AMÉLIORATION DE LA PRÉCISION D'ESTIMATEURS SUR PETITS DOMAINES À L'AIDE D'UN MODÈLE À EFFETS FIXES

Ketty ATTAL-TOUBERT, Olivier SAUTORY

INSEE - 18, Boulevard Adolphe Pinard 75675 Paris Cedex 14, FRANCE

RÉSUMÉ

Dans ce papier, nous présentons une méthode d'estimation sur petits domaines pour un vecteur de paramètres d'intérêt. L'estimateur proposé repose sur un modèle à effets fixes, qui signifie qu'il existe une certaine "régularité" sur les composantes du vecteur, qui ne dépendent que d'un petit nombre de paramètres. Il peut être vu comme un estimateur *composite*, car il s'exprime comme une moyenne pondérée entre l'estimateur *direct* et l'estimateur *lissé* donné par le modèle. Sa variance peut être largement inférieure à celle de l'estimateur direct. La méthode a été appliquée aux données de l'enquête emploi de l'INSEE pour estimer des taux d'activité régionaux par sexe et tranche d'âge.

SUMMARY

In this paper we present a method to estimate a vector of population parameters for small areas. The proposed estimator is based upon a fixed-effect model, which means that some "regularity" exists between the components of the vector, which depend only on few parameters. It can be seen as a *composite* estimator, since it is obtained as a weighted average of the *direct* estimator and the *smoothed* estimator provided by the model. Its variance can be largely smaller than the variance of the direct estimator. The method has been applied to the French Labour Force Survey to estimate regional rates of activity by sex and group of age.

Les méthodes d'estimation sur petits domaines

Parmi les nombreuses classifications des méthodes d'estimation sur petits domaines, on peut retenir celle-ci (voir par exemple [4]) : les **estimateurs directs** reposent sur les données de l'enquête provenant uniquement du domaine ; les **estimateurs synthétiques** reposent sur l'hypothèse selon laquelle le petit domaine ressemble d'une certaine façon à un domaine plus grand qui le contient ; ils utilisent donc de l'information concernant la variable étudiée (et d'autres variables) provenant d'autres domaines ; les **estimateurs combinés** s'obtiennent en faisant la moyenne pondérée d'un estimateur direct et d'un estimateur synthétique.

La méthode présentée ici, proposée par J.-C. Deville [2], peut se ranger dans cette dernière catégorie de méthodes d'estimation.

Le problème - Les hypothèses

Soit Y un vecteur de p paramètres d'intérêt, dont on veut estimer les valeurs sur R régions, notées $(Y_r, r = 1 \dots R)$, à partir d'un échantillon tiré d'une population. On note $\hat{Y}_r$ le vecteur estimateur *direct* de Y_r, supposé (approximativement) sans biais. On note V_r la matrice estimée de covariance d'échantillonnage du vecteur $\hat{Y}_r$, supposée connue. V_r dépend en particulier de la taille de l'échantillon dans la région r, et peut donc être élevée si cette taille est faible.

On peut donc écrire :

$$\hat{Y}_r = Y_r + e_r$$

où les e_r sont indépendants et vérifient $E\, e_r = 0$, $Ve_r = V_r$.

Avec une hypothèse de normalité sur les e_r, la densité du vecteur $\hat{Y}$ s'écrit :

$$f(\hat{Y} / Y) = \prod_r f(\hat{Y}_r / Y_r) = \prod_r \frac{1}{(2\pi)^{\frac{p}{2}} \sqrt{\det(V_r)}} \exp\left(-\frac{1}{2} (\hat{Y}_r - Y_r)' (V_r)^{-1} (\hat{Y}_r - Y_r) \right)$$

D'autre part, on suppose qu'il existe une certaine "régularité" concernant ces Y_r, que l'on va traduire sous la forme d'un modèle d'analyse factorielle (ACP) à effets fixes (voir H. Caussinus, [1]). Selon ce modèle, les Y_r sont des vecteurs aléatoires indépendants, de même matrice de covariance $\sigma^2 \Gamma$ (Γ connue), et dont les espérances Y_r^* appartiennent à un sous-espace L de $\mathbb{R}^p$ de dimension q fixée.

On peut donc écrire :

$$Y_r = Y_r^* + \varepsilon_r \,,\; Y_r^* \in L$$

où les ε_r sont des vecteurs aléatoires indépendants vérifiant $E\,\varepsilon_r = 0$, $V\varepsilon_r = \sigma^2 \Gamma$.

Avec une hypothèse de normalité sur les ε_r, la densité du vecteur Y s'écrit :

$$g(Y / Y^*, \sigma^2) = \prod_r g(Y_r / Y_r^*, \sigma^2) = \prod_r \frac{1}{(2\pi)^{\frac{p}{2}} \sqrt{\det(\sigma^2 \Gamma)}} \exp\left(-\frac{1}{2} (Y_r - Y_r^*)' (\sigma^2 \Gamma)^{-1} (Y_r - Y_r^*) \right)$$

Ceci peut s'interpréter dans une optique bayésienne : on modélise une "loi a priori" sur les paramètres Y_r, dépendant en particulier d'un paramètre inconnu σ^2.

Le modèle complet s'écrit donc :

$$(M) \quad \begin{cases} \hat{Y}_r = Y_r + e_r & Ve_r = V_r \\ Y_r = Y_r^* + \varepsilon_r \qquad Y_r^* \in L & V\varepsilon_r = \sigma^2 \Gamma \end{cases}$$

Les estimations

Il s'agit d'estimer les Y_r (par $\tilde{Y}_r$), les Y_r^*, L et σ^2.

D'après les résultats établis en [1], si on utilise la méthode du maximum de vraisemblance en supposant la distribution des Y_r gaussienne :

- L est estimé par le sous-espace propre $\hat{L}$ de dimension q obtenu en effectuant l'ACP des Y_r avec la métrique Γ^{-1} ; Γ est la matrice diagonale de taille p contenant les inverses des variances empiriques des variables composant le vecteur Y.
- Y_r^* est estimé par la projection orthogonale (avec la métrique Γ^{-1}) de $\tilde{Y}_r$ sur $\hat{L}$, que l'on notera également Y_r^* pour ne pas alourdir les notations : Y_r^* est donc une combinaison linéaire des q vecteurs engendrant les q premiers axes principaux de l'ACP des Y_r.

Pour estimer les Y_r, on écrit la densité du modèle complet :

$$\prod_r f(\hat{Y}_r / Y_r)\ g(Y_r / Y_r^*, \sigma^2)$$

La méthode du maximum de vraisemblance conduit à :

$$\tilde{Y}_r = V_r\left(V_r + \sigma^2\ \Gamma\right)^{-1}\ Y_r^* + \sigma^2\ \Gamma\left(V_r + \sigma^2\ \Gamma\right)^{-1}\hat{Y}_r \qquad (1)$$

Il reste à estimer σ^2. Pour cela, comme dans les estimations de type "Bayes empirique", on se place dans le "modèle marginal" :

$$\hat{Y}_r = Y_r^* + e_r + \varepsilon_r$$

On égalise le carré de la norme du vecteur $\hat{Y} - Y^*$ à son espérance, ce qui donne :

$$\sigma^2 = \frac{1}{R\ Tr(\Gamma)}\left[\sum_r \left(\hat{Y}_r - Y_r^*\right)'\left(\hat{Y}_r - Y_r^*\right) - \sum_r T_r\left(V_r\right)\right] \quad (2)$$

L'algorithme de résolution est donc le suivant :

(i) effectuer l'ACP normée des $\hat{Y}_r$ (i.e. avec la métrique "inverse des variances"), ce qui donne une estimation de $\hat{L}$ et des Y_r^*

(ii) utiliser la formule (2) pour estimer σ^2

(iii) utiliser la formule (1) pour estimer les $\tilde{Y}_r$

(iv) retourner en (i) en utilisant les $\tilde{Y}_r$ au lieu des $\hat{Y}_r$ pour effectuer l'ACP, ... jusqu'à convergence de σ^2.

Propriétés de $\tilde{Y}_r$

- $\tilde{Y}_r$ est un estimateur combiné, moyenne de l'estimateur direct $\hat{Y}_r$ et de l'estimateur "lissé" Y_r^*, d'autant plus proche de $\hat{Y}_r$ que la précision de l'estimateur direct (matrice V_r^{-1}) est grande par rapport à la "qualité" de l'ajustement par le modèle ACP (matrice $(\sigma^2\ \Gamma)^{-1}$).

- $E_{(M)}(\tilde{Y}_r) = Y_r$

- $V_{(M)}(\tilde{Y}_r) = V_r\left(V_r + \sigma^2\ \Gamma\right)^{-1}\sigma^2\ \Gamma$

Commentaires sur la méthode

Une particularité de la méthode proposée, par rapport aux méthodes courantes d'estimation sur petits domaines, est qu'elle n'utilise pas d'information auxiliaire. Elle repose sur une hypothèse de régularité des composantes du vecteur des variables d'intérêt, qui est révélée par l'analyse en composantes principales : c'est l'idée que les p composantes du vecteur peuvent être réduites à q (= 2 ou 3 en général) facteurs synthétiques, qui déterminent le vecteur "régularisé", ou "lissé". Dans une région donnée, les écarts trop importants entre le vecteur estimé directement sur l'échantillon et le vecteur "régularisé" sont corrigés (i.e. réduits), et ce d'autant plus que la taille de la région est petite.

La méthode présentée ressemble à celle proposée par Fay-Herriot [3], à la différence près que ces auteurs modélisent les Y_r sous la forme d'un modèle de régression, au niveau régional, sur un certain nombre de totaux de variables auxiliaires supposés connus.

Application

Cette méthode a été appliquée aux données issues de l'enquête emploi de l'INSEE, pour estimer des taux d'activité régionaux par sexe et tranche d'âge. Le tableau suivant présente les résultats obtenus pour la région Limousin (région la plus petite en termes d'effectifs).

Notations :

$\hat{Y}$: estimateur direct de Y, $\sigma_{\hat{Y}}$: précision de $\hat{Y}$ (écart-type estimé)

Y^* : estimateur lissé donné par le modèle, obtenu au bout des 18 itérations de l'algorithme de résolution (on a utilisé les deux premiers axes principaux de l'ACP), σ_{Y^*} : précision de Y^*

$\tilde{Y}$: estimateur composite, $\sigma_{\tilde{Y}}$: précision de $\tilde{Y}$

Rapport : rapport des variances (en %) $\left(\sigma_{\tilde{Y}} / \sigma_{\hat{Y}}\right)^2 \times 100$

REGION : Limousin

Taux	**Effectifs**	$\hat{\mathbf{Y}}$	$\sigma_{\hat{Y}}$	$\mathbf{Y}^*$	σ_{Y^*}	$\tilde{\mathbf{Y}}$	$\sigma_{\tilde{Y}}$	**Rapport**
15-24-F	364	27,3	2,76	26,4	1,03	26,6	0,95	12,0
15-24-H	377	30,9	2,92	33,5	1,40	32,9	1,25	18,3
25-39-F	593	84,1	1,52	81,6	1,52	82,9	1,06	48,9
25-39-H	554	96,4	0,78	96,2	0,33	96,1	0,30	15,1
40-49-F	412	85,6	1,94	83,3	1,60	84,8	1,19	37,6
40-49-H	415	95,7	1,10	96,7	0,39	96,6	0,36	10,9
50-59-F	286	62,6	3,13	65,1	2,23	63,8	1,80	33,0
50-59-H	276	79,3	2,50	83,1	1,49	81,9	1,24	24,6
60ET+F	931	4,4	0,70	4,2	0,33	4,2	0,29	17,7
60ET+H	731	5,1	0,85	6,2	0,71	5,8	0,53	39,7

Bibliographie

[1] H. CAUSSINUS (1984). Analyses en composantes principales. Quelques réflexions sur la part des modèles probabilistes, *Publications du Laboratoire de Statistique et Probabilités*, Université Paul Sabatier, Toulouse.

[2] J.-C. DEVILLE (1985). Nature et fonction des modèles pour l'analyse des données socio-démographiques, *Actes des Quatrièmes Journées Internationales, Analyse de Données et Informatique*, INRIA, Versailles.

[3] R. E. FAY III & R. A. HERRIOT (1979). Estimates of income for small places : an application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, pp. 269 - 277.

[4] M.-P. SINGH, J. GAMBINO, H.-J. MANTEL (1994). Les petites régions : problèmes et solutions, *Techniques d'enquête*, 20, n° 1, pp. 3 - 23.