Techniques d'enquête, décembre 1993
Vol. 19, n° 2, pp. 183–192
Statistique Canada

Plans de sondage à deux degrés optimaux pour des estimateurs de ratios: application au contrôle de qualité du recensement français de 1990

JEAN-CLAUDE DEVILLE[1]

RÉSUMÉ

Cette étude est basée sur l'utilisation de modèles de superpopulation pour anticiper la variance d'une mesure par sondage de ratios *avant* l'enquête. On arrive, en utilisant des modèles simples qu'on voudrait néanmoins assez réalistes, à des expressions plus ou moins complexes qu'on parvient à optimiser, parfois rigoureusement, quelquefois de façon approximative. La solution du dernier des problèmes évoqués fait apparaître un facteur assez peu étudié en matière d'optimisation de plan de sondage: le coût lié à la mobilisation d'une information individuelle.

MOTS CLÉS: Contrôle de qualité du recensement; modèle de superpopulation; optimisation d'un plan à deux degrés; enquête à objectifs multiples.

1. INTRODUCTION

Le contrôle par sondage de la qualité des données du recensement Français a posé quelques problèmes à la fois intéressants et nouveaux. Trois d'entre eux sont traités dans cet article. Nous les étudierons en termes généraux et nous décrirons ensuite leur application au cas du recensement.

Dans tous les cas, le problème est celui de l'optimisation d'un sondage à deux degrés où les unités primaires sont des districts de collecte du recensement. Ceux-ci sont repérés par un indice k variant dans une population U de districts, qui est concrétement une unité de traitement des bulletins collectés.

Le premier problème consiste à estimer la fréquence d'un caractère dans la population de bulletins (le fait de comporter une erreur). Désirant avoir une précision donnée pour cette estimation on cherche à minimiser le coût du sondage avec une fonction de coût de la forme:

$$C_T = mC_o + nC_1, \tag{1.1}$$

où m est le nombre d'unités primaires (districts) échantillonnées, C_o le coût de traitement d'une U.P., n le nombre d'unités finales (bulletins) échantillonnées, et C_1 le coût de traitement d'une unité finale. Le problème est assez habituel dans le cas de l'estimation d'une moyenne (voir par exemple W. Cochran (1977)). Il reçoit ici une solution plus complète tenant compte de la grande variabilité de taille des unités primaires.

Le second problème est plus original et possède une portée plus générale. La population finale (ici les bulletins) est composée de G groupes ($g = 1$ à G) distincts. On désire avoir une estimation de la fréquence du caractère dans chacun des groupes, avec une précision donnée pour chacun d'eux. La contrainte réside dans le fait que les unités primaires seront communes à tous les groupes, l'échantillonnage au sein de chaque U.P. portant alors sur chaque groupe.

L'objectif est alors de minimiser le coût du sondage celui-ci ayant la forme:

$$C_T = mC_o + \sum_{g=1}^{G} n_g C_g, \tag{1.2}$$

où n_g est le nombre total d'unités finales du groupe g et C_g le coût de traitement d'une unité finale du groupe g. En pratique les groupes sont constitués par les différents types de bulletins utilisés dans le recensement.

Le troisième problème a sa source dans le contrôle de la codification. On possède, *a priori*, une mesure de la difficulté de codification de chaque bulletin. Formellement donc, on dispose au niveau de chaque individu i de la population, d'une variable quantitative X_i, telle que la probabilité - en un sens à préciser - pour que l'individu possède le caractère à mesurer soit à peu près proportionnelle à X_i. On cherche à utiliser cette information pour minimiser le coût du contrôle (mesure de la fréquence du caractère "codification erronnée") sous la requête d'une précision donnée du sondage.

Dans chacun des cas, on utilise des modèles de superpopulation plausibles et simples qui permettent d'évaluer la variance anticipée du sondage. On a, en quelque sorte, une illustration presque typique d'un "échantillonnage assisté par un modèle" dans l'esprit du livre de Särndal, Swensson, Wretman (1992).

[1] Jean-Claude Deville, Chef de la Division des Méthodes Statistiques et Sondages, Institut National de la Statistique et des Études Économiques, 18, boul. Adolphe Pinard, 75675 Paris, CEDEX 14.

2. ESTIMATION OPTIMALE DE LA PROPORTION D'ENREGISTREMENTS ERRONNÉS À L'AIDE D'UN PLAN DE SONDAGE À DEUX DEGRÉS

Les unités primaires k (districts) comportent chacune un nombre connu N_k d'individus (des bulletins). Parmi ceux-ci D_k possèdent le caractère (être erronné). Le but est donc d'estimer:

$$P = \sum_U D_k \Big/ \sum_U N_k .$$

Le sondage consistera à tirer un échantillon s d'unités primaires (U.P.) avec des probabilités d'inclusion π_k au premier ordre et $\pi_{k\ell}$ au second ordre à déterminer. Ensuite, si l'unité primaire k est tirée dans s on vérifiera n_k individus tirés par sondage aléatoire simple sans remise. Soit d_k le nombre de bulletins erronnés qu'on y relèvera.

L'estimateur $\hat{P}_k$ de $P_k = D_k/N_k$ sera $\hat{P}_k = d_k/n_k$ et $\hat{D}_k = N_k\hat{P}_k$ estimera D_k sans biais. L'estimateur de P sera:

$$\hat{P} = \frac{\sum_s \frac{\hat{D}_k}{\pi_k}}{\sum_s \frac{\hat{N}_k}{\pi_k}}. \tag{2.1}$$

C'est le ratio des estimateurs sans biais de D et de N, le nombre total de bulletins. Bien que ce nombre soit connu, il est bien évident que l'estimateur (4.1) est plus précis que $1/N \sum_s \hat{D}_k/\pi_k$.

On a:

$$\operatorname{Var}(\hat{P}) = E \operatorname{Var}(\hat{P} \mid s) + \operatorname{Var} E(\hat{P} \mid s). \tag{2.2}$$

Or:

$$\operatorname{Var}(\hat{P} \mid s) = \hat{N}^{-2} \sum_s \frac{N_k^2}{\pi_k^2} \frac{P_k(1-P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right)$$

avec $$\hat{N} = \sum_s \frac{N_k}{\pi_k}.$$

D'où:

$$E \operatorname{Var}(\hat{P} \mid s) \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k} \frac{P_k(1-P_k)N_k}{N_k - 1}\left(\frac{1}{n_k} - \frac{1}{N_k}\right). \tag{2.3}$$

Par ailleurs,

$$E(\hat{P} \mid s) = \frac{\sum_s \frac{D_k}{\pi_k}}{\sum_s \frac{N_k}{\pi_k}}.$$

La variance de cette quantité s'obtient par linéarisation en introduisant la variable $Z_k = D_k - PN_k = N_k(P_k - P)$.

On obtient:

$$\operatorname{Var} E(\hat{P} \mid s) \simeq N^{-2} \operatorname{Var}\left(\sum_s \frac{z_k}{\pi_k}\right).$$

Soit, compte tenu de ce que $\sum_U Z_k = 0$:

$$\operatorname{Var} E(\hat{P} \mid s) = N^{-2}\left(\sum_k \frac{Z_k^2}{\pi_k} + \sum\sum_{k \neq l} \frac{Z_k Z_\ell}{\pi_k \pi_\ell}\pi_{k\ell}\right). \tag{2.4}$$

La somme des quantités (2.3) et (2.4) nous donne la variance de l'estimateur (2.1).

2.1 Introduction d'un modèle

La variance de $\hat{P}$ est difficile à manipuler et, de plus, contient des paramètres inconnus. On se tire de la difficulté en faisant de nécessaires hypothèses qui se traduisent par un modèle de superpopulation. On supposera plus loin que les paramètres de ce modèle sont susceptibles d'être estimés à partir d'un essai préliminaire portant sur une toute petite partie de la population. On note E_ξ l'espérance sous le modèle (resp Var_ξ pour la variance) dont tous les aléas sont supposés indépendants du processus d'échantillonnage.

Le modèle suit les spécifications suivantes:

a) D_k suit une loi binomiale (N_k, p_k). P_k est donc, *sous le modèle*, un estimateur de p_k.

b) p_k est lui même aléatoire. On suppose les p_k indépendantes et de même loi avec:

$$E_\xi p_k = P,$$

$$\operatorname{Var}_\xi p_k = \sigma^2$$

pour tout k, quelle que soit, en particulier, la valeur de N_k.

En conditionnant, dans le modèle, par les p_k on a évidemment:

$$E_\xi(D_k \mid p_k) = N_k p_k,$$

$$\operatorname{Var}_\xi(D_k \mid p_k) = N_k p_k(1 - p_k).$$

La *variance anticipée* de $\hat{P}$ est la quantité $E_\xi \operatorname{Var}\hat{P}$. C'est à elle que nous allons nous intéresser désormais. Pour l'évaluer on remarque que:

a) $E_\xi(P_k - P)^2 = E_\xi(E_\xi(P_k - p_k + p_k - P)^2 \mid p_k)$

$$= \frac{P(1-P) - \sigma^2}{N_k} + \sigma^2,$$

b) $E_\xi P_k(1 - P_k) = E_\xi(E_\xi((P_k - P_k^2) \mid p_k))$

$$= E_\xi p_k(1 - p_k)\frac{N_k - 1}{N_k}$$

$$= (P(1 - P) - \sigma^2)\frac{N_k - 1}{N_k},$$

c) $E_\xi Z_k Z_\ell = 0$ à cause de l'indépendance des Z_k et des Z_ℓ, ce qui nous débarrasse d'un terme bien encombrant en même temps que des $\pi_{k\ell}$.

En récollant tous les morceaux de (2.3) et (2.4) un petit miracle algébrique se produit et nous avons l'expression:

$$E_\xi \mathrm{Var}\,\hat{P} \simeq N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \quad \text{avec } \tau^2 = P(1 - P) - \sigma^2 \text{ (quantité positive par nature)} \qquad (2.1.1)$$

Remarque:

Le miracle algébrique s'explique bien si on ne cherche pas à obtenir la variance sous le plan de sondage uniquement. Elle est d'ailleurs la conséquence d'un modèle un peu plus général que celui que nous avons posé.

Supposons que nous voulions estimer le total $N\bar{Y} = \sum_U Y_i$ d'une variable Y et que pour cela nous réalisions un tirage à deux degrés: un premier degré où des unités primaires k sont tirées avec des probabilités π_k, un second où n_k unités finales sont tirées par sondage aléatoire simple.

Nous posons un modèle où:

$$Y_i = \bar{Y} + \alpha_k + \epsilon_i,$$

avec α_k variable liée à l'U.P. d'indice k. Les α_k sont indépendantes de même loi d'espérance nulle de variance σ^2. Les ϵ_i sont également indépendantes centrées de variance égale à τ^2. Avec $\pi_i^* = \pi_k n_k / N_k$ (N_k taille de l'U.P. numéro k), l'estimateur de Horvitz-Thompson du *total* vaut $\hat{Y} = \sum Y_i/\pi_i^*$ la somme étant étendue à l'échantillon. Sous le modèle, et conditionnellement à l'échantillon on a:

$$\mathrm{Var}_\xi(\hat{Y} \mid s) = \sum_s \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right).$$

L'espérance sous le plan de cette expression redonne la formule (2.1.1).

2.2 Recherche d'un plan de sondage optimal

La variance maximum de $\hat{P}$ est fixée par les critères retenus pour le contrôle de qualité. Le sondage étant répété pour chacune des unités de traitement il est tout à fait naturel de chercher à minimiser l'espérance du coût du sondage donné en (2.1.1) soit:

$$E \sum_s (C_o + n_k C_1) = \sum_U \pi_k (C_o + n_k C_1). \qquad (2.2.1)$$

Le problème d'optimisation s'écrit donc:

$$\text{Minimiser} \sum_U \pi_k (C_o + n_k C_1)$$

sous les contraintes:

$$N^{-2} \sum_U \frac{N_k^2}{\pi_k}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \leq V_o$$

et pour tout k, $n_k \leq N_k$.

Associons un multiplicateur de Lagrange λ à la première contrainte - qui sera évidemment saturée - et des multiplicateurs μ_k aux autres. On obtient les solutions:

$$C_o + n_k C_1 = \lambda \frac{N_k^2}{\pi_k^2}\left(\sigma^2 + \frac{\tau^2}{n_k}\right) \qquad (2.2.2)$$

et, pour tout k:

$$C_1 \pi_k = \lambda \frac{N_k^2}{\pi_k} \cdot \frac{\tau^2}{n_k^2} + \mu_k \qquad (2.2.3)$$

avec

$$\mu_k = 0 \quad \text{si} \quad n_k < N_k \quad \text{et} \quad \mu_k > 0 \quad \text{si} \quad n_k = N_k.$$

Pour l'utilisation des multiplicateur de Lagrange, voir par exemple Luenberger (1973).

Pour toutes les unités primaires où $\mu_k = 0$ (les plus grosses) on obtient:

$$n_k = \frac{\tau}{\sigma}\left(\frac{C_o}{C_1}\right)^{1/2} = n^*. \qquad (2.2.4)$$

Chaque unité primaire reçoit donc la même allocation, ce qui correspond à l'idée qu'on a besoin de la même précision dans chacune d'elle. Retournons à l'équation (2.2.3). On constate alors que, toujours pour ces unités primaires, les probabilités d'inclusion π_k doivent être proportionnelles aux tailles N_k soit:

$$\pi_k = \lambda^{1/2} C_1^{-1/2} \frac{\tau}{n^*} N_k. \qquad (2.2.5)$$

C'est la justification habituelle d'un sondage autopondéré avec un premier degré tiré avec des probabilités proportionnelles à une mesure de taille (Voir par exemple Cochran 1977).

Comme n_k ne dépend pas de N_k on ne pourra avoir $n_k = N_k$ et $\mu_k > 0$ que si $N_k \leq n^*$. L'équation (2.2.2) nous permet alors d'obtenir les probabilités d'inclusion à un facteur près:

$$\pi_k = \lambda^{1/2} N_k \left(\frac{\sigma^2 + \tau^2/N_k}{C_o + C_1 N_k}\right)^{1/2} = \lambda^{1/2} N_k^{1/2} \left(\frac{N_k\sigma^2 + \tau^2}{N_k C_1 + C_o}\right)^{1/2}. \tag{2.2.6}$$

Les relations (2.2.5), valide si $N_k \geq n^*$ et (2.2.6) valide si $N_k \leq n^*$ établissent que π_k est proportionnelle à une variable connue $T_k = f(N_k)$ dont le graphique est donné à la figure 1.

Pour spécifier entièrement le sondage il reste à trouver le nombre m d'unités primaires à tirer. Or, $T = \sum_U T_k$ est aussi une quantité connue.

En se restreignant à un échantillonnage de taille fixe on aura donc $\pi_k = m\, T_k/T$. On trouve m en portant cette valeur dans la contrainte de variance soit:

$$N^2 V_o m = T \sum_U \frac{N_k^2}{T_k}(\sigma^2 + \tau^2/n_k).$$

Si, en première approximation, on prend $T_k = N_k$, on obtient la formule simplifiée:

$$m V_o = \sigma^2 + \tau^2/n^*.$$

On a ainsi obtenu une solution complète au problème.

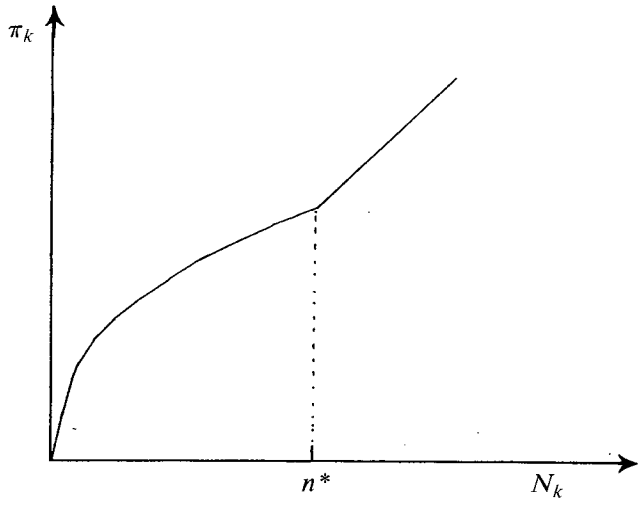

Figure 1. Graphique de π_k en fonction de N_k

3. ESTIMATION OPTIMALE DANS LE CAS D'UN SONDAGE À DEUX DEGRÉS OU LES UNITÉS PRIMAIRES SONT STRATIFIÉES

La dure réalité des choses nous amène à compliquer un peu le problème car on doit, en fait, contrôler indépendamment plusieurs types de bulletins. Ceci amène à poser un problème assez général qui est le suivant:

Pour chaque unité primaire (ici les districts d'une unité de traitement) on connaît les effectifs N_{kg}, d'unités secondaires appartenant à G groupes. La "population" de l'UP numéro k vaut $N_{k+} = \sum_g N_{kg}$; celle du groupe g vaut $N_{+g} = \sum_k N_{kg}$. Comme dans ce qui précède on cherche avec quelle probabilité d'inclusion π_k échantillonner l'UP numéro k, le nombre d'UP à tirer et l'allocation n_{kg} de l'échantillon parmi les différents groupes dans l'UP k, sachant que ces n_{kg} unités sont tirées par un sondage aléatoire simple parmi les N_{kg} unités tirables.

3.1 Recherche d'un plan optimal à l'aide d'un modèle

On postule, dans chacun des groupes, un modèle identique à celui formulé à la section (2.1) (ou sous une forme plus générale dans la remarque qui la termine).

Pour $g = 1$ à G on aura donc:

$$v_g = E_\xi \operatorname{Var}(\hat{P}_g) = N_{+g}^{-2} \sum_U \frac{N_{kg}^2}{\pi_k}(\sigma_g^2 + \tau_g^2/n_{kg}). \tag{3.1.1}$$

La fonction de côut est donnée par la forme générale (1.2). On va chercher à minimiser l'espérance du coût de sondage:

$$C_T = \sum_U \pi_k \left(C_o + \sum_g n_{kg} C_g\right), \tag{3.1.2}$$

sous les contraintes $V_g \leq \mathcal{V}_g$, où les quantités $\mathcal{V}_g$, sont fixées de façon extérieure, par exemple par la qualité des données qu'on veut obtenir et la rigueur du contrôle.

Sous cette forme, le problème peut s'avérer assez complexe. Nous allons écrire un "Lagrangien" général:

$$L = \lambda C_T + \sum_g \lambda_g V_g.$$

Le problème posé fixe $\lambda = 1$ et les λ_g sont des multiplicateurs à déterminer. Une variante simple consiste à fixer les λ_g: on désire alors minimiser une combinaison linéaire donnée des variances sous une contrainte de coût. Dans toutes les hypothèses, on obtient par dérivation par rapport aux n_{kg} (considérées comme des variables réelles):

$$\lambda \pi_k^2 C_g = \lambda_g N_{+g}^{-2} N_{kg}^2 \tau_g^2/n_{kg}^2. \tag{3.1.3}$$

Les π_k étant, pour l'instant, destinées à être connues à un facteur près, on peut écrire:

$$\pi_k \, n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g \frac{N_{kg}}{N_{+g}}. \qquad (3.1.4)$$

Par sommation sur k on en déduit que:

$$E\, n_{+g} = \sum_U \pi_k \, n_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g. \qquad (3.1.5)$$

La taille totale de l'échantillon dans chaque groupe est donc directement liée au multiplicateur λ_g.

La dérivation du Lagrangien par rapport aux π_k nous donne de nouvelles relations qui se simplifient miraculeusement si on utilise aussi (3.1.4). On obtient:

$$C_o = \sum_g C_g \left(\frac{\sigma_g}{\tau_g}\right)^2 n_{kg}^2, \qquad (3.1.6)$$

où encore, si on introduit les nombres

$$n_g^* = \left(\frac{C_o}{C_g}\right)^{1/2} \frac{\tau_g}{\sigma_g},$$

on écrit:

$$\sum_g \left(\frac{n_{kg}}{n_g^*}\right)^2 = 1. \qquad (3.1.7)$$

Comme on s'en rend compte en jetant un oeil à la formule (2.2.4), les n_g^* sont les nombres d'unités secondaires à tirer par UP s'il n'y a qu'un seul groupe; n_{kg} sera toujours inférieur à n_g^*.

De (2.1.4), (3.1.5) et (3.1.7) on tire les relations:

$$\pi_k^2 = \frac{1}{C_o} \sum_g \lambda_g \sigma_g^2 \left(\frac{N_{kg}}{N_{+g}}\right)^2. \qquad (3.1.8)$$

Ainsi, les π_k sont proportionnelles aux quantités T_k telles que $T_k^2 = \sum_g \lambda_g \sigma_g^2 N_{kg}^2 / N_{+g}^2$ qui apparaissent comme la mesure de taille adéquate. Les relations (3.1.4) montrent que, à k fixé, les n_{kg} sont proportionnelles à $n_g^* \lambda_g^{1/2} \sigma_g N_{kg}/N_{+g}$, ce qui compte tenu de (3.1.7) conduit à:

$$n_{kg} = n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}} T_k^{-1}. \qquad (3.1.9)$$

3.2 Solutions explicites dans deux cas particuliers

a) Si les λ_g étaient connus, c'est-à-dire si on minimisait $\sum_g \lambda_g v_g$ sous une contrainte de coût, alors (3.1.2) et (3.1.9) nous permettraient de calculer les T_k. En reportant:

$$\pi_k = m\, T_k / T \left(T = \sum_U T_k,\ m \quad \begin{matrix}\text{nombre d'unités}\\ \text{primaires à tirer}\end{matrix}\right)$$

dans la contrainte de budget $C_T \leq C_T^*$, on trouve:

$$C_T^* = \frac{m}{T} \sum_U \left(C_o T_k + \sum_g C_g n_g^* \lambda_g^{1/2} \sigma_g \frac{N_{kg}}{N_{+g}}\right)$$

soit:

$$m = C_T^* \Big/ \left(C_o + \sum_g C_g n_g^* \cdot \frac{\lambda_g^{1/2} \sigma_g}{T}\right).$$

Si un seul des λ_g est différent de zéro, on peut vérifier assez facilement qu'on retrouve le résultat donné à la fin de la section (2.2).

b) Le problème initial (min C_T sous $V_g \leq \mathcal{V}_g$) se résoud assez facilement dans deux cas particuliers:

b1 - *Dispersion maximale* des groupes. Pour toute UP k, on a $N_{kg} = N_{k+}$ pour un certain k. Le problème est décomposé en G problèmes distincts, chacun d'eux étant du type étudié dans la section 2.

b2 - *Dispersion minimale*: La répartition est la même dans toutes les UP; autrement dit on a pour tout k et g

$$N_{kg} = N_{k+} \frac{N_{+g}}{N} \quad \text{avec} \quad \left(N = \sum_g N_{+g}\right),$$

T_k est alors proportionnelle à N_{k+}, et les n_{kg} sont des quantités $n_g^* u_g$ indépendantes de k.

Avec $\pi_k = mN_{k+}/N$, on obtient en écrivant $V_g = \mathcal{V}_g$:

$$m\mathcal{V}_g = \sigma_g^2 + \tau_g^2 / n_g^* u_g$$

soit:

$$m = \frac{\sigma_g^2}{\mathcal{V}_g} + u_g^{-1} \frac{\tau_g^2}{n_g^* \mathcal{V}_g}.$$

On obtient ainsi G-1 relations linéaires entre les u_g^{-1} ce qui permet, en principe, de résoudre complètement le problème sachant que la somme des u_g^2 vaut 1.

3.3 Un algorithme numérique permettant de trouver la solution optimale dans le cas général

Une résolution numérique itérative du problème peut se faire de la façon suivante:

Étape 1: On fixe une allocation approximative de l'échantillon dans chaque groupe, soit n_{+g} unités dans le groupe g. Pour y arriver on peut, par exemple, se servir de la solution approximative avec les hypothèses du point a) ou du point b).

Étape 2: La valeur des λ_g est déterminée par les relations (3.1.5):

$$\lambda_g = C_g n_{+g}^2 / \tau_g^2.$$

Étape 3: Les π_k sont déterminées par les relations (3.1.8). La somme des π_k fixe, en particulier, le nombre d'UP à tirer.

Étape 4: Les n_{kg} sont déterminés par les relations (3.1.4). On peut ensuite itérer par retour à l'étape 2 en espérant que cet algorithme converge vers la solution d'optimisation.

Remarque: La probabilité de tirer une unité de type g vaut

$$\pi_k n_{kg} \Big/ N_{kg} = \left(\frac{\lambda_g}{C_g}\right)^{1/2} \tau_g / N_{+g}.$$

Elle ne dépend pas de l'Unité Primaire k et est donc la même pour chaque unité d'un groupe g donné (sondage à probabilités égales). On en déduit la taille n_{+g} de l'échantillon dans le groupe g, ou du moins, son espérance mathématique. Pratiquement, il arrive que l'on fixe "autoritairement" les tailles des échantillons. Ceci revient à déterminer les λ_g, ou, implicitement, des variances ∇_g. Ce résultat est assez naturel lui aussi.

4. ESTIMATION OPTIMALE À L'AIDE D'UNE MESURE DE LA DIFFICULTÉ À CODER UN ENREGISTREMENT

Il s'agit d'estimer la proportion de bulletins présentant une erreur de codification dans l'"univers" U de tous les bulletins codifiés une semaine donnée dans une Direction Régionale. Le caractère particulier du problème est le suivant: tous les bulletins i sont déjà précodifiés ce qui permet, grâce à des informations tirées de l'essai de recensement, d'attribuer à chacun d'eux une variable numérique positive X_i qui traduit sa "difficulté". Cette variable a été calibrée de façon à ce que Y_i (qui vaut 1 en cas d'erreur et 0 sinon) ait une "espérance" proportionnelle à X_i.

Toujours pour les mêmes raisons de coût du contrôle, on est amené à envisager un sondage à deux degrés:

- au premier degré de sondage on tirera un échantillon s_1 de districts k (les Unités Primaires) à probabilités inégales π_k à déterminer. On notera $\pi_{k\ell}$ les probabilités d'inclusion double pour cet échantillonnage.
- au second degré de sondage, on tirera un échantillon s_k d'unités finales (les bulletins) dans l'unité primaire échantillon k. On notera $\pi_{i|k}$ la probabilité d'inclusion de l'unité dans l'unité primaire k, $\pi_{ij|k}$ la probabilité d'inclusion du couple (i,j) dans les unités primaires et $s = U_{k\in s_1} s_k$ l'échantillon d'unités finales.

On notera $X_k = \sum_{i\in k} X_i$ le total des X_i dans l'unité primaire k, $X = \sum_{k\in U_o} X_k = \sum_U X_i$ et on adoptera des notations analogues pour toutes les variables. (U_o désigne la population des Unités Primaires - districts, U la population des Unités finales - bulletins).

Le but est d'estimer une quantité de la forme $R = \sum_U Y_i / \sum_U W_i$ où W_i est une variable connue pour chaque bulletin. Cela pourra être $W_i = 1$ ou $W_i = X_i$ selon la mesure qui semble la plus adéquate du taux d'erreur.

4.1 Choix d'estimateur et variance

a) Au niveau de l'Unité Primaire numéro k il est naturel d'estimer le total Y_k des Y_i pour $i\in k$ par le ratio:

$$\hat{Y}_k = X_k \left(\sum_{s_k} Y_i/\pi_{i|k} \right) \Big/ \left(\sum_{s_k} X_i/\pi_{i|k} \right) = X_k \hat{a}_k.$$

Ici $\hat{a}_k$ estime $a_k = Y_k/X_k$ avec un faible biais.

b) Pour estimer le ratio Y/X on utilisera:

$$\hat{a} = \frac{\sum_{s_1} \frac{\hat{Y}_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}} = \frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}.$$

c) Si on veut estimer R, on remarquera que:

$$R = \frac{Y}{X} \cdot \frac{X}{W},$$

où X et W sont des totaux connus (difficulté totale et nombre total de bulletins, par exemple). Comme la variable X_i a été choisie pour sa bonne corrélation avec Y_i, un estimateur *a priori* intéressant de R sera:

$$\hat{R} = \hat{a} \frac{X}{W}$$

de sorte que la seule véritable question porte sur l'estimation de $a = \sum_k a_k X_k / X$.

d) On aura:

$$\mathrm{Var}(\hat{a}) = \mathrm{Var}E(\hat{a} \mid s_1) + E\,\mathrm{Var}(\hat{a} \mid s_1).$$

Pour le premier terme, compte tenu du fait que $\hat{a}_k$ estime (à peu près) sans biais a_k, on peut écrire:

$$\mathrm{Var}E(\hat{a} \mid s_1) \simeq \frac{1}{X^2} \mathrm{Var}\left(\sum_{s_1} \frac{(a_k - a)X_k}{\pi_k} \right)$$

$$= \frac{1}{X^2} \left(\sum_k \frac{(a_k - a)^2 X_k^2}{\pi_k^2} \right.$$

$$\left. + \sum\sum_{k \neq l} (a_k - a)(a_\ell - a) \frac{X_k X_\ell \pi_{k\ell}}{\pi_k \pi_\ell} \right). \quad (4.1.1)$$

Pour le second terme, on a, *conditionnellement à* s_1:

$$\mathrm{Var}\left(\frac{\sum_{s_1} \hat{a}_k \frac{X_k}{\pi_k}}{\sum_{s_1} \frac{X_k}{\pi_k}}\right) = \left(\sum_{s_1} \frac{X_k}{\pi_k}\right)^{-2} \cdot \sum_{s_1} \mathrm{Var}(\hat{a}_k) \frac{X_k^2}{\pi_k^2}.$$

L'espérance de cette quantité vaut approximativement

$$X^{-2} \sum_k E\ \mathrm{Var}(\hat{a}_k \mid s_1) \frac{X_k^2}{\pi_k}, \qquad (4.1.2)$$

avec:

$$\mathrm{Var}(\hat{a}_k \mid s_1) = \mathrm{Var} \frac{\sum_{s_k} \frac{Y_i}{\pi_{i|k}}}{\sum_{s_k} \frac{X_i}{\pi_{i|k}}} \simeq \frac{1}{X_k^2} \mathrm{Var} \sum_{s_k} \frac{Y_i - a_k X_i}{\pi_{i|k}}$$

$$= \frac{1}{X_k^2}\left(\sum_{i \in k} \frac{(Y_i - a_k X_i)^2}{\pi_{i|k}}\right.$$

$$\left. + \sum\sum_{k \neq l} \frac{(Y_i - a_k X_i)(Y_j - a_k X_j)\pi_{ij|k}}{\pi_{i|k}\ \pi_{j|k}}\right).$$

Comme dans les parties précédentes, nous arrivons à des formules complexes et, finalement, inutilisables. Un modèle va nous simplifier un peu l'existence.

4.2 Intervention d'un modèle

Il aura la même structure que ceux qui ont déjà servi antérieurement:

a) Les a_k seront des variables aléatoires indépendantes de même espérance et de même variance:

$$E_\xi\, a_k = a \qquad \mathrm{Var}_\xi\, a_k = \sigma^2.$$

La variance prend en compte l'influence de l'opérateur ou de l'opératrice, qu'on renonce à isoler, mais aussi celle du jour de la semaine, de l'heure dans la journée, de certains jours du mois *etc.*...

b) Conditionnellement à a_k, les Y_i de l'Unité Primaire k sont des variables de Bernoulli indépendantes avec $E_\xi(Y_i \mid k) = a_k X_i$

$$\mathrm{Var}_\xi(Y_i \mid k) = a_k X_i - a_k^2 X_i^2.$$

Remarque:
La variable X_i n'a pas de véritable sens concret et n'est d'ailleurs définie qu'à un facteur d'échelle près. En revanche aX_i et σX_i ont une interprétation physique invariante, car ce sont des probabilités. Dans tout ce qui suit il faudra toujours garder en tête que les résultats devront être invariants si les X_i sont multipliés par un facteur arbitraire à condition que a et σ soient divisés par le même facteur. En particulier $\mathrm{Var}(\hat{a})$ n'a pas de sens "concret". Seule $\mathrm{Var}(\hat{a}X)$ en a un.

Comme précédemment nous allons étudier la variance anticipée, espérance sous modèle de la somme de (4.1.1) et (4.1.2).

Pour le premier terme, l'espérance des produits croisés est nulle, comme de bien entendu. L'espérance sous modèle de ce terme est donc:

$$X^{-2}\sigma^2 \sum_k \frac{X_k^2}{\pi_k}.$$

Pour le second terme on trouve: (vu les définitions données au 4.2.a et 4.2.b):

$$X^{-2} \sum_k \frac{X_k^2}{\pi_k} \cdot \frac{1}{X_k^2} \sum_i E_\xi \frac{(a_k X_i - a_k^2 X_i^2)}{\pi_{i|k}}$$

$$= X^{-2} \sum_k \frac{1}{\pi_k} \sum_i \frac{aX_i - (a^2 + \sigma^2)X_i^2}{\pi_{i|k}}.$$

Globalement donc:

$$E_\xi\, \mathrm{Var}(\hat{a}X) = \sigma^2 \sum_{k \in U_o} \frac{X_k^2}{\pi_k}$$

$$+ \sum_{k \in U_o} \frac{1}{\pi_k} \sum_{i \in k} \frac{aX_i - (a^2 + \sigma^2)X_i^2}{\pi_{i|k}}.$$

Ici pas de miracle algébrique. *Pour simplifier* nous admettons que $(a^2 + \sigma^2)X_i^2$ est négligeable devant aX_i. Numériquement on peut attendre $aX_i = 2$ à 5×10^{-2} et $(a^2 + \sigma^2)X_i^2 = 3$ à 30×10^{-4}. D'où notre approximation

$$E_\xi\, \mathrm{Var}(\hat{a}X) \simeq \sigma^2 \sum_{k \in U_o} \frac{X_k^2}{\pi_k} + a \sum_{k \in U_o} \frac{1}{\pi_k} \sum_{i \in k} \frac{X_i}{\pi_{i|k}}.$$

4.3 Optimisation du plan de sondage

Nous utiliserons la fonction de coût suivante:

$$C = \sum_{s_1} (C_o + C_1 n_k).$$

Ici, $n_k = \sum_{i\in k} \pi_{i|k}$ est la taille de l'échantillon tiré dans le district k (supposé de taille fixe à s_1 fixé). Son espérance vaut:

$$C_T = \sum_{k\in U_o} \pi_k (C_o + C_1 n_k).$$

Posons

$$\pi_{i|k} = n_k P_i \left(\text{avec } \sum_{i\in k} P_i = 1\right) \quad \text{et} \quad Q_k = \pi_k n_k.$$

Le problème d'optimisation est maintenant:

$$\text{Min: } C_o \sum_k \pi_k + C_1 \sum_k Q_k$$

$$\text{sous: } \sigma^2 \sum_k \frac{X_k^2}{\pi_k} + a \sum_k \frac{1}{Q_k} \sum_{i\in k} \frac{X_i}{P_i} \leq \mathcal{V}_o.$$

Sous cette forme, on constate avec plaisir qu'on peut minimiser les termes en $\sum_i X_i/P_i$ indépendamment du reste. Autrement dit, n_k n'a pas d'incidence sur ce terme. Laissons l'optimisation du second degré de tirage pour plus tard et notons seulement S_k^{*2} la valeur optimisée de $\sum_i X_i/P_i$. Avec un multiplicateur de Lagrange λ on obtient par dérivation par rapport aux π_k puis au Q_k:

$$^*C_o = \lambda\sigma^2 \frac{X_k^2}{\pi_k^2} \text{ soit } \boxed{\pi_k \text{ proportionnel à } X_k} \quad (4.3.1)$$

$$^*C_1 = \lambda a \frac{S_k^{*2}}{Q_k^2} \text{ d'où } \boxed{n_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{S_k^*}{X_k}} \quad (4.3.2)$$

En particulier on tirera les Unités Primaires avec des probabilités proportionnelles à leur difficulté totale, ce qui est un résultat standard (voir par exemple Särndal, Swensson, Wretman 1992, chapitre 12).

Passons maintenant au tirage infra-district (deuxième degré de sondage).

Commençons par un cas simple et naïf: on tire les bulletins individuellement. La minimisation conduit à P_i proportionnelle à $\sqrt{X_i}$. Un calcul simple nous montre qu'alors $S_k^* = \sum_{i\in k} \sqrt{X_i}$. Ceci nous permet de calculer n_k grâce à (4.3.2) et notre problème est entièrement résolu.

En fait les choses sont plus compliquées. Pour des raisons assez naturelles, on ne sélectionnera que les bulletins que de ménages entiers. Autrement dit le sondage au second degré est un sondage en *grappes*. Les valeurs de P_i seront les mêmes, soit P_m, pour tous les membres d'une même grappe (ménage) m.

Notons par X_m la somme des X_i des individus i du ménage m. Le problème est donc de minimiser $\sum X_m/P_m$ sous $\sum n_m P_m = 1$ avec n_m taille du ménage m. On trouve facilement la solution:

$$P_m = \sqrt{\bar{X}_m} \Big/ \sum n_m \sqrt{\bar{X}_m},$$

avec $\bar{X}_m = X_m/n_m$, *difficulté moyenne des bulletins du ménage m*. Par suite on trouve $S_k^* = \sum n_m \sqrt{\bar{X}_m}$.

Cette solution nous permet de déterminer le nombre n_k *d'unités finales* à tirer grâce à (4.3.2). Le nombre de grappes (*ménages*), en revanche n'est pas déterminé. Cette difficulté était prévisible. La fonction de coût, en effet, ne fait pas intervenir cette contrainte. Pour obtenir le nombre m_k de grappes à tirer, on s'arrangera de façon à ce que l'espérance du nombre d'unités finales soit égale à n_k. Elle vaut:

$$m_k \left(\sum n_m \sqrt{\bar{X}_m}\right) \Big/ \sum \sqrt{\bar{X}_m}$$

d'où:

$$m_k = n_k \frac{\sum \sqrt{\bar{X}_m}}{\sum n_m \sqrt{\bar{X}_m}}.$$

Compte tenu de (4.3.2) on a aussi:

$$m_k = \left(\frac{C_o}{C_1}\right)^{1/2} \frac{a^{1/2}}{\sigma} \frac{\sum \sqrt{\bar{X}_m}}{X_k}$$

et la probabilité de tirer un ménage vaut alors:

$$\frac{m_k \sqrt{\bar{X}_m}}{\sum \sqrt{\bar{X}_m}}.$$

Après quelques manipulations algébriques, on trouve la valeur de la variance optimale:

$$E_\xi \text{ Var } (\hat{a}X)_{\text{OPT}} = \frac{(\sigma X)^2}{m} \left(1 + \frac{a}{\sigma} \frac{a^{-1/2} S^*}{X} \left(\frac{C_1}{C_o}\right)^{1/2}\right).$$

Cette forme respecte le caractère homogène des différents facteurs. On a, en particulier $a^{-1/2} S^*/X = a^{1/2} S^*/aX$; le dénominateur est interprétable comme un nombre total d'erreurs dans un lot, tandis que le numérateur est homogène à une taille.

Nous avons obtenu une solution complète du problème.

Remarque 1:
Dans les deux cas qui ont été traités S_k^* est multiplié par $C^{1/2}$ si les X_i sont multipliés par C. La formule qui donne n_k est donc bien invariante à l'échelle de mesure.

Remarque 2:
La solution dans le cas du tirage en grappes privilégie le tirage de petites grappes dont les unités finales on un fort indice de difficulté.

Remarque 3:
Ici comme dans les parties précédentes nous déterminons les probabilités d'inclusion simple mais pas les probabilités d'inclusion double. L'algorithme de tirage, qui fixe ces dernières, est donc sans influence. Ceci est relativement naturel si nous nous disons que l'information auxiliaire utilisée pour optimiser le tirage déterminera les π_k et $\pi_{i|k}$ mais ne peut pas avoir d'influence sur les probabilités doubles.

5. APPLICATIONS AUX CONTRÔLES PAR SONDAGE DE LA QUALITÉ DU RECENSEMENT DE 1990 EN FRANCE

5.1 Problème de contrôle de la saisie

Les techniques d'échantillonnage décrites aux paragraphes 2 et 3 ont été motivées par la nécessité de contrôler la saisie du recensement de 1990. Pour comprendre la nature des problèmes statistiques une description des principes d'exploitation est utile.

L'unité de base pour la collecte est le district, correspondant, en ville, a un pâté de maisons et, à la campagne, à un village ou une réunion de hameaux. Il peut comporter une population variant de zéro à environ deux mille habitants. La moyenne est de 150 logements pour 350 habitants environ.

À mesure de l'achèvement de la collecte et de sa vérification, les différents bulletins du recensement, notamment les bulletins individuels (BI) et les feuilles de logement (FL), sont soigneusement comptés pour chaque district. Les données récapitulatives des districts sont saisies sur support informatique, tandis que les bulletins groupés, dans une chemise de district, partent pour la saisie.

Des ensembles de districts groupant environ 100,000 logements sont constitués. Ce sont les unités de traitement (UT). Chaque UT est saisie par une entreprise à façon pour le compte de l'INSEE.

L'INSEE, le "client" en termes de théorie du contrôle, vérifie la qualité du travail de chaque façonnier en contrôlant par sondage un certain nombre de bulletins dans chaque UT.

Le but du sondage décrit au paragraphe 2 est d'estimer la proportion de bulletins erronnés dans chaque UT avec une précision (écart-type) de un point. La proportion maximum de bulletins erronnés ne saurait excéder 4%. Un test de recensement portant sur environ 400 districts permet d'estimer les valeurs des deux paramètres du modèle. On trouve:

$$\sigma^2 \simeq P^2 \simeq 14.10^{-4}$$

$$\tau^2 \simeq P \simeq 4.10^{-2}.$$

La fonction de coût (1.1) a pu être évaluée en temps de travail. Les mesures faites dans les ateliers ont permis d'estimer à 5 minutes le temps de manipulation d'une chemise de district (du moment où on va la chercher dans une étagère au moment où on la range) et à 30 secondes le temps de saisie d'un BI. Avec des données numériques, l'optimisation du plan avec les hypothèses du paragraphes 1 conduit à contrôler 40 districts par lot de traitement et 16 bulletins à l'intérieur de chaque district.

Après discussion de cette solution avec l'équipe responsable du recensement, il est apparu qu'il fallait, en fait, contrôler deux types de documents: les bulletins individuels (BI) et les feuilles de logement (FL). On avait été conduit à négliger ces dernières, en première approximation, parce qu'elles sont moins susceptibles de recéler des erreurs et que leur temps de codification est plus court (la moitié environ) que celui nécessaire pour un BI. Toutefois, dans certains districts, par exemple dans les communes très touristiques, on trouve une forte majorité de résidences secondaires et donc beaucoup de FL pour très peu de BI. Cette situation demande une étude particulière, dont la théorie a été faite au paragraphe 3.

Dans le cas du recensement le nombre G de groupes vaut 2 ($g = 1$ pour les BI et $g = 2$ pour les FL). Les données numériques relatives aux deux groupes étaient les suivantes:

- $P_1 = 0{,}04 \quad \sigma_1 = P_1 \quad \tau_1^2 = P_1(1 - P_1) - \sigma_1^2 = P_1 - 2P_1^2,$
- $P_2 = 0{,}01 \quad \sigma_2 = P_2 \quad \tau_2^2 = P_2 - P_2^2,$
- $\mathcal{V}_1 = (0{,}0075)^2 \quad \mathcal{V}_2 = (0{,}0150)^2.$

Pour la fonction de coût on a pris $C_o = 5$ minutes, $C_1 = 0{,}5$ minute et $C_2 = 0{,}25$ minute. L'optimisation du problème selon les hypothèses de la partie 3.2.b a conduit à examiner 73 districts par unité de traitement. La solution pratique à consisté à traiter 15 bulletins individuels par distcrit ainsi que les FL associées. Pour les districts comportant moins de 15 BI, l'intégralité des BI était traitée. Pour les districts vides de BI on traitait 4 FL (si ce nombre était inférieur au nombre de FL du district).

Remarque:
La technique de la partie 2 semble avoir un domaine d'application assez fréquent. Elle a, en particulier, été utilisée pour l'échantillonnage de l'enquête française de 1992 sur les migrations en ce qui concerne la population de nationalité étrangère. Pour les agglomérations de moins de 20,000 habitants, l'échantillon était à deux degrés, le premier degré de sondage étant constitué par les 90 départements où ce type d'agglomération existe. La population étrangère (sur la base du recensement) était divisée en 8 groupes de nationalités pour lesquels on devrait obtenir des indicateurs ayant la même précision.

5.2 Problèmes liés à la codification

La seconde étape d'élaboration des données est dite opération COLIBRI (pour Codification en Ligne des Bulletins du Recensement des Individus). Recevant des bulletins toujours groupés en districts, les opérateurs et opératrices des Directions Régionales de l'INSEE procédaient à leur codification pour constituer le sondage au quart.

Physiquement, chaque opérateur (trice) travaille devant un écran qui lui indique l'identifiant du prochain logement à inclure dans l'échantillon au quart dont il doit codifier tous les BI.

Le contrôle de la qualité de la codification est également réalisé par sondage. L'unité de contrôle est l'ensemble du travail réalisé en une semaine dans une Direction Régionale. L'opération dure un peu plus d'un an dans les 22 Directions Régionales soit plus de mille sondages. L'unité à contrôler est le ménage (c'est-à-dire l'ensemble des BI d'un ménage tiré pour figurer dans l'échantillon de contrôle). L'objectif est d'estimer la proportion de bulletins comportant une erreur. Pour cela, on détecte automatiquement ceux pour lesquels apparaît une divergence entre les deux codifications. Une opération de réconciliation permet de chiffrer le nombre d'erreurs. La théorie de ce contrôle a fait l'objet de la partie 4 de cet article. L'indice de difficulté des bulletins a été élaboré à partir des données déjà saisies pour une étude faite à partir du précédent recensement et d'un test. Les modalités pratiques et les enseignements tirés de ces contrôles sont détaillés dans G. Badeyan (1992).

L'application pratique et numérique de la théorie repose sur des hypothèses concernant les ordres de grandeurs des différents paramètres (ce qui demande qu'on puisse les raccrocher à une interprétation physique simple). Dans la phase de préparation du recensement, sans mesures préalables très précises, on a utilisé les valeurs $\sigma/a = 0{,}5$ et $C_1/C_o = 0{,}1$.

À la suite de diverses hypothèses sur les autres paramètres et de discussions entre experts, il a été décidé un contrôle portant sur 50 districts chacun d'eux étant contrôlé pour environ 20 BI (par région et par semaine). Cet ordre de grandeur initial pouvait, évidemment, être modulé dans la suite du contrôle, les paramètres du modèle pouvant être réestimés après chacun d'eux.

Remarque finale:

Ce problème fait apparaître des résultats un peu surprenants sur lesquels il est utile de réfléchir un peu.

Dans un premier cas, nous avons supposé qu'on pouvait isoler chaque bulletin. On tirait alors ceux-ci avec des probabilités proportionnelles à leur difficulté individuelle. On supposait, dans une certaine mesure, que le coût d'utilisation de l'information individuelle était nul.

Dans le second cas, la réalité du contrôle, ce coût était considéré comme infini et la seule information ayant un coût négligeable était celle relative à l'ensemble du ménage. La solution fait alors apparaître des probabilités de tirage des individus (BI) qui est fonction de la difficulté moyenne de codification des bulletins de l'ensemble du ménage auquel appartient cet individu.

Il en irait de même en ce qui concerne le tirage des districts. Si on sait y distinguer les BI, on les tire avec des probabilités proportionnelles à la difficulté totale; à l'intérieur des districts, on tirera des BI difficiles avec une plus grosse probabilité. Supposons, au contraire, qu'on ne sache pas distinguer les BI à l'intérieur des districts. Ce serait le cas, par exemple, si la désignation des BI à contrôler ne pouvait pas se faire en temps réel par suite d'une organisation du traitement inadéquate. On tirerait alors les districts proportionnellement à leur difficulté moyenne: à l'intérieur des districts, on serait obligé de réaliser des sondages aléatoires simples.

Dans le premier cas le sondage privilégiera les gros districts à l'intérieur desquels on tirera plutôt les BI difficiles. Dans le second cas, on privilégiera les petits districts difficiles à l'intérieur desquels on tirera des bulletins à probabilités égales. *Dans les deux cas*, on cherchera à augmenter la probabilité de sonder des BI difficiles. La différence réside simplement dans la possibilité (c'est-à-dire le coût) de mobiliser l'information au moment où on en a besoin.

REMERCIEMENTS

L'auteur tient à remercier de leurs commentaires très positifs le rédacteur en chef, le rédacteur associé et l'arbitre qui ont examiné ce travail. Il en va de même de Claude Thelot dont les remarques sont à l'origine d'un certain nombre de développements et de Gérard Badeyan qui a mis en place à l'INSEE les techniques ici préconisées. Il remercie Françoise Hitier sans qui ce texte n'aurait jamais pu exister.

BIBLIOGRAPHIE

BADEYAN, G. (1992). Communication aux secondes Journées de Méthodologie Statistique, 17 et 18 juin 1992, INSEE, Paris.

COCHRAN, W. (1977). *Sampling Techniques*, (3ième édition). New York: Wiley.

DESABIE, J. (1965). *Théorie et Pratique des Sondages*. Paris: Dunod.

LUENBERGER, D.G. (1973) *Introduction To linear and Non-linear Programming*. New York: Addison-Wesley.

SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.