Techniques d'enquête, juin 1989
Vol. 15, n° 1, pp. 3–13
Statistique Canada

Estimation robuste pour petits domaines à l'aide du modèle des effets aléatoires

PHILLIP S. KOTT[1]

RÉSUMÉ

Dans cet article, l'auteur utilise un modèle des effets aléatoires pour construire un estimateur convergent selon le plan pour un petit domaine. Il évalue ensuite l'erreur quadratique moyenne de cet estimateur *sans* supposer que la composante d'effet aléatoire du modèle est juste. À l'aide des données d'une enquête par sondage complexe, l'auteur montre comment cette méthode d'estimation de l'erreur quadratique moyenne, bien que probablement trop incertaine pour être appliquée directement, peut servir à déterminer si l'estimateur pour petits domaines proposé ici est supérieur à l'estimateur classique fondé sur un plan.

MOTS CLÉS : Population finie; modèle; erreur quadratique moyenne; convergent selon le plan; randomisation.

1. INTRODUCTION

Supposons que nous avons un échantillon probabiliste de valeurs unitaires et que l'on nous demande d'estimer la moyenne d'un petit domaine inclus dans la population visée par l'échantillon. Scott et Smith (1969) ont défini à cette fin un estimateur bayesien et ont montré qu'il était possible de construire cet estimateur essentiellement à l'aide de deux critères : l'absence de biais et la variance minimum. Nous utiliserons ici cette approche, qui est appelée parfois modèle des effets aléatoires ou modèle des composantes de variance.

La plupart des auteurs qui se sont penchés sur l'estimation pour petit domaine, outre Scott et Smith, (notamment, Fay et Herriot 1979; Battese et Fuller 1971; Ghosh et Meeden 1986; Prasad et Rao 1986; Fuller et Harter 1987; et Stroud 1987) supposent que le plan de sondage est non informatif et donc «neutre». On pose la même hypothèse pour les estimateurs synthétiques de moyennes de petits domaines, dont il n'est aucunement question dans cet article (pour des exemples de ces estimateurs, voir Gonzalez et Hora, 1978).

Il faut supposer qu'un plan de sondage non informatif dissimule probablement l'apport le plus important de la randomisation à l'inférence statistique. Comme la plupart des modèles statistiques utilisés en inférence pour population finie sont soit erronés ou (au mieux) incomplets, il est souhaitable qu'une méthode d'estimation ait la propriété suivante: si l'échantillon est suffisamment grand, l'estimateur devrait approcher presque à coup sûr la valeur du paramètre qu'il vise à estimer, quel que soit le «véritable» modèle. Ce souhait trouve sa pleine expression dans le critère de convergence selon le plan, défini par Isaki et Fuller (1982).

La convergence selon le plan est une propriété asymptotique. On doit donc souvent définir par hypothèse un ou plusieurs modèles lorsque vient le temps de choisir parmi diverses méthodes d'estimation convergente selon le plan. Cela est particulièrement vrai dans le cas de l'estimation pour un petit domaine, où l'échantillon peut être particulièrement petit et le plan de sondage peut échapper à tout contrôle. Néanmoins, le fait de concentrer son attention sur des estimateurs convergents selon le plan offre une garantie minime mais réelle en ce qui a trait à la validité du modèle. C'est pourquoi Särndal (1984) s'est surtout intéressé aux estimateurs convergents selon le plan pour petit domaine et c'est ce que nous nous proposons de faire dans cet article.

[1] Phillip S. Kott, Senior Mathematical Statistician, Survey Research Branch, National Agricultural Statistics Service, USDA, S-4801, Washington, D.C., 20250, É.-U.

Dans la section 2, nous définissons un estimateur d'effets aléatoires convergent selon le plan pour la moyenne de population d'un petit domaine. Dans la section suivante, nous présentons un estimateur robuste (mais instable) pour les erreurs quadratiques moyennes de modèle et de plan de l'estimateur pour un petit domaine. L'estimateur est robuste en ce sens qu'il ne dépend pas du modèle fragile, mais nécessaire, qui relie les petits domaines entre eux. Enfin, nous présentons un exemple empirique dans la section 4 et une analyse dans la section 5.

2. L'ESTIMATEUR

Définissons tout d'abord le modèle *de base* (ou modèle des effets fixes):

$$y_{gi} = \theta_g + \epsilon_{gi}, \tag{1}$$

où les ϵ_{gi} sont des variables aléatoires non corrélées de moyenne nulle et de variance $\text{var}(\epsilon_{gi}) = \delta_g^2$. L'indice inférieur désigne une unité appartenant au domaine g. N_g unités de la population se trouvent dans le domaine g et il y a m domaines.

Prenons un domaine particulier j. Le problème consiste à estimer la moyenne de domaine:

$$\bar{y}_{jP} = \sum_{i=1}^{N_j} y_{ji}/N_j.$$

Soit p_{ji} la probabilité d'échantillonnage de l'unité ji et n_j le nombre d'unités prélevées dans le domaine j. Nous savons tous que si nous utilisons une méthode d'estimation linéaire non biaisée selon le plan et efficace selon le modèle pour $\bar{y}_{jP}$ p_{ji} sera égale à n_j/N_j et l'estimateur sera égal à $\sum_{i=1}^{n_j} y_{ji}/n_j$, où les unités sont identifiées à nouveau de manière que $j1, \ldots, jn_j$ se trouvent dans l'échantillon.

Malheureusement, il arrive souvent en pratique que l'on doive estimer une moyenne de domaine à l'aide d'un échantillon qui n'a pas été nécessairement constitué à cette fin. Par conséquent, il se peut que les probabilités d'échantillonnage dans le domaine j ne soient pas toutes égales à n_j/N_j. Un estimateur très utilisé dans les circonstances est

$$d_j = \sum_{i=1}^{n_j} w_{ji}\, y_{ji}, \tag{2}$$

où

$$w_{ji} = p_{ji}^{-1} / \sum_{k=1}^{n_j} p_{jk}^{-1},$$

désigne le *poids d'échantillonnage* de l'unité ji. Cet estimateur a été proposé par Brewer (1963) et Hajek (1971).

L'estimateur d_j est de toute évidence non biaisé selon le modèle (1), en ce sens que $E_\epsilon(d_j - \bar{y}_{jP}) = 0$. Il est aussi *convergent* selon de nombreux plans de sondage, c'est-à-dire

$$\underset{n_j \to \infty}{plim_\pi}(d_j - \bar{y}_{jP}) = 0,$$

où π désigne l'espace-probabilité issu du processus d'échantillonnage aléatoire plutôt que du modèle (1).

Isaki et Fuller (1982) exposent les conditions suffisantes pour que d_j soit convergent selon le plan et il l'est effectivement selon la plupart des plans de sondage d'usage courant, à l'exception notamment de l'échantillonnage systématique effectué à partir d'une liste pré-établie (voir Kott 1986). Au lieu de la convergence selon le plan, on parle souvent, pour un estimateur, de la propriété d'être *asymptotiquement non biaisé selon le plan* (Brewer 1979). L'estimateur d_j est toujours asymptotiquement non biaisé selon le plan

L'inconvénient de cet estimateur d_j est qu'il peut ne pas être très efficace lorsque n_j est faible. Pour éliminer cet inconvénient, on peut notamment «tirer avantage» des autres domaines en considérant le paramètre fixe θ_j comme la réalisation d'une variable aléatoire, qui satisfait le modèle *d'enchaînement*:

$$\theta_j = \mu + \tau_j, \tag{3}$$

où $E(\tau_j) = 0$, et $E(\tau_j \tau_g) = \sigma^2$ lorsque $j = g$ et 0 dans le cas contraire. C'est ce qu'on appelle parfois le «modèle des effets aléatoires» car l'effet j, θ_j, qui était jusque-là fixe, est maintenant considéré comme une variable aléatoire.

En combinant les équations (1) et (3), on obtient le modèle des composantes de variance en forme abrégée:

$$y_{ji} = \mu + \tau_j + \epsilon_{ji}, \tag{4}$$

De nombreux analystes utilisent dès le départ l'équation (4). Nous avons fait la distinction entre le modèle de base et le modèle d'enchaînement pour mieux illustrer le fait que les analystes prêtent souvent plus de crédibilité au modèle de base (surtout lorsqu'on suppose dans le modèle d'enchaînement que tous les $\delta_g^2 = \delta^2$, comme ce sera le cas bientôt).

N'importe quel estimateur de la forme:

$$f_j(\alpha,c) = (1 - a)d_j + \alpha\,\hat{\mu},$$

où

$$c = (c_1, \ldots, c_{j-1}, 0, c_{j+1}, \ldots, c_m),$$

$$\hat{\mu} = \sum_{g=1}^{m} c_g \bar{y}_{gS},$$

$$\bar{y}_{gS} = \sum_{i=1}^{n_g} y_{gi}/n_g,$$

et

$$\sum^{m} c_g = 1$$

est non biaisé selon le modèle (4). (Nota: bien que les variables c et $\hat{v}$ dépendent du domaine j, on a supprimé les indices correspondants pour plus de simplicité.)

Si nous supposons que tous les δ_g^2 sont égaux à δ^2, il est facile de montrer, par la méthode des multiplicateurs de Lagrange, que les valeurs de α^2, et de c_g qui minimisent la variance de modèle de $f_j(\alpha, c) - \bar{y}_{jP}$ sont

$$\alpha^* = \frac{\sum_{i=1}^{n_j} w_{ji}^2 - 1/N_j}{\sum_i w^2{}_{ji} + \sum_g c_g^{*2}/n_g + (1 + \sum_g c_g^{*2})\,(\sigma^2/\delta^2)}, \tag{5}$$

et

$$c_g^* = \frac{[(\sigma^2/\delta^2) + n_g^{-1}]^{-1}}{\sum_h [(\sigma^2/\delta^2) + n_h^{-1}]^{-1}}, \quad \text{pour } g \neq j. \tag{6}$$

Dans la pratique, σ^2 et δ^2 sont rarement connues. Ghosh et Meeden (1986) ont proposé d'estimer le rapport σ^2/δ^2 à partir de l'échantillon et d'une manière qui soit conforme au *modèle* (lorsque $m \to \infty$) par la formule

$$L = \max\left\{0, \left[\frac{\sum_g n_g(\bar{y}_{gS} - \bar{y}_S)^2/(m-1)}{\sum_g \sum_i (y_{gi} - \bar{y}_{gS})^2/(n-m)} - 1\right](m-1)/(n - \sum_g n_g^2/n)\right\} \tag{7}$$

où

$$\bar{y}_S = \sum n_g \bar{y}_{gS}/n$$

et

$$n = \sum n_g.$$

Posons $\alpha'(L)$ et $c'(L)$ comme les membres de droite des équations (5) et (6) respectivement, où L remplace σ^2/δ^2. Ensuite, désignons

$$e_j = f_j[a'(L), c'(L)]$$

comme *l'estimateur des effets aléatoires*, où $\hat{\mu}$ dans $e_j = f_j(.,.)$ équivaut à $\mu'(L) = \sum c_g'(L)\bar{y}_{gS}$. Lorsque m augmente, e_j se confond de plus en plus avec $f_j(a^*, c^*)$.

Si le modèle (4) est juste et que tous les $\delta_j^2 = \delta^2 > 0$, alors L doit être positif pour une valeur de m suffisamment grande. Même si le modèle n'est pas valide, pour autant que L a une borne inférieure positive $|\mu'(L)|$ est bornée et $n_j \sum_{i=1}^{n_j} w_{ji}^2$ est bornée lorsque n_j (*mais non m*) prend une valeur arbitrairement élevée, alors e_j est convergent selon le plan lorsque d_j l'est car

$$\underset{n_j \to \infty}{plim_\pi} [\alpha'(L)] = 0,$$

de sorte que e_j converge vers d_j qui est convergent selon le plan.

3. ERREUR QUADRATIQUE MOYENNE DE MODÈLE ET DE PLAN

Selon certains plans de sondage, l'estimateur de la variance de plan de d_j est aussi un estimateur non biaisé selon le modèle de la variance de d_j pris comme estimateur de $\bar{y}_{jP}$ selon le modèle de base (pour simplifier l'exposé, nous omettrons désormais les mots «pris comme estimateur de $\bar{y}_{jP}$»). Cependant, on doit souvent prendre en compte un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de d_j (à supposer, comme nous le ferons, qu'il existe une telle erreur). Cela est particulièrement vrai lorsque $\sum_{k=1}^{n_j} p_{jk}^{-1} \neq N_j$. Kott (1987) montre comment (lorsque cela s'impose) on peut redresser cet estimateur de manière à en faire simultanément un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de d_j et un estimateur non biaisé de la variance de d_j selon le modele *de base*. Désignons cet «estimateur de variance» redressé par $v(d_j)$.

Nous pouvons maintenant parler des erreurs quadratiques moyennes de modèle et de plan de l'estimateur des effets aléatoires, e_j. Bien qu'il fallait supposer que les δ_j^2 étaient tous égaux pour déterminer e_j, il n'est pas nécessaire de faire de même lorsqu'on évalue la précision de e_j. De fait, il n'est même pas nécessaire de supposer que le modèle d'enchaînement défini par l'équation (3) est valide! Il suffit de supposer que m est assez grand pour que L puisse être considéré comme (virtuellement) indépendant des unités du domaine de j. Par ailleurs, on peut redéfinir L en retirant les unités du domaine j des sommes qui figurent dans le membre de droite de l'équation (7).

D'une façon ou d'une autre, $E_\epsilon[(d_j - \bar{y}_{jP})(\bar{y}_{jP} - \mu'(L))] = 0$. En conséquence,

$$E_\epsilon[\{d_j - \mu'(L)\}^2] = \mathrm{var}_\epsilon(d_j - \bar{y}_{jP}) + E_\epsilon[\{\bar{y}_{jP} - \mu'(L)\}^2].$$

Il est maintenant facile de montrer que selon le modèle de base défini en (1),

$$v(e_j) = [1 - 2\alpha'(L)]\, v(d_j) + [\alpha'(L)]^2 [d_j - \mu'(L)]^2$$

est un estimateur non biaisé de l'erreur quadratique moyenne de modèle de e_j *étant donné* L et $\mu'(L)$. Puisque $\alpha'(L)$ est asymptotiquement nul lorsque n_j tend vers l'infini, $v(e_j)$ est aussi un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de e_j lorsque $v(d_j)$ est un estimateur convergent selon le plan de l'erreur quadratique moyenne de plan de d_j.

Il n'est pas nécessaire que L converge vers σ^2/δ^2 ou que $\mu'(L)$ converge vers μ pour que $v(e_j)$ possède les propriétés décrites ci-dessus. De fait, il n'est pas nécessaire de donner quelque interprétation que ce soit aux limites de L et $\mu'(L)$ puisque les propriétés en question ont été définies indépendamment du modèle exprimé par l'équation (3).

Les statisticiens ont souvent une confiance beaucoup plus grande dans le modèle de base (équation 1) que dans le modèle d'enchaînement (équation 3), surtout lorsque ce dernier s'accompagne de l'hypothèse que les variances (δ_g) sont les mêmes d'un domaine à l'autre. Il est donc rassurant de savoir que l'on peut estimer la précision de e_j sans recourir à l'équation (3) ou exiger que les δ_g soient égales.

Malheureusement, $\mathrm{v}(e_j)$ est instable et peut même être négatif lorsque $\alpha'(L)$ est supérieur à 0.5. Néanmoins, une simple comparaison des valeurs relatives de $\mathrm{v}(d_j)$ et de $\mathrm{v}(e_j)$ pour les m domaines ($j = 1, \ldots, m$) représente une méthode robuste pour choisir l'un ou l'autre des deux estimateurs, d_j et e_j.

4. EXEMPLE EMPIRIQUE

Dans le cadre de la Continuing Survey of Food Intakes by Individuals (CSFII), le Human Nutrition Information Service (HNIS) a recueilli des données sur la quantité de nourriture consommée quotidiennement par des femmes de 19 à 50 ans en 1985; pour cela, l'organisme a procédé à un échantillonnage à plusieurs degrés stratifié. Les données recueillies visaient 60 groupes d'aliments et 27 éléments nutritifs. Voir Human Nutrition Information Service (1985) pour plus de détails sur l'enquête et le plan de sondage.

Nous allons nous borner ici à estimer la quantité moyenne de lait et de produits laitiers (1 des 60 groupes d'aliments) consommés par des femmes âgées de 19 à 34 ans et de 35 à 50 ans dans douze domaines s'excluant mutuellement. Ces domaines sont définis en fonction de deux critères de classification: la région (Nord-Est, Middle-West, Sud et Ouest) et le niveau d'urbanisation (centre-ville, banlieue, secteur non métropolitain). En ce qui concerne la consommation moyenne par groupe d'aliments, HNIS a publié des données distinctes pour les deux groupes d'âge au niveau national seulement. Pour ce qui est de la consommation moyenne d'éléments nutritifs, des données ont été publiées pour chaque groupe d'âge selon la région et le niveau d'urbanisation mais non en fonction des deux critères à la fois.

Le plan de sondage de la CSFII prévoyait un échantillon à plusieurs degrés stratifié indépendant pour chacun des domaines. Tout d'abord, les unités primaires d'échantillonnage (villes ou municipalités) ont été choisies au moyen d'un échantillonnage *avec* remise avec probabilité proportionnelle à la taille; ensuite, nous avons prélevé un sous-échantillon aléatoire de segments aréolaires duquel a été tiré un sous-échantillon aléatoire de ménages, plus petit. Nous avons ensuite procédé à un autre sous-échantillonnage. En effet, lorsqu'un ménage de l'échantillon de la CSFII comptait plusieurs femmes du même groupe d'âge, nous en choisissions une aléatoirement.

Pour chaque groupe, d_j défini en (2) représente l'estimateur classique (fondé sur un plan) de la moyenne d'un domaine. Le programme SESUDAAN (Shah 1980) permet de calculer des estimateurs convergents selon le plan pour tous les d_j et leurs erreurs-types moyennes de plan ($\sqrt{\mathrm{EQM}(d_j)}$). Cependant, lorsque ces estimateurs sont élevés au carré, ils ne sont pas nécessairement des estimateurs non biaisés de la variance de d_j selon le modèle défini en (1).

Afin de vérifier cela, nous allons nous en tenir non seulement à un seul groupe d'âge mais aussi à un seul domaine et nous allons supprimer l'indice inférieur j. Désignons les strates par $h = 1, \ldots, H$ les unités primaires d'échantillonnage (U.P.É.) dans, $k = 1, \ldots, K_h$ et les femmes choisies dans hk par $i = 1, \ldots, n_{hk}$. L'estimateur pour la consommation moyenne est

$$d = \sum_{h=1}^{H} \sum_{k=1}^{K_h} \sum_{i=1}^{n_{hk}} w_{hki}\, y_{hki}.$$

Avant de poursuivre, nous devons définir de nouvelles relations. Soit

$$x_{hk} = \sum_{i=1}^{n_{hk}} w_{hki},$$

$$z_{hk} = \sum_{i=1}^{n_{hk}} w_{hki}^2,$$

$$f_{hk} = \sum_{i=1}^{n_{hk}} w_{hki}\,(y_{hki} - d),$$

et

$$f_h = \sum_{k=1}^{K_h} f_{hk}/K_h.$$

En supposant que la taille de la population du domaine est suffisamment élevés pour ne pas être prise en considération (cette hypothèse permettant aussi d'affirmer avec une assez grande certitude qu'aucune personne n'a été échantillonée deux fois), la variance de modèle de d est

$$\begin{aligned} \mathrm{var}_\epsilon(d) &= \delta^2 \sum_h \sum_k \sum_i w_{hki}^2 \\ &= \delta^2 \sum_h \sum_k z_{hk}. \end{aligned}$$

L'estimateur SESUDAAN (linéarisation) de l'erreur quadratique moyenne de plan de d est

$$\mathrm{v}^*(d) = \sum_{h=1}^{H} (K_h/[K_h - 1]) \sum_{k=1}^{K_h} (f_{hk} - f_h)^2.$$

Après de nombreuses transformations, il est possible de montrer que l'espérance de modèle de cet estimateur est

$$\begin{aligned} E_\epsilon[v^*(d)] = \delta^2 \Bigg[& \sum_h \sum_k z_{hk} \\ & - 2 \sum_h (K_h/[K_h - 1]) \left(\sum_k^{K_h} z_{hk}\, x_{hk} - \sum_k^{K_h} z_{hk} \sum_k^{K_h} x_h/K_h \right) \\ & + \left(\sum_h \sum_k z_{hk} \right) \sum_h (K_h/[K_h - 1]) \left(\sum_k^{K_h} x_{h\,k}^2 - \left\{ \sum_k^{K_h} x_{hk} \right\}^2 / K_h \right) \Bigg]. \end{aligned}$$

Tableau 1

Valeurs estimées pour les domaines, par groupe d'âge

Domaine	Taille de l'échantillon	dj	ej	v(dj)	v(ej)	$\alpha'(L)$
	Femmes 19-34 ans					
N - C	68	220.6	222.1	683.0	367.5	.233
N - B	95	195.7	203.1	568.8	367.8	.225
N - R	12	219.1	223.8	5266.7	-1349.5	.630
M - C	55	270.7	258.6	2021.5	1152.5	.251
M - B	107	277.2	267.8	625.8	509.6	.164
M - R	73	301.1	285.9	4027.1	2754.3	.187
S - C	66	212.4	215.7	3011.6	1700.1	.220
S - B	112	156.8	167.9	472.8	457.3	.146
S - R	81	117.0	139.3	592.0	868.9	.184
O - C	39	403.0	333.2	2064.2	5438.4	.364
O - B	74	205.0	209.6	1704.0	1018.3	.207
O - R	13	120.0	190.7	3533.5	3924.3	.652
	Femmes 35-50 ans					
N - C	44	205.3	197.4	1716.1	318.4	.425
N - B	67	135.0	153.1	1068.8	698.0	.326
N - R	21	206.1	195.4	579.2	56.6	.550
M - C	88	89.0	139.5	470.3	2559.9	.482
M - B	87	200.3	196.1	2128.5	1049.2	.258
M - R	38	304.9	250.7	6065.3	3973.9	.415
S - C	47	136.1	159.6	266.7	592.6	.421
S - B	93	161.0	167.7	1492.5	809.1	.244
S - R	77	128.8	146.3	1023.4	790.9	.263
O - C	23	205.5	193.9	7497.1	-1067.6	.580
O - B	88	245.1	229.1	2484.7	1432.2	.263
O - R	11	132.1	173.3	743.3	1344.1	.734

Domain Codes

N - Nord-Est; M - Middle-West; S - Sud; O - Ouest; C - centre-ville; B - banlieue; R - secteur non métropolitain

D'après Kott (1987),

$$\mathrm{v}(d) = \mathrm{v}^*(d)\ \mathrm{var}_\epsilon(d)/E_\epsilon[\mathrm{v}^*(d)]$$

est à la fois un estimateur convergent selon le plan de l'erreur quadratique moyenne de d (suivant certaines conditions) et un estimateur non biaisé selon le modèle de la variance de modèle de d.

Le tableau 1 donne les valeurs de n_j, d_j, $\alpha'(L)$, e_j, v(d_j) et v(e_j) qui ont été calculées pour les 12 domaines dans chacun des deux groupes d'âge (l'indice de domaine j figure de nouveau avec d_j et e_j). En utilisant l'équation (5), on obtient une valeur de L égale à 0.055 pour les femmes de 19 à 34 ans et à 0.037 pour les femmes de 35 à 50 ans. Ces résultats donnent à penser que les femmes d'un domaine ont peu de choses en commun si ce n'est qu'elles appartiennent au même groupe d'âge. Néanmoins $\alpha'(L)$ n'est supérieur à 0.5 que pour les cinq domaines (sur un total de 24) pour lesquels l'échantillon compte moins de 25 femmes.

La valeur estimée v(e_j) est négative dans deux cas sur 24 et inférieure à v(d_j) dans 18 cas sur 24, soit dans neuf cas pour chaque groupe d'âge. Ces derniers chiffres nous permettent

de croire que les e_j sont effectivement de meilleures estimations que les d. De façon formelle, si nous considérons les 24 écarts d_j, comme indépendants d'un domaine à l'autre (ce qui n'est pas tout à fait le cas), l'hypothèse que les erreurs quadratiques moyennes réelles de modèle (ou de plan) de $v(e_j) - v(d_j)$, sont égales et que la variable aléatoire $v(e_j) - v(d_j)$ est aussi susceptible de prendre une valeur positive que négative ne tient plus du tout.

Si nous substituons e_j à d_j l'erreur quadratique moyenne devrait diminuer d'environ 40.6 % (selon la formule $\sum \{v(e_j) - v(d_j)\} / \sum v(d_j)$). L'erreur-type devrait pas conséquent diminuer de 22.9 %. Comme nous additionnons 24 variables aléatoires quasi indépendantes, nous avons beaucoup plus confiance dans cette estimation que dans n'importe quelle estimation particulière $v(e_j)$ (ou $v(d_j)$ dans les circonstances).

5. ANALYSE

Soit $n_j^* = 1/\sum_{i=1}^{n_j} w_{ji}^2$ la *taille effective de l'échantillon* dans le domaine j. Soulignons que $n_j^* \leq n_j$ l'égalité étant vérifiée si et seulement si tous les poids d'échantillonnage dans j égalent $1/n_j$. Pour un rapport σ^2/δ^2, connu, la seule chose qui distingue l'estimateur optimal défini ici $f_j(a^*,c^*)$, du meilleur estimateur linéaire sans biais de Scott et Smith (1969) est que $1/n_j^*$ remplace $1/n_j$ dans la formule qui sert à calculer α^* (équation (5)). Cette substitution a pour effet d'accroître w_{ji} lorsque les poids j ne sont pas tous égaux; en d'autres termes, cette substitution crée une dépendance plus forte par rapport aux données d'échantillon qui ne sont pas du domaine j. Cela est inévitable car lorsqu'on s'efforce d'obtenir un estimateur convergent selon le plan, on n'utilise pas l'échantillon du domaine j de la façon la plus efficace. Nous pourrions toujours pondérer cet échantillon d'une manière conforme à partir d'autres domaines en nous servant de poids d'échantillon pour calculer $\mu'(L)$, mais cela ne ferait que réduire l'efficacité de modèle de l'estimateur sans améliorer aucune caractéristique axée sur le plan.

Par l'équation (7), nous sommes sûrs que L ne peut être inférieur à zéro. Cela implique que $\alpha'(L)$ ne peut être plus grand que $\sum_{g=j} n_g / (\sum_{g=j} n_g + n_j^*)$. Si $\alpha'(L)$ était égal à sa borne supérieure et que $n_j^* = n_j$, e_j se réduirait alors à la moyenne simple des y_{gi} pour tout l'échantillon. Cela a du sens car lorsque le modèle défini en (4) est exact et que $\sigma^2 = 0$, l'estimateur le plus efficace de $\mu + \gamma_j = \mu$ est la moyenne de l'échantillon global.

Par contre,si $n_j^* < n_j$ et $L = 0$, on calculera e_j en pondérant plus fortement les unités qui n'appartiennent pas au domaine j que celles qui y appartiennent, ce qui est peu rationnel. Une façon ponctuelle de contourner le problème serait de définir une borne supérieure de $1 - (n_j/\sum n_g)$ (ou moins) pour $\alpha'(L)$. Une autre solution serait de renoncer à l'estimation pour un petit domaine lorsque la valeur de $\alpha'(L)$ calculée selon la formule proposée plus haut, dépasse $1 - (n_j/\sum n_g)$. Pour que cela se réalise, il faudrait que L, la valeur estimée pour σ^2/δ^2, soit très faible. Dans l'exemple empirique de la section précédente, la valeur de L se situait entre 0.03 et 0.06 et pourtant $\alpha'(L)$ était encore bien en-deçà de $1 - (n_j/\sum n_g)$.

Le modèle complet défini par l'équation (4) peut s'avérer inexact de deux façons: ou bien le modèle des effets fixes dans chaque domaine (équation (1)) est erroné, ou bien le modèle d'enchaînement (équation (3)) l'est. Dans la réalité, les deux modèles sont susceptibles d'être erronés. Le modèle des effets fixes ne tient pas compte des effets de stratification ou de grappe ni des effets, aussi ténus soient-ils, de la présence de plusieurs femmes du même groupe d'âge dans un ménage. Dans l'un ou l'autre cas, ces effets ne devraient pas être notables. De plus, en introduisant des poids d'échantillonnage dans l'estimateur d_j et en faisant en sorte que les estimateurs de l'erreur quadratique moyenne soient convergents selon le plan, nous avons fait tout ce que nous pouvions pour parer au rejet du modèle des effets fixes.

Par ailleurs, le modèle d'enchaînement ne devrait pas nous inspirer une grande confiance. Ce modèle n'est guère plus qu'un artifice statistique qui, notamment, ne tient pas compte de la corrélation qui pourrait exister entre les quantités d'aliments consommés par des femmes qui vivent dans la même région mais dans des secteurs qui n'ont pas le même niveau d'urbanisation ou *vice versa*.

Comme nous l'avons dit plus haut, une façon de choisir l'un ou l'autre des deux estimateurs $v(e_j) - v(d_j)$ qui est indépendante du modèle d'enchaînement, est de dénombrer simplement les fois que δ_j et e_j est négatif. Cependant, l'estimateur $v(e_j)$ est instable et ne devrait pas servir en pratique à estimer l'erreur quadratique moyenne.

Tandis que les estimations de l'erreur quadratique moyenne de e_j ont instables, les $v(d_j)$ sont à peine meilleures. Au mieux, le nombre de «degrés de liberté» se rattachant à $v(d_j)$ est égal à la différence entre le nombre d'UPÉ et le nombre de strates dans j. En ce qui concerne l'échantillon de la CSFII, le nombre de degrés de liberté varie de 2 à 7.

Comme les statisticiens doivent de plus en plus indiquer les erreurs types estimées à côté des moyennes estimées qu'ils publient, il est urgent de trouver des estimateurs plus stables que $v(d_j)$ et $v(e_j)$. On pourrait, par exemple, ajuster $v(d_j)$ et $v(e_j)$, au moyen d'une fonction d'estimation de la variance, cet ajustement pouvant s'appliquer à l'un ou à l'autre estimateur ou aux deux à la fois. Toutefois, il s'agit là d'une méthode *ponctuelle* qui ne peut guère plus que donner des valeurs proches des valeurs estimées (entièrement dépendantes du modèle) des erreurs quadratiques moyennes de d_j et de e_j (voir Prasad et Rao 1986, pour une analyse probante de la question) en faisant la moyenne des effets du rejet du modèle.

Une solution intéressante serait de combiner les estimateurs (stables mais biaisés) de l'erreur quadratique moyenne fondés sur le modèle avec les estimateurs convergents selon le plan qui ont été définis dans cet article, un peu comme on le fait pour les moyennes avec e_j. Cependant, d'autres recherches s'imposent pour que nous sachions comment appliquer cette solution.

REMERCIEMENTS

L'auteur tient à remercier le Human Nutrition Information Service, qui a permis l'accès à sa base de données, ainsi que Joe Goldman, qui a aidé à constituer les séries de données qui ont servi à l'analyse empirique. Des remerciements sont également adressés à John Herbert et à deux arbitres anonymes pour leurs précieux commentaires sur des versions préliminaires de cet article.

BIBLIOGRAPHIE

BATTESE, G.E., et FULLER, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 500-505.

BREWER, K. R. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* 5, 93-105.

BREWER, K. R. (1979). A class of robust sampling designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

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

FULLER, W. A., et HARTER, R. M. (1987). The multivariate components of variance model for small area estimation. Dans *Small Area Statistics*, Éds. R. Platek, J. N. K. Rao, C. E. Särndal et M. P. Singh. New York: John Wiley and Sons.

GHOSH, M., et MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, 81, 1058-1069.

GONZALEZ, M.E., et HORA, C. (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, 7-15.

HAJEK, J. (1971). Comment. Dans *Foundations of Statistical Inference*, Éds. V. P. Godambe et D. A. Sprott. Toronto: Holt, Rinehart, et Winston.

HUMAN NUTRITION INFORMATION SERVICE (1985). *CSFII - Nationwide Food Consumption Survey Continuous Survey of Food Intake by Individuals: Women 19-50 Years and Their Children 1-5 Years, 1 Day*. NFCS, CSFII Report No. 85-1. Washington: United States Department of Agriculture.

ISAKI, C. T., et FULLER, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOTT, P. S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika*, 73, 485-491.

KOTT, P. S. (1987). Estimating the conditional variance of a design consistent regression estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 486-491.

PRASAD, N. G. N., et RAO, J. N. K. (1986). On the estimation of mean square error of small area predictions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 108-116.

SÄRNDAL, C. E. (1984). Design consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association*, 79, 624-631.

SCOTT, A., et SMITH, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association*, 64, 830-840.

SHAH, B. V. (1981). *SESUDAAN: Standard Error Program for Computing of Standardized Rates from Sample Survey Data*. Research Triangle Park: Research Triangle Institute.

STROUD, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. Dans *Small Area Statistics*, Éds. R. Platek, J. N. K. Rao, C. E. Särndal et M. P. Singh. New York: John Wiley and Sons.