2_{naïf}$
  - Remplace θ_i (inconnu) par $\hat{\theta}_i$ dans le coefficient de détermination idéal
  - Trop petit … Reflète le modèle combiné et non le modèle de liaison

Coefficient de détermination

- On peut ré-écrire $R^2_{idéal}$:

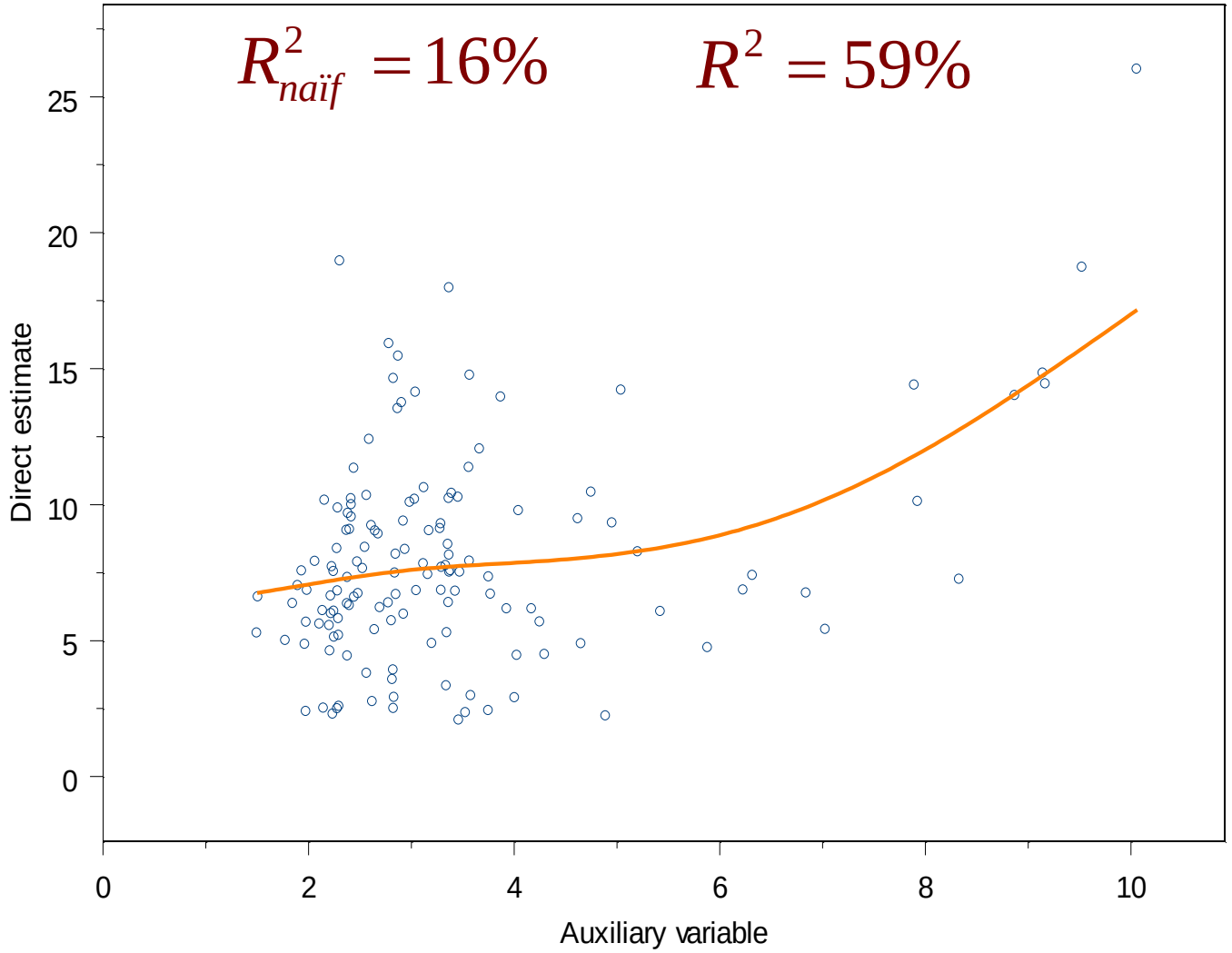
$$R^2_{idéal} = 1 - \frac{\hat{\sigma}^2_{v*}}{\left(\frac{m-p}{m-1}\right)\hat{\sigma}^2_{v*} + \hat{\boldsymbol{\beta}}'_{*}\mathbf{S}_{zz'}\hat{\boldsymbol{\beta}}_{*}} = f(\hat{\boldsymbol{\beta}}_{*}, \hat{\sigma}^2_{v*})$$

- Coefficient de détermination proposé:

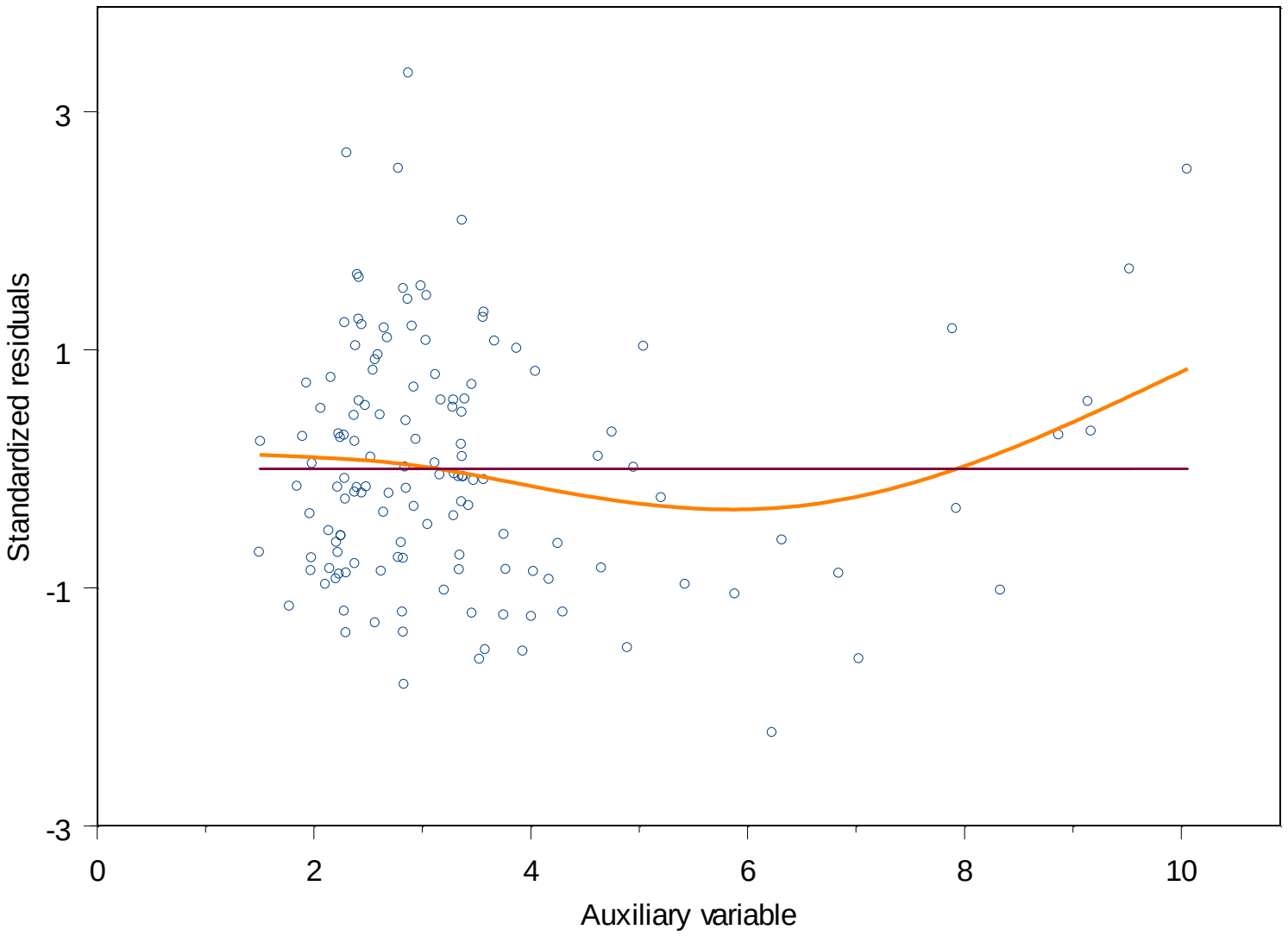
$$R^2 = 1 - \frac{\hat{\sigma}^2_{v}}{\left(\frac{m-p}{m-1}\right)\hat{\sigma}^2_{v} + \hat{\boldsymbol{\beta}}'\mathbf{S}_{zz'}\hat{\boldsymbol{\beta}}} = f(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_{v})$$

Résultats pour Mai 2011

Scatter plot of domain-level data



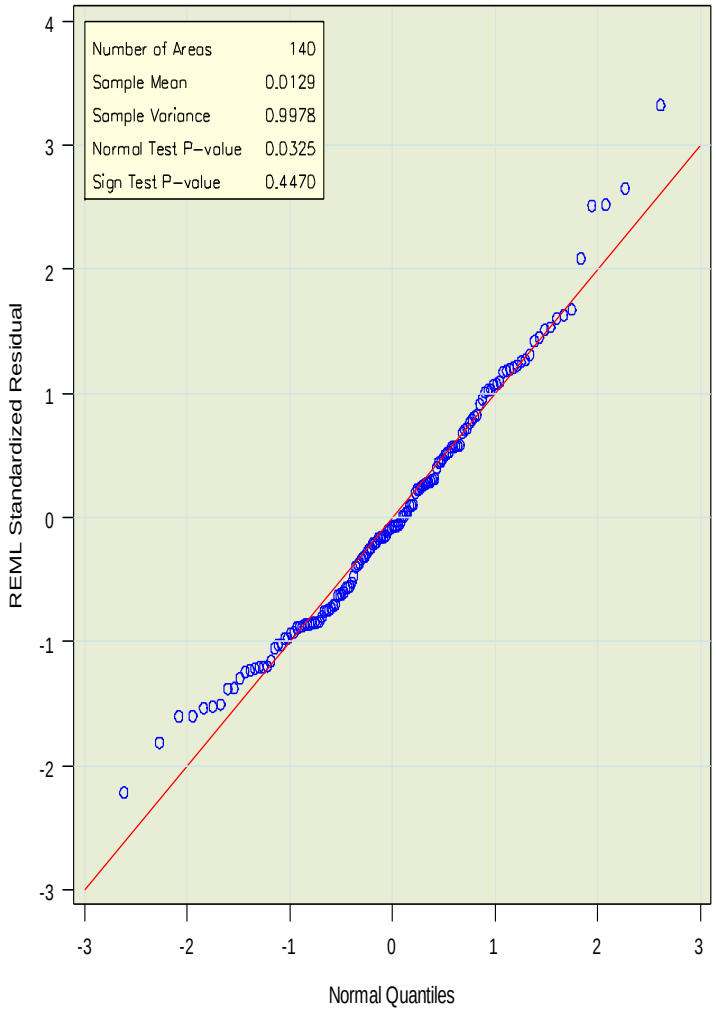
Graph of standardized residuals vs the auxiliary variable



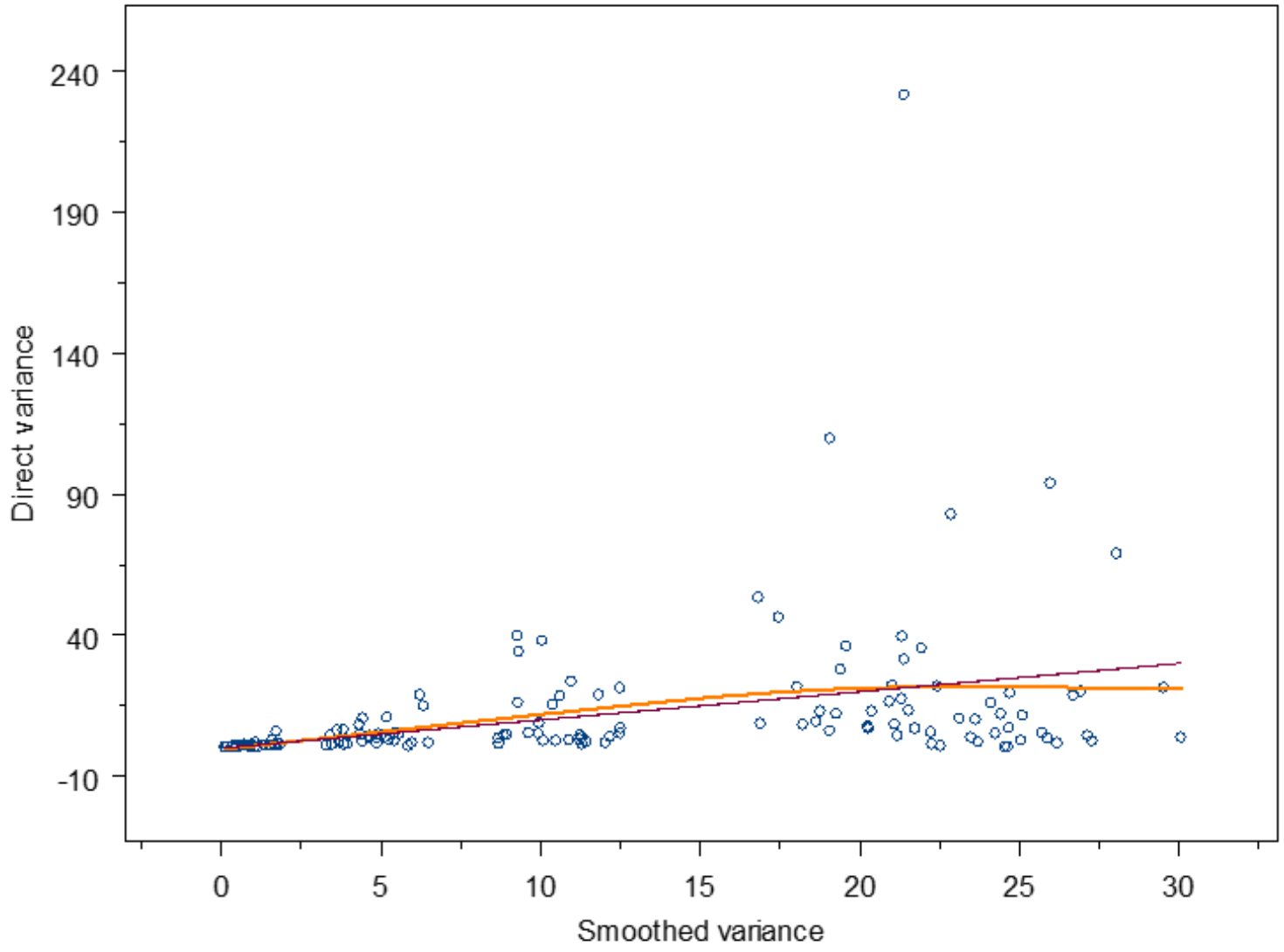
REML Method

QQ-Plot of Standardized Residuals vs. Standard Normal Quantiles

Parameter Definition: unemployment_rate

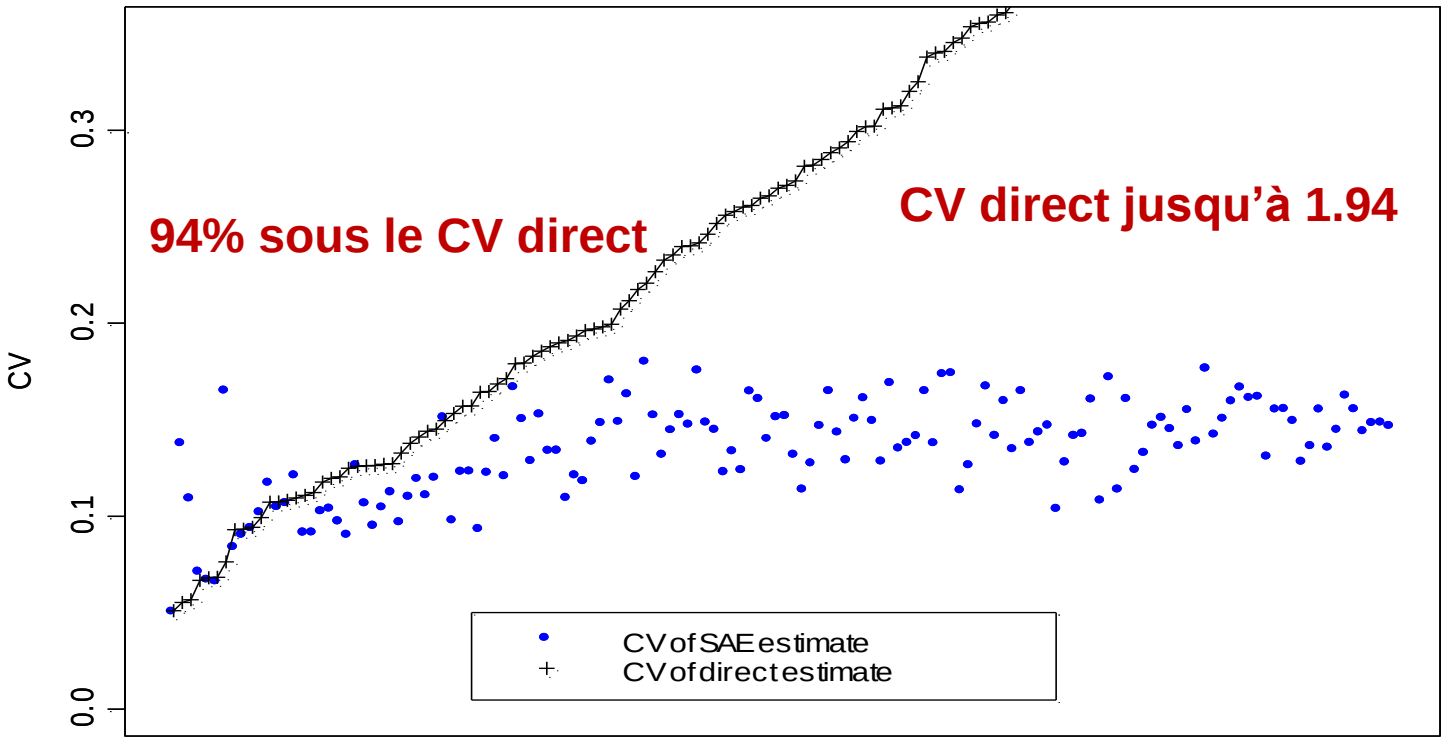


Direct variance vs. smoothed variance



CV	Moy.	Min	Max
CV EBLUP	0.13	0.05	0.18
CV direct	0.34	0.05	1.94

CV of SAE and direct estimates



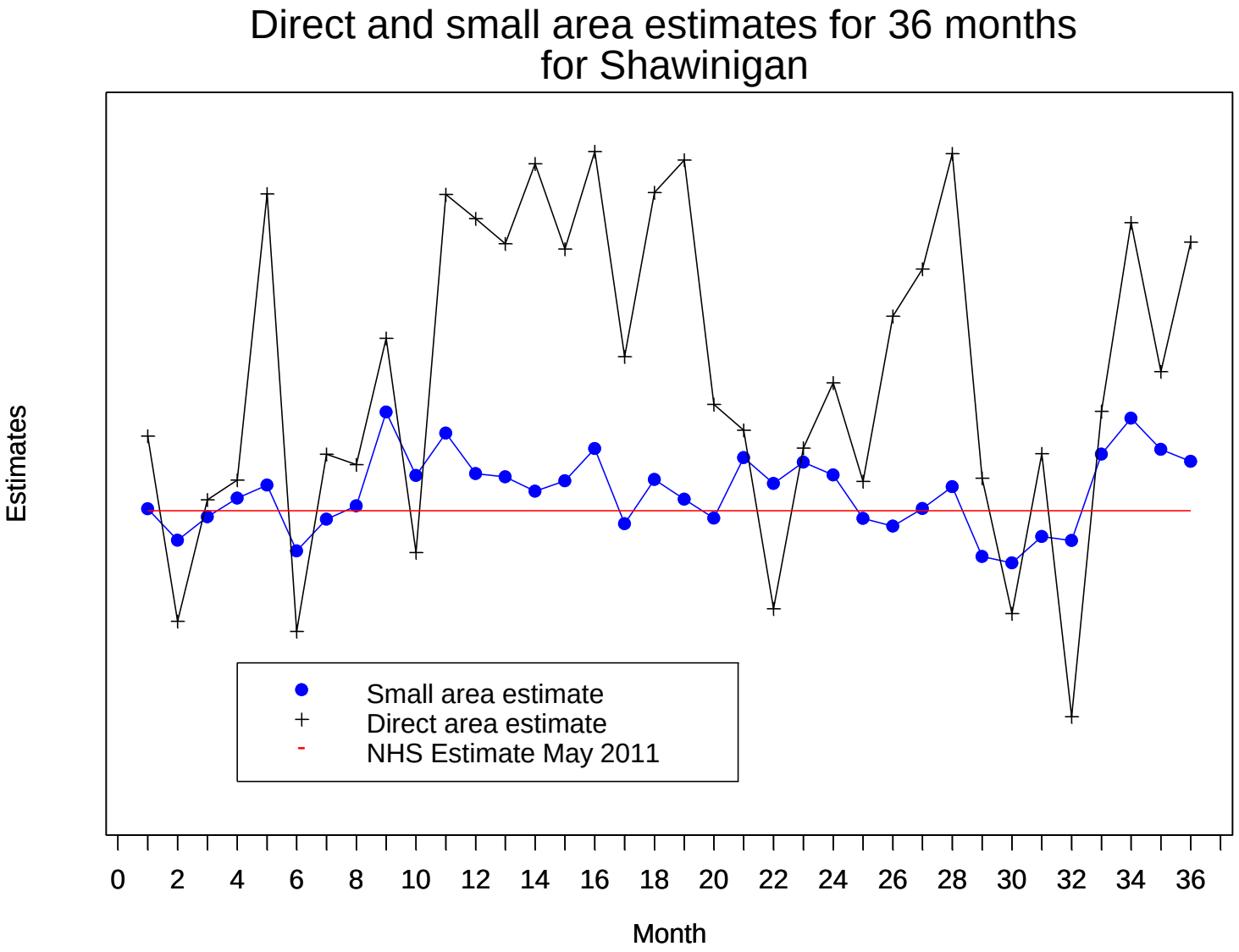
Comparaison avec les estimations de l' ENM, $n_{ENM} >>> n_{EPA}$

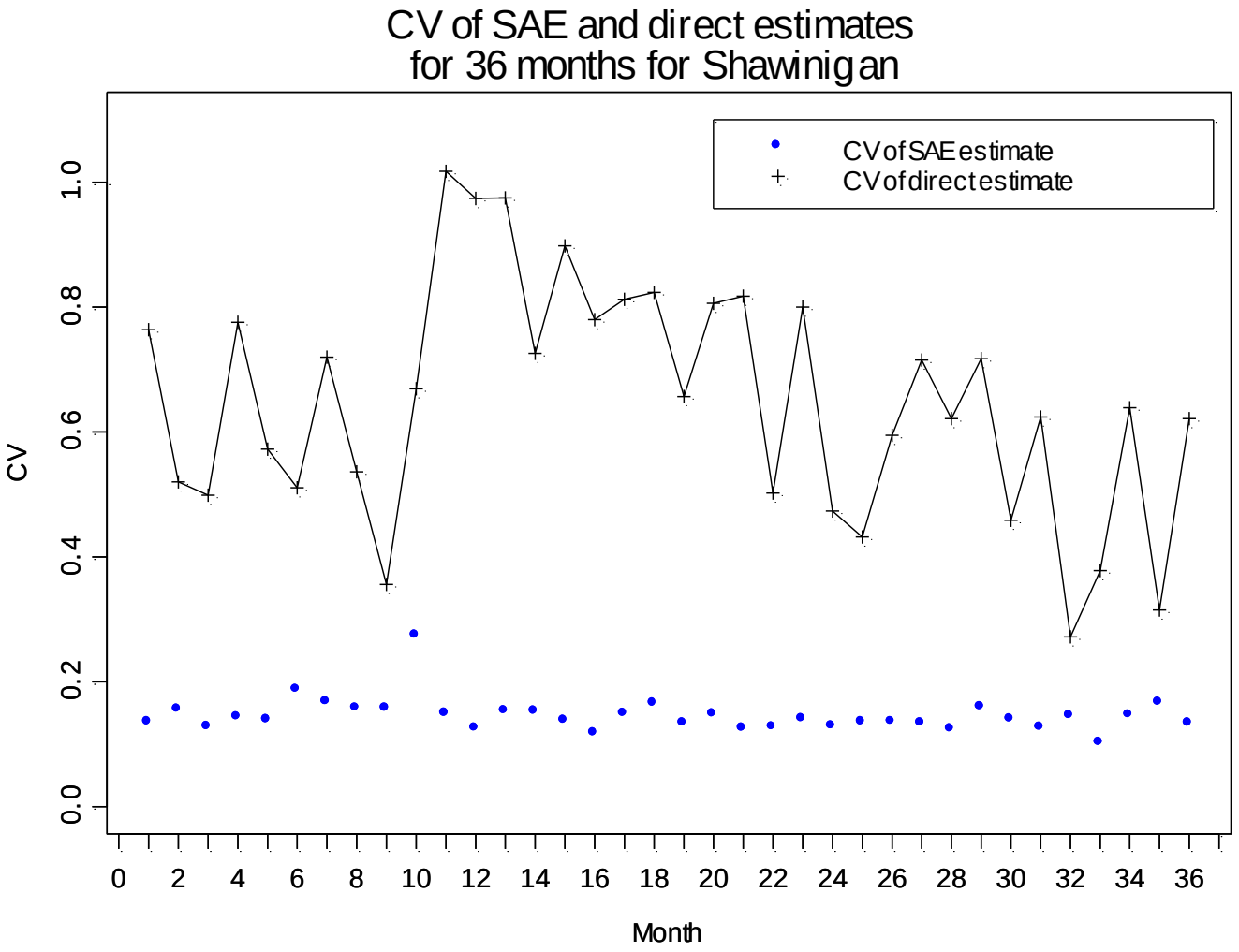
Taille d'échantillon	Dif. Rel. Abs. Direct EPA vs ENM	Dif. Rel. Abs. EBLUP vs ENM
28 plus petites villes	52.4%	17.0%
28 villes suivantes	41.7%	20.5%
28 villes suivantes	33.0%	22.8%
28 villes suivantes	21.5%	17.9%
28 plus grandes villes	10.1%	8.8%
Toutes les villes	31.7%	17.4%

Taille d'échantillon	Dif. Rel. Abs. Direct EPA vs ENM	Dif. Rel. Abs. EBLUP vs ENM (avec lissage)	Dif. Rel. Abs. EBLUP vs ENM (sans lissage)
28 plus petites villes	52.4%	17.0%	24.9%
28 villes suivantes	41.7%	20.5%	20.6%
28 villes suivantes	33.0%	22.8%	19.6%
28 villes suivantes	21.5%	17.9%	18.9%
28 plus grandes villes	10.1%	8.8%	9.5%
Toutes les villes	31.7%	17.4%	18.7%

Résultats pour Mai 2011 à Avril 2014 (36 mois)

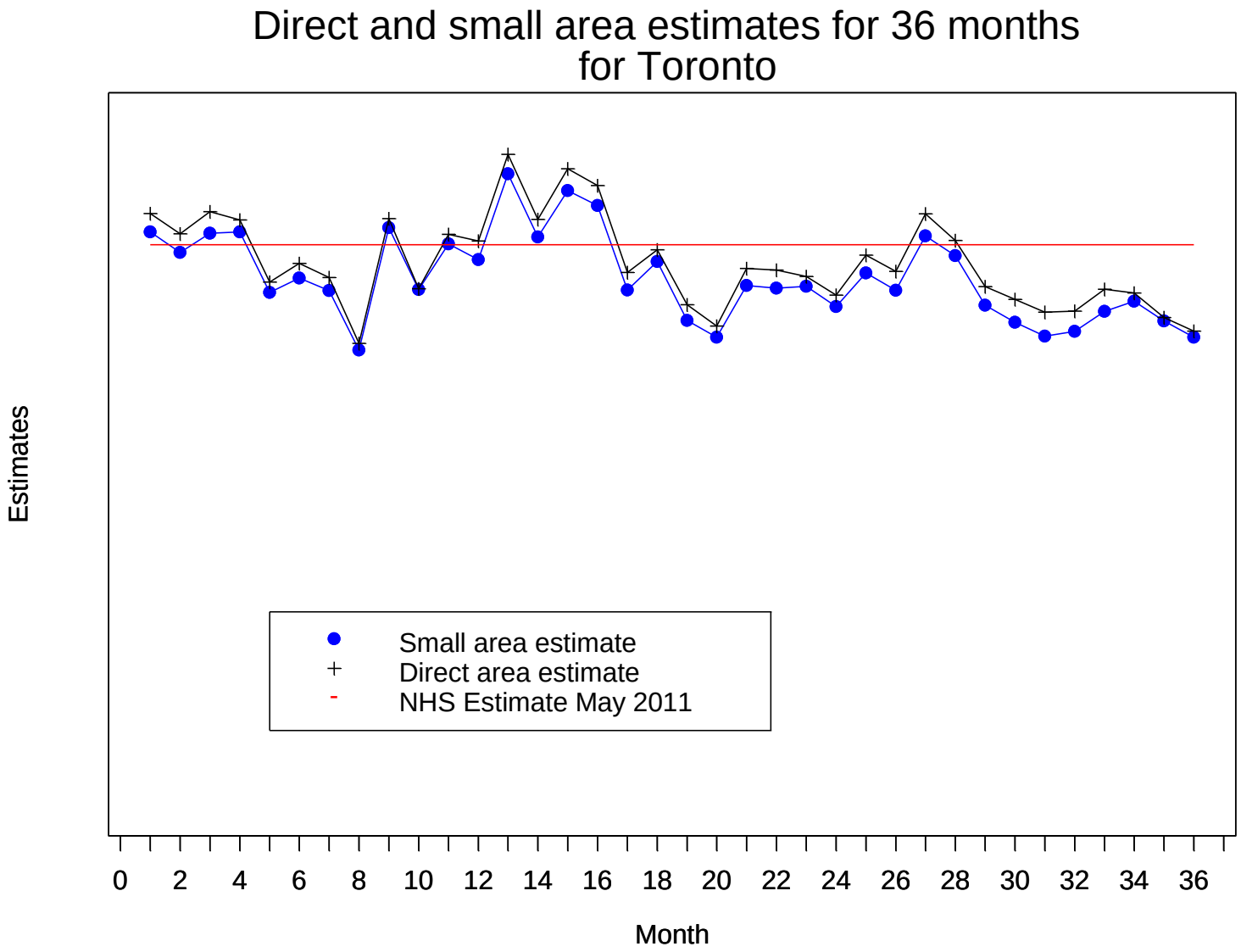
Taille d'échantillon moyenne = 39

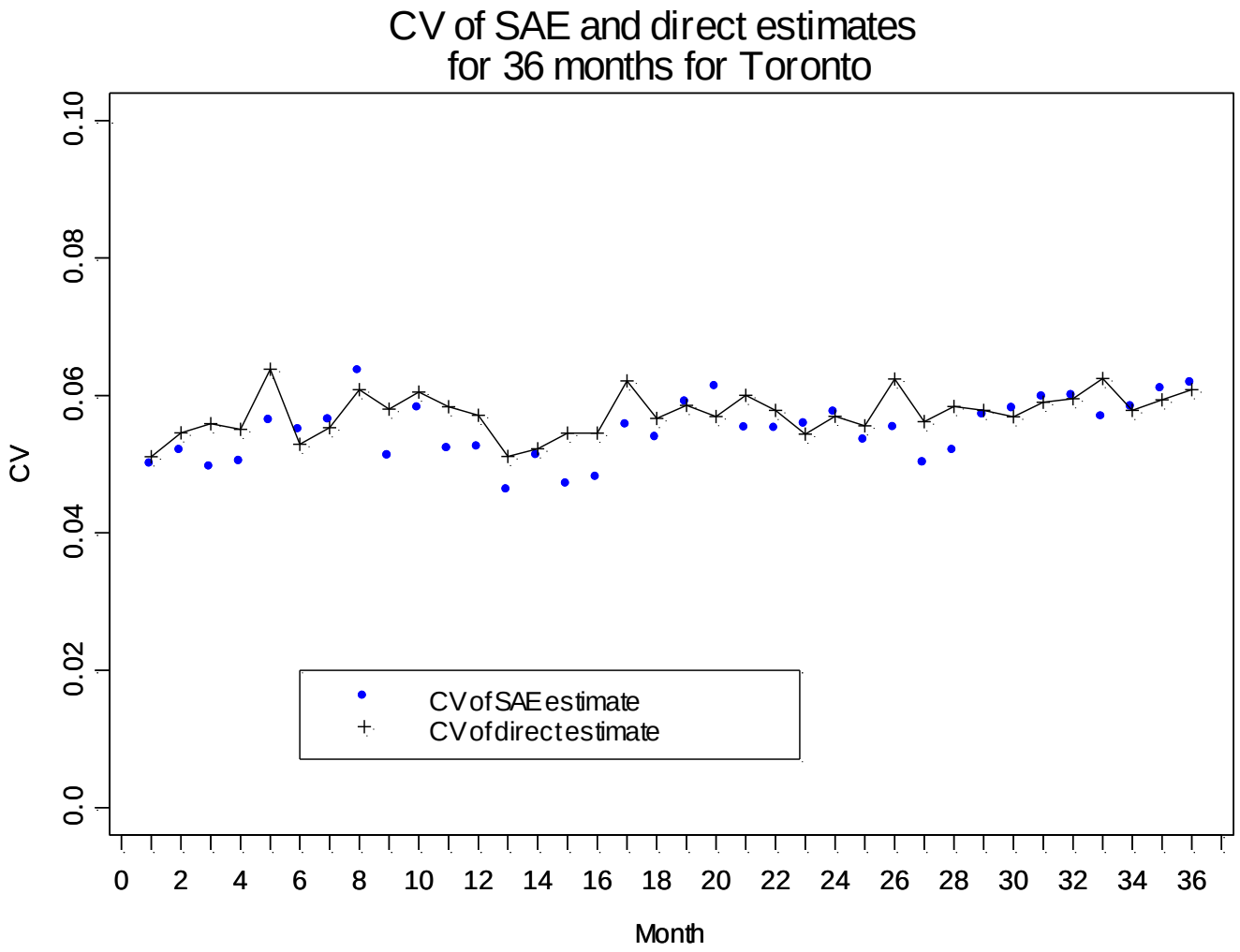




CV EBLUP moyen	**0.14**
CV direct moyen	**0.65**

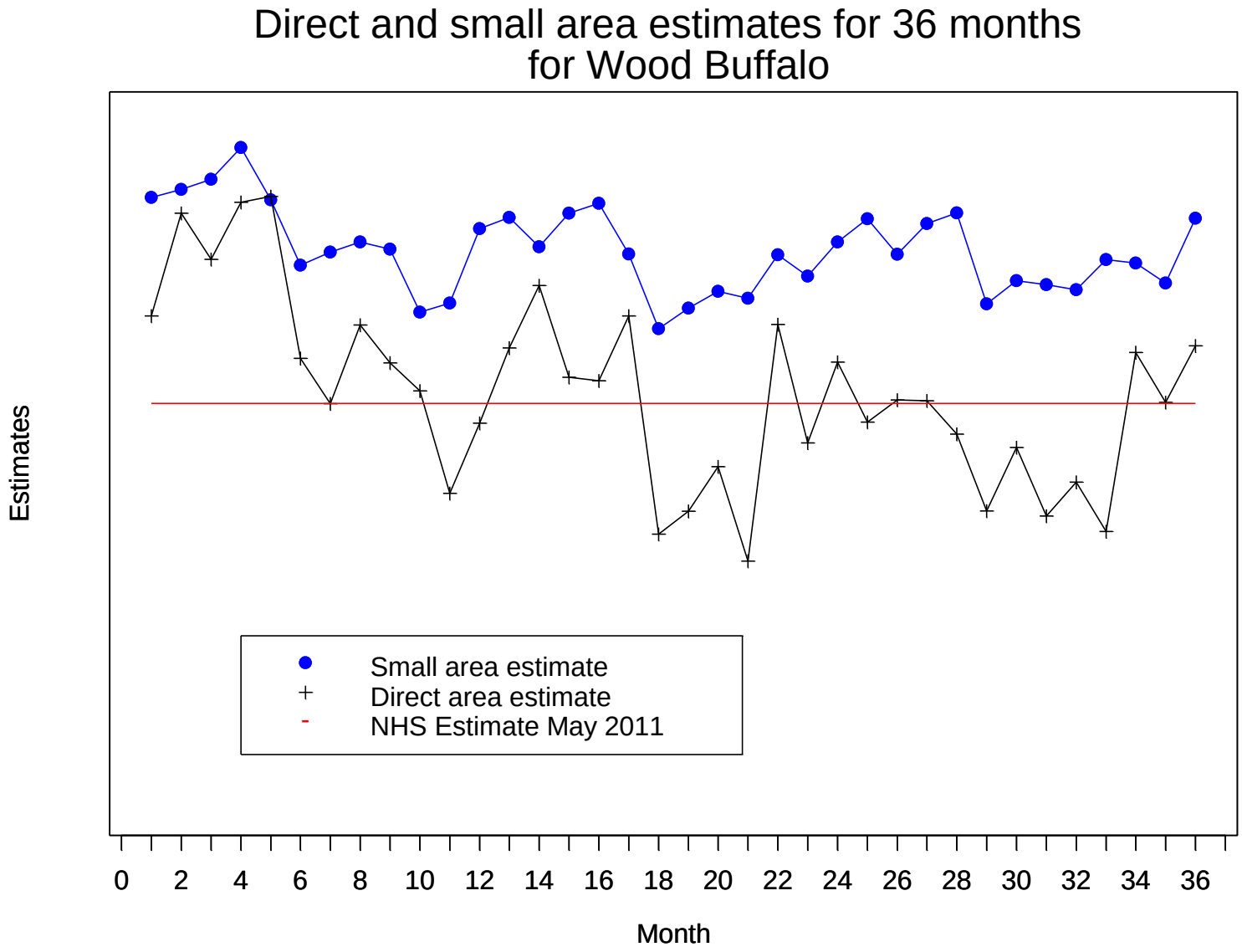
Taille d'échantillon moyenne = 3768

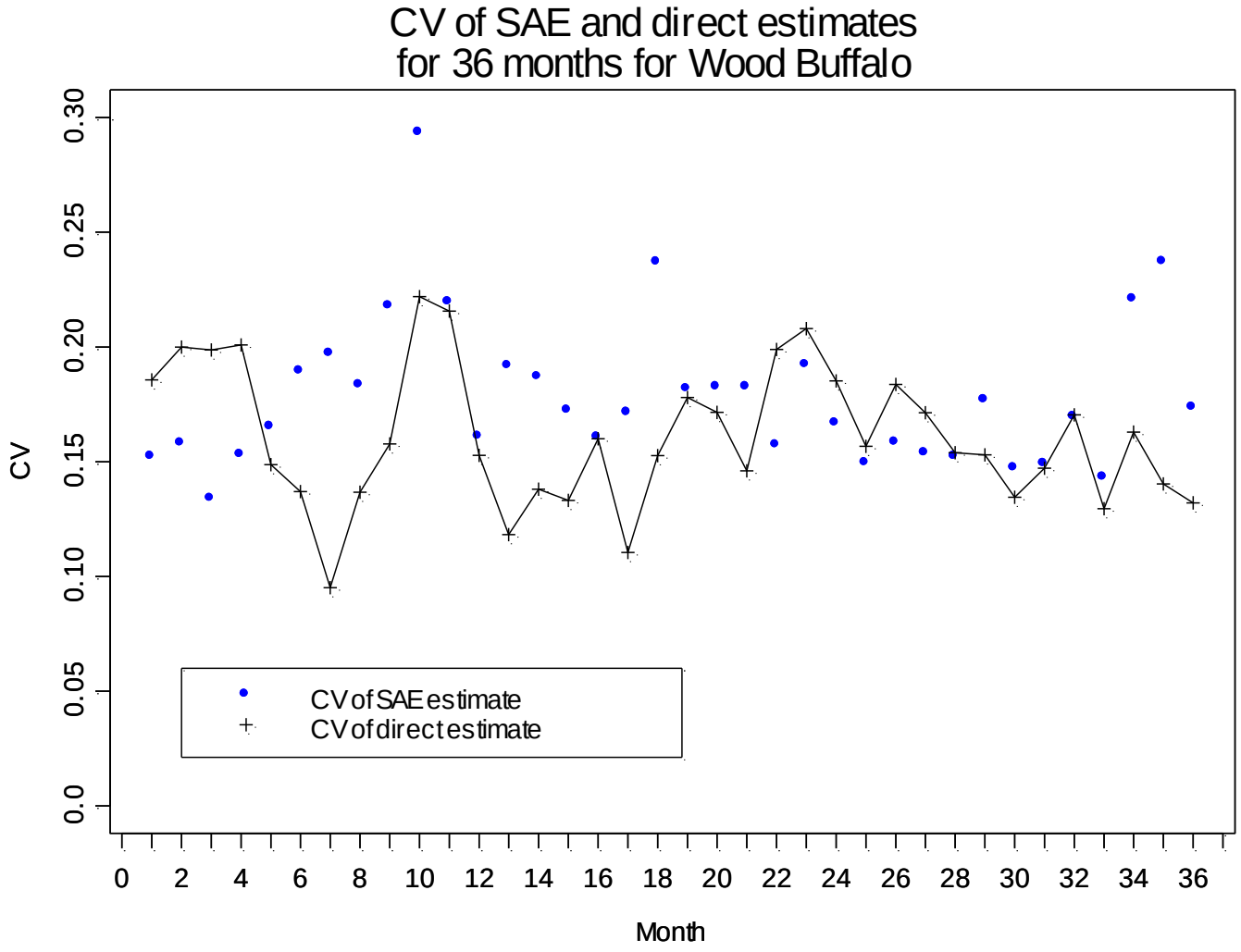




CV EBLUP moyen	**0.055**
CV direct moyen	**0.057**

Taille d'échantillon moyenne = 541





CV EBLUP moyen	**0.18**
CV direct moyen	**0.16**

Conclusion

- L'application du modèle de Fay-Herriot à l'EPA semble prometteuse
- Pour quelques autres applications, nous n'avons pu obtenir de meilleures estimations que les estimations directes. Pourquoi?
  - Non-normalité extrême des résidus (valeurs influentes) **Méthodes robustes pourraient être utiles**
- Un autre défi:
  - Non-linéarité ou valeurs influentes dans le modèle de lissage
  - **Ne pas lisser $\hat{\psi}_i$ a quelques fois donné des améliorations**