Estimation sur petits domaines par scission des poids

Toky Randrianasolo[1,2], Yves Tillé[2] et Jimmy Armoogum[1]

Université Paris-Est, IFSTTAR-DEST[1]
Université de Neuchâtel[2]

XI^es J.M.S.
Paris XIV^e, 25 Janvier 2012

Plan

Motivation

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

Estimation directe

- L'estimation directe au niveau d'un domaine consiste à construire un estimateur n'utilisant aucune information extérieure au domaine donné.
- Un individu ne contribue donc qu'à son propre domaine.

Estimateurs directs classiques

1. Estimateur d'Horvitz-Thompson du total $t_y^d = \sum_{k \in \mathcal{U} \cap A_d} y_k$ au niveau d'un domaine A_d :

$$\hat{t}_{y,\pi}^d = \sum_{k \in S \cap A_d} \frac{y_k}{Pr(k \in S)}$$

2. 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.

Principe de scission de poids

- Un individu peut contribuer à n'importe quel domaine : nous construisons donc un poids $w_{kd} = w_k q_{kd}$ dépendant du k^{e} individu 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 :

$$\sum_{d=1}^{D} \underbrace{\sum_{k \in S} w_{kd} \mathbf{x}'_k}_{\mathbf{t}_x^d} = \sum_{k \in S} w_k \mathbf{x}'_k = \mathbf{t}_x \Rightarrow \sum_{d=1}^{D} q_{kd} = 1.$$

Matrice Q

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

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

et pour tout domaine A_d,

$$\sum_{k \in S} w_k q_{kd} \mathbf{x}'_k = \mathbf{t}_x^d.$$

Calcul de la matrice $\mathbf{Q}$

- 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.

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 :

$$\widehat{t}_{y,\text{extra}}^{d} = \sum_{k \in S} w_k q_{kd} y_k$$

tel que

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

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 $\mathbf{Q}$.
- 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 :

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

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

- Nous retrouvons

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

Choix de α_d

- Nous voulons déterminer la valeur de α_d qui minimise la variance de notre estimateur composite.
- Le poids α_d dépend de l'effectif du domaine A_d dans l'échantillon :

$$\begin{aligned} \alpha_d(n_d) &\simeq \frac{1}{1+\frac{\sigma^2_{d,1}}{\sigma^2_{d,2}}\frac{n_d}{n}} \\ &\simeq \frac{1}{1+\vartheta_d\frac{n_d}{n}} \end{aligned}$$

où $\sigma^2_{d,1}/n$ est la variance de la partie extra-contribution de notre estimateur composite, et $\sigma^2_{d,2}/n_d$ la variance de la partie domaine.

Calcul de précision

1. Chaque échantillonnage se fait en tirant un échantillon de taille d'espérance égale à n avec des probabilités inégales. Nous obtenons ensuite pour chaque échantillonnage une quantité

$$\varepsilon = \frac{\hat{t} - t}{t} \times 100$$

où $\hat{t}$ représente le total estimé et t le vrai total.

2. L'erreur quadratique moyenne s'obtient par :

$$\mathrm{MSE}(\hat{t}) = \left(\frac{t}{100}\right)^2 \mathbb{E}(\varepsilon^2),$$

3. et la racine relative de la MSE par :

$$\mathrm{RRMSE}(\hat{t}) = \frac{\sqrt{\mathrm{MSE}(\hat{t})}}{t}.$$

Application aux données Suisses

- Population Suisse au niveau des municipalités provenant du recensement de 2003. Nous disposons de 2896 communes comme unités d'observation suivant 22 variables.
- Variables auxiliaires :
  1. le nombre de communes par canton,
  2. **POPTOT** : la population totale,
  3. **H00PTOT** : le nombre de ménages total et
  4. **Pop020**+**Pop2040** : la population de moins de 40 ans.
- Variables d'intérêt :
  1. **Pop020** : la population âgée entre 0 et 19 ans,
  2. **Pop2040** : la population âgée entre 20 et 39 ans,
  3. **Pop4065** : la population âgée entre 40 et 64 ans et
  4. **Pop65P** : la population âgée de 65 ans et plus.

Simulation

- 100 simulations pour des valeurs des constantes $\vartheta = (\vartheta_1, \ldots, \vartheta_d, \ldots, \vartheta_D)$ égales à $0 \times \mathbf{1}_D, \ldots, 10 \times \mathbf{1}_D$ où $\mathbf{1}_D$ est un vecteur unité d'ordre D, pour un taux de sondage à 8%.
- Les RRMSEs diminuent lorsque les ϑ augmentent. Nous prenons alors des valeurs de ϑ très grandes égales à $10 \times \mathbf{1}_D, 20 \times \mathbf{1}_D, \ldots, 100 \times \mathbf{1}_D$.

Résultats pour la variable **Pop65P** (1)

RRMSE pour la variable **Pop65P**										
ϑ	0	1	2	3	4	5	6	7	8	9
ZH	6.53	5.92	5.46	5.16	4.88	4.56	4.31	4.11	4.05	3.88
BE	2.62	2.29	1.82	1.67	1.85	1.61	1.62	1.77	1.46	1.60
LU	3.74	3.44	3.32	3.03	3.02	2.91	3.03	3.06	2.65	2.53
UR	11.89	11.67	11.25	10.91	11.30	10.68	10.90	10.88	10.61	10.53
SZ	2.59	2.34	2.37	2.11	1.99	1.81	2.15	2.05	2.03	1.86
OW-NW	3.24	3.00	2.89	2.65	2.51	2.29	2.49	2.72	2.45	2.31
GL	11.91	11.57	11.19	10.91	11.04	10.55	10.79	10.67	10.40	10.53
ZG	6.47	6.61	6.71	6.74	6.75	6.48	6.22	6.34	6.32	6.52
FR	2.83	2.58	2.43	2.58	2.41	2.25	2.65	2.31	2.56	2.47
SO	3.88	3.50	3.14	2.96	2.95	2.70	2.79	2.67	2.40	2.60
BS-BL	2.58	2.68	2.69	2.69	2.55	2.62	2.41	2.30	2.40	2.34
SH	1.58	1.43	1.31	1.15	1.35	1.20	1.44	1.17	1.12	1.27
AR-AI	11.08	10.84	10.30	9.98	10.05	9.64	9.68	9.88	9.17	9.27
SG	5.00	4.59	4.28	3.91	3.70	3.38	3.55	3.49	3.39	3.12
GR	4.07	3.76	3.39	3.32	3.45	2.96	3.81	3.25	3.19	3.45
AG	4.26	4.20	4.07	4.01	3.77	3.88	3.43	3.18	3.60	3.15
TG	6.07	5.58	5.36	4.90	4.80	4.39	4.60	4.57	4.13	3.98
TI	2.01	1.43	1.91	1.34	1.75	1.61	1.84	1.56	1.47	1.46
VD	1.11	1.08	1.01	1.16	1.13	1.21	1.30	1.15	1.44	1.37
VS	5.42	5.00	4.60	4.24	4.48	3.91	4.28	4.02	3.68	3.78
NE	2.90	2.70	2.53	2.35	2.44	2.23	2.40	2.33	2.13	2.26
GE	3.84	3.76	3.79	3.76	3.65	3.67	3.59	3.30	3.37	3.38
JU	8.49	8.09	7.63	7.39	7.65	7.04	7.52	7.08	6.93	7.33

Résultats pour la variable **Pop65P** (2)

RRMSE pour la variable **Pop65P**										
ϑ	10	20	30	40	50	60	70	80	90	100
ZH	3.74	3.08	2.48	2.38	2.14	2.22	8.00	9.45	13.35	9.56
BE	1.56	1.30	1.63	1.63	1.63	1.62	12.21	12.11	16.96	12.18
LU	2.66	2.53	2.88	2.56	2.49	2.80	11.90	9.92	14.94	11.94
UR	10.49	9.50	9.57	9.32	9.01	9.04	19.46	18.16	25.46	19.23
SZ	2.06	2.46	2.79	2.85	2.70	3.50	17.59	15.90	17.68	17.69
OW-NW	2.38	2.57	2.54	2.70	2.62	3.11	17.51	16.20	23.33	17.59
GL	10.13	9.26	8.87	8.76	8.95	8.47	19.16	18.89	25.40	19.01
ZG	6.42	5.69	5.11	5.48	5.18	5.01	18.01	16.08	24.34	17.97
FR	2.38	2.83	2.81	2.87	3.40	3.12	14.73	11.76	19.10	14.74
SO	2.57	2.48	2.36	2.63	2.85	2.87	17.49	17.36	24.47	17.47
BS-BL	2.26	2.12	1.99	1.96	1.76	1.70	10.13	8.32	10.13	10.06
SH	1.34	1.33	1.60	1.33	1.52	1.81	8.99	8.82	12.45	8.93
AR-AI	9.01	7.99	8.16	7.32	7.05	7.28	18.54	17.76	24.77	18.52
SG	3.19	2.62	2.67	2.56	2.45	2.41	11.61	12.89	19.00	13.86
GR	3.12	3.54	3.56	3.66	3.66	4.31	14.73	13.94	19.88	14.40
AG	3.36	3.04	3.15	2.66	2.80	2.77	17.52	16.53	24.35	17.58
TG	3.95	3.32	3.38	2.94	2.78	2.53	17.61	14.46	24.11	17.51
TI	1.63	1.96	2.18	2.05	2.08	2.29	17.10	17.41	24.45	17.44
VD	1.36	2.11	2.20	2.05	2.12	1.89	13.65	12.88	19.05	13.40
VS	3.64	3.07	3.12	2.44	2.74	2.75	15.83	15.11	21.66	15.58
NE	2.24	2.04	2.02	2.14	2.01	2.24	9.86	9.76	13.58	9.87
GE	3.16	2.93	2.61	2.61	2.41	2.23	8.37	6.92	8.26	
JU	6.85	6.27	6.70	6.41	5.86	6.18	18.13	17.43	24.55	17.97

Conclusion et perspectives

- Pour chaque domaine, les RRMSEs en fonction de ϑ_d ont une certaine forme parabolique admettant un minimum. Il est intéressant de voir à quoi correspond la valeur de ϑ_d avec laquelle le minimum des RRMSEs est atteint.
- Le calage d'une matrice sur les lignes et les colonnes peut être vu comme un calage d'une nouvelle matrice largement plus grande mais uniquement sur les colonnes. De ce fait, le calcul de la variance de l'estimateur composite peut se faire par la technique de la linéarisation répétée deux fois.