Comparaison de précision entre les estimateurs par scission et les estimateurs classiques pour petits domaines

Toky Randrianasolo, Yves Tillé

6 novembre 2012

Université de Neuchâtel unine

IFSTTAR

Plan

UNIVERSITÉ DE NEUCHÂTEL

Présentation basée sur :

- Présentation aux **XI**[es] **JMS** : T. Randrianasolo, Y. Tillé, J. Armoogum (2012). Estimation sur petits domaines par scission des poids,
- Article soumis pour publication : T. Randrianasolo, Y. Tillé (2012). Small Area Estimation by Splitting the Sampling Weights.

Motivation

- Une méthode utilisant des poids.
- Une méthode fournissant des estimations locales cohérentes avec des estimations globales.

UNIVERSITÉ DE NEUCHÂTEL

Estimation directe

- L'estimation directe au niveau d'un domaine consiste à construire un estimateur n'utilisant aucune information extérieure au domaine donné.
- Une unité ne contribue donc qu'à son propre domaine.
- Estimateur calé du total $t_y^d = \sum_{k \in \mathcal{U} \cap A_d} y_k$ au niveau d'un domaine A_d :

$$\hat{t}_{y,w}^d = \sum_{k \in S \cap A_d} w_k y_k ,$$

où les w_k sont les poids de calage.

UNIVERSITÉ DE NEUCHÂTEL

Principe de scission de poids

- Une unité peut contribuer à n'importe quel domaine : nous construisons donc un poids $w_{kd} = w_k q_{kd}$ dépendant de la k^e unité et du domaine A_d,
- w_k est un poids obtenu à partir d'un calage global sur le vecteur de totaux connus $\mathbf{t}_x$.
- Nous voulons des estimations locales cohérentes avec des estimations globales :

$$\sum_{d=1}^{D} \underbrace{\sum_{k \in S} w_{kd} \mathbf{x}_k^\top}_{\mathbf{t}_x^d} = \sum_{k \in S} w_k \mathbf{x}_k^\top = \mathbf{t}_x \Rightarrow \sum_{d=1}^{D} q_{kd} = 1.$$

UNIVERSITÉ DE NEUCHÂTEL

Matrice $\mathbf{Q}$

- Nous considérons la matrice $\mathbf{Q} = \{(q_{kd})_{d=1...D}^{k=1...n}\}$ telle que

$$\sum_{d=1}^{D} q_{kd} = 1,$$

et pour tout domaine A_d,

$$\sum_{k \in S} w_k q_{kd} \mathbf{x}_k^\top = \mathbf{t}_x^d.$$

UNIVERSITÉ DE NEUCHÂTEL

Calcul de la matrice **Q**

- Nous utilisons un algorithme consistant à répéter deux calages successifs sur une matrice initialisée à $\mathbf{Q}^{\{0\}}$:
    1. on cale les colonnes sur les totaux des domaines,
    2. on ajuste les lignes de la matrice pour qu'elle soit stochastique.
- Nous nous arrêtons lorsque la somme des lignes vaut 1 après un calage des colonnes.

UNIVERSITÉ DE NEUCHÂTEL

Estimateur de type extra-contribution

- L'estimateur de type extra-contribution du total $t_y^d = \sum_{k \in \mathcal{U} \cap A_d} y_k$ au niveau d'un domaine A_d est donné par :

$$\hat{t}_{y,q}^d = \sum_{k \in S} w_k q_{kd} y_k$$

tel que

$$\sum_{d=1}^{D} \hat{t}_{y,q}^d = \hat{t}_{y,w}.$$

UNIVERSITÉ DE NEUCHÂTEL

Estimateur composite

- L'estimation composite consiste à mélanger une estimation directe et une estimation de type extra-contribution construite à l'aide de la matrice **Q**.
- L'estimateur composite du total $t_y^d = \sum_{k \in \mathcal{U} \cap A_d} y_k$ au niveau d'un domaine A_d est construit en pondérant à l'aide de α_d :

$$\begin{aligned} \hat{t}_{y,c}^d &= \sum_{k \in S} c_{kd} w_k y_k \\ &= \alpha_d \sum_{k \in S} h_{kd} q_{kd} w_k y_k + (1 - \alpha_d) \sum_{k \in S \cap A_d} h_{kd} w_k y_k, \end{aligned}$$

où $c_{kd} = g_{kd} h_{kd}$, les h_{kd} sont les ajustements résultant du calage de la matrice **G** avec $g_{kd} = \alpha_d q_{kd} + (1 - \alpha_d) \mathbb{1}_{\{k \in A_d\}}$.

- Nous retrouvons $\sum_{d=1}^{D} \hat{t}_{y,c}^d = \hat{t}_{y,w}$.

UNIVERSITÉ DE NEUCHÂTEL

Choix de α_d

- Estimateur BLUP : estimateur de type composite

$$\tilde{t}^{d}_{y,\mathrm{BLUP}} = \gamma_d \left(N_d \overline{y}_d + (\mathbf{t}^d_x - N_d \overline{\mathbf{x}}_d)^\top \tilde{\beta} \right) + (1 - \gamma_d) {\mathbf{t}^d_x}^\top \tilde{\beta}.$$

- En passant par le modèle de régression à erreur emboîtée, une estimation de α_d peut être obtenue

$$\hat{\alpha}_d = 1 - \hat{\gamma}_d.$$

UNIVERSITÉ DE NEUCHÂTEL

Calcul de précision

- 1000 échantillons de taille d'espérance égale à n sont tirés avec des probabilités inégales, parmi une population donnée.
- Pour un estimateur donné $\hat{t}_y^d$, le $\%\mathrm{RRMSE}^d_{\mathrm{sim}}(\hat{t}_y^d)$ est obtenu par

$$\%\mathrm{RRMSE}^d_{\mathrm{sim}}(\hat{t}_y^d) = 100 \times \frac{\sqrt{\mathrm{MSE}^d_{\mathrm{sim}}(\hat{t}_y^d)}}{t_y^d},$$

avec

$$\mathrm{MSE}^d_{\mathrm{sim}}(\hat{t}_y^d) = \mathrm{Biais}_{\mathrm{sim}}(\hat{t}_y^d)^2 + \mathrm{Var}_{\mathrm{sim}}(\hat{t}_y^d).$$

UNIVERSITÉ DE NEUCHÂTEL

Application aux données Suisses

- Population Suisse provenant du recensement de 2003.
- Variables auxiliaires :
  1. **POPTOT** : la population totale,
  2. **H00PTOT** : le nombre de ménages total et
  3. **Pop020+Pop2040** : la population de moins de 40 ans.
- Variable d'intérêt :
  1. **Pop65P** : la population âgée de 65 ans et plus.

UNIVERSITÉ DE NEUCHÂTEL

Comparaison des %RRMSE$_{sim}$ des estimateurs du total de la population de 65 ans et plus

	n_d	Calage global	Synthétique	EBLUP	Extra-contribution	Composite
ZH	31	8.03	37.39	46.82	6.63	2.60
BE	30	12.21	32.24	23.23	2.49	1.58
LU	12	18.58	41.73	21.46	3.54	2.68
UR	1	85.10	15.20	14.81	11.71	9.07
SZ	5	41.44	6.12	8.63	2.33	3.23
OW-NW	3	56.66	5.00	5.41	3.01	3.00
GL	1	83.82	11.23	13.53	11.75	8.32
ZG	4	41.61	6.94	4.22	6.83	5.26
FR	9	28.96	19.31	31.71	2.63	3.16
SO	9	29.04	1.99	20.54	3.71	2.62
BS-BL	11	15.48	39.76	42.96	2.74	1.69
SH	2	39.66	49.74	19.18	1.46	1.77
AR-AI	3	55.10	13.56	6.34	10.89	6.84
SG	15	18.28	27.07	14.66	4.84	2.51
GR	7	34.33	18.21	24.42	3.93	3.80
AG	20	19.28	5.91	6.39	4.36	2.95
TG	9	28.22	6.08	7.92	5.92	3.22
TI	11	27.03	2.25	20.80	1.80	2.09
VD	20	14.82	18.70	9.90	1.05	1.99
VS	10	27.50	16.37	10.83	5.29	2.77
NE	6	29.19	46.15	17.71	2.80	2.31
GE	10	15.90	50.85	65.33	3.97	2.38
JU	3	60.31	6.14	25.51	8.29	5.65

Source : Recensement Suisse de 2003.

UNIVERSITÉ DE NEUCHÂTEL

Discussion, conclusion et perspectives

- L'estimateur proposé peut s'écrire comme un système de poids et peut être appliqué à n'importe quelle variable d'intérêt.
- La constante α_d dépend de la variable d'intérêt. Dans le cas de plusieurs variables d'intérêt d'un même thème, on pourrait prendre la moyenne des constantes.
- La prochaine étape consiste à calculer la variance des estimateurs proposés.

UNIVERSITÉ DE NEUCHÂTEL

Merci pour votre attention

IFSTTAR
Département Économie et Sociologie des Transports
2 rue de la Butte Verte
F-93166 Noisy-le-Grand

Mèl. toky.randrianasolo@ifsttar.fr

UNINE
Pierre à Mazel 7
CH-2000 Neuchâtel

Mèl. toky.randrianasolo@unine.ch

UNIVERSITÉ DE NEUCHÂTEL