Calcul de précision et plan de sondage : application aux enquêtes camerounaises auprès des ménages (ECAM 2 et ECAM 3)

Justin Bem
Martin Mba
Ludovic Subran [1]

Cet article aborde l'estimation de variance dans le cas de plans de sondage et de statistiques complexes. Recours est fait à la linéarisation et aux techniques de réplication pour estimer la précision des enquêtes camerounaises auprès des ménages, ECAM 2 et ECAM 3. Les résultats obtenus montrent que la qualité de l'enquête du point de vue de la précision des estimations est très satisfaisante au vu des normes internationales. Ils permettent d'établir des recommandations pour optimiser le plan de sondage des enquêtes auprès des ménages à venir au Cameroun, concernant la taille et l'allocation de l'échantillon et les critères de stratification. Ainsi il est démontré que les coûts des enquêtes futures relativement à ceux des enquêtes ECAM 2 et 3 pourraient être réduits tout en gardant un même niveau de précision.

Introduction

La plupart des enquêtes auprès des ménages sont réalisées à partir de plans d'échantillonnage complexes. Ce sont généralement des enquêtes stratifiées, à plusieurs degrés et/ou à probabilités de tirage inégales ; l'objectif poursuivi restant la précision des estimateurs et la maîtrise des coûts de collecte. Cependant, l'estimation statistique devient plus complexe du fait notamment du système de pondération et des techniques d'estimation de variance (nécessité de calculer des probabilités d'inclusion double, d'établir des formules de variance pour des statistiques non linéaires ou définies implicitement).

De fait, le calcul de la variance des estimateurs est rarement mis en œuvre en Afrique subsaharienne. Aussi, les utilisateurs de données d'enquêtes ne disposent pas souvent d'indicateurs objectifs du degré de précision des statistiques et donc de leur qualité. Cet article se propose d'examiner le problème de la mesure de précision des indicateurs. Dans la première partie, les principales techniques d'estimation de la variance de différents indicateurs statistiques sont exposées, tandis qu'en seconde partie les plans de sondage des enquêtes camerounaises auprès des ménages (ECAM 2 et 3) et leurs incidences sur la précision des indicateurs d'intérêt pour le suivi de la pauvreté sont analysés. Ce travail permet d'évaluer le coût d'opportunité

[1] *Justin Bem : Institut Sous-régional de Statistique et d'Economie Appliquée (ISSEA), Cameroun. Martin Mba : Institut National de la Statistique, Cameroun. Ludovic Subran : Institut Nationale de la Statistique et des Etudes Economiques (INSEE), Paris.*
Les auteurs remercient vivement Thomas LUMLEY de l'Université de Washington pour les avoir orientés dans leur démarche méthodologique et la manipulation du package 'survey' du logiciel R.

informationnel des plans de sondage retenus par l'Institut National de la Statistique (INS) du Cameroun.

Nous montrons que le niveau de précision de certaines statistiques de l'enquête ECAM 2 est assez fin et que les coûts des enquêtes futures relativement à ceux des enquêtes ECAM 2 et 3 pourraient être réduits en réduisant la taille des échantillon tout en gardant un même niveau de précision.

Cadre théorique

Il est possible de résumer à quatre cas les problèmes d'estimation de variance sur des données d'enquête, selon la complexité de la statistique et du plan de sondage :

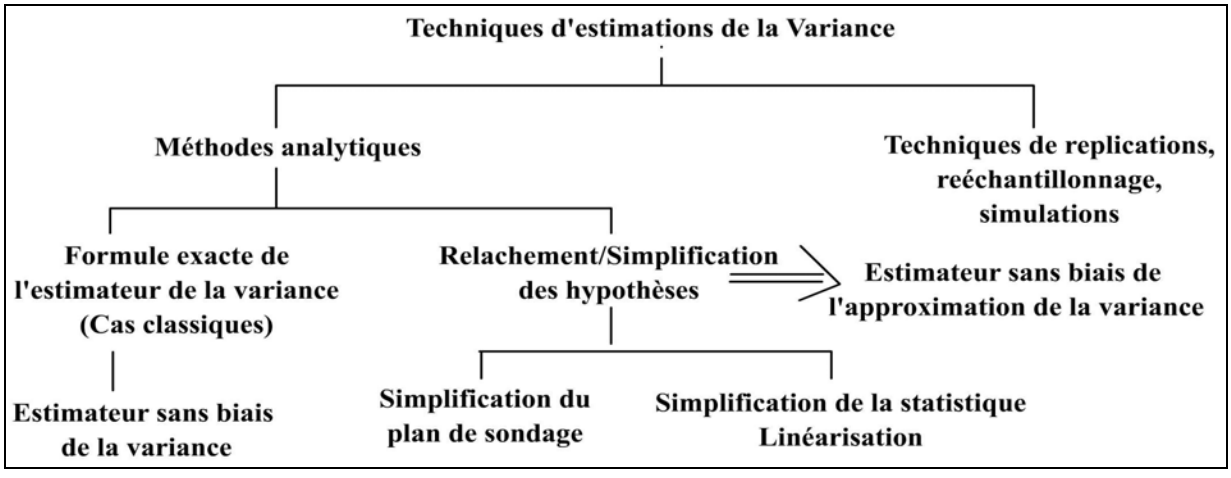
Tableau 1 :
Plan de sondage et estimateurs :

	Plan simple	Plan complexe
Statistique linéaire (ex : consommation moyenne)	a	b
Statistique complexe (ex : indice de Gini)	c	d

Source : Wolter (2003).

La théorie des sondages offre des méthodes d'estimation sans biais de variance dans le cas (a). Plusieurs ouvrages, dont celui d'Ardilly (1994), en font une présentation exhaustive. Quant aux cas (b), (c) et (d), ils sont complexes et la théorie ne fournit que des méthodes d'estimation sans biais de l'approximation de la variance. Deux groupes de méthodes sont concurrentes pour l'estimation : les méthodes analytiques, principalement la linéarisation, et les techniques de réplication. Le schéma suivant récapitule les techniques d'estimation de précision d'indicateurs. Premièrement, nous abordons le cas classique d'estimation sans biais de la variance de statistiques simples. Deuxièmement des méthodes de linéarisation (linéarisation de Taylor simple pour des statistiques linéaires et méthode des équations estimantes pour les statistiques plus complexes) permettant d'estimer sans biais une approximation de la variance sont présentées. Enfin, les techniques de réplication sont discutées.

Figure 1 :
Techniques d'estimation de variance :



Source : European Communities (2002).

Quelques résultats généraux pour l'estimation de variance

Soient *U* une population de taille *N* et *S* un échantillon de *n* unités tirées de cette population.

π_k est la probabilité d'inclusion dans l'échantillon de l'unité k, et π_{kl} est la probabilité d'inclusion double, simultanée, des unités k et l. Y_k est la valeur de la variable d'intérêt pour l'unité k de la population.

La somme de la variable d'intérêt s'écrit :

$$Y = \sum_{k \in U} Y_k$$

L'estimateur de Horvitz-Thompson (HT) du total Y est donné par :

$$\hat{Y} = \sum_{k \in S} \frac{Y_k}{\pi_k}$$

L'estimateur de HT est sans biais, et une estimation sans biais de sa variance est l'estimateur de Yates-Grundy défini par :

$$\hat{V}\left(\hat{Y}\right) = \frac{1}{2} \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \left(\pi_k \pi_l - \pi_{kl}\right) \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right)^2$$

Dans le cas d'un sondage aléatoire simple (SAS) sans remise, $\pi_k = \frac{n}{N}$, $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$. L'estimateur de variance se simplifie en :

$$\hat{V}_{SAS}\left(\hat{Y}\right) = \frac{N\left(N-n\right)}{n\left(n-1\right)} \sum_{k \in S} \left(Y_k - \hat{Y}\right)^2$$

Pour d'autres plans de sondage, l'adaptation n'est pas aisée du fait du calcul des probabilités d'inclusion double. Par exemple, pour un sondage à plusieurs degrés, les formules se compliquent. Pour ce type de tirages, un des modes de calcul de précision privilégié est le calcul récursif :

$$V(.) = \underbrace{E_1(V_{2/1}(.))}_{\text{2e degré}} + \underbrace{V_1(E_{2/1}(.))}_{\text{1er degré}}$$

La formule de Raj ci-dessous donne l'estimateur de la variance dans le cas où les unités secondaires sont tirées indépendamment d'une unité primaire à l'autre :

$$\hat{V}\left(\hat{Y}\right) = \hat{V}_1\left(\sum_{i \in S_1} \frac{T_i}{\pi_i} \right) + \sum_{i \in S_1} \frac{\hat{V}_i}{\pi_i^2}$$

où S_1 est l'ensemble des unités primaires (UP) échantillonnées, T_i et V_i sont respectivement la somme de *Y* dans la *i*ème UP et sa variance, et π_i est la probabilité de tirage de l'UP i. Si les vraies valeurs des totaux par UP (des T_i) sont inconnues, elles sont remplacées par leurs estimateurs sans biais.

Cette formule également se simplifie dans le cas d'un SAS à chaque degré et devient :

$$\hat{V}(\hat{Y}) = \underbrace{M^2\left(1 - \frac{m}{M}\right)\frac{s_1}{m}}_{a} + \underbrace{\frac{M}{m}\sum_{i=1}^{M} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{s_{2,i}^2}{n_i}}_{b}$$

où M représente le nombre total d'unités primaires dans la population, m le nombre d'unités primaires tirées, s_1 est l'estimateur de la variance entre les totaux des unités primaires et $s_{2,i}$ la variance estimée au sein de l'unité primaire i

Le terme a est la composante du premier degré, mesurant la dispersion entre les totaux des UP, *b* celle du deuxième degré mesurant la dispersion à l'intérieur des UP. Les formules ne se compliquent pas dans le cas d'un sondage stratifié puisque celui-ci peut-être considéré comme un sondage à deux degrés dont le premier degré est un tirage exhaustif. La formule de Raj se limite au deuxième terme, la variance inter-strate étant nulle.

L'effet du plan de sondage (ou *design effect*) mesure l'erreur faite en estimant la variance d'une statistique en ignorant le plan de sondage, c'est-à-dire en supposant à tort qu'il s'agit d'un SAS. Pour la moyenne de la variable *Y* et le plan de sondage complexe *P*, il est défini par :

$$Deff = V_P\left(\bar{Y}\right) \Big/ V_{SAS}\left(\bar{Y}\right)$$

Il est nécessaire d'avoir une valeur opérationnelle du design effect, aussi est-il estimé. L'effet du plan de sondage peut être dû à la stratification, à des probabilités inégales de tirage et à l'effet de grappe (ou *cluster effect*). A tort, l'effet de grappe est souvent supposé être la seule cause du design effect[2]. Dans le cas d'un plan non stratifié à deux degrés avec un SAS à chaque degré, c'est vrai, et alors :

(Cluster Effect) $$\rho = Deff - 1 \Big/ \left(\bar{n} - 1\right)$$

où $\bar{n}$ est le nombre d'unités statistiques tirées dans chaque unité primaire. En revanche, lorsque le plan est en plus stratifié, ou que le tirage des UP est fait selon des probabilités inégales (par exemple proportionnelles à leur taille), cette formule n'est plus vérifiée et il n'existe plus d'expression mathématique simple de ρ. Il est alors d'usage, par

[2] *Effet de grappe et design effet sont même souvent confondus. Pourtant Deff est toujours positif et s'interprète en fonction de sa position par rapport à 1, alors que ρ peut être positif ou négatif.*

convention, de garder la formule ci-dessus, mais en considérant $\bar{n}$ comme le rapport de la taille totale de l'échantillon sur le nombre total d'UP tirées.

Dans le cas d'un plan de sondage seulement stratifié, avec un SAS dans chaque strate, le Deff est inférieur à 1. En effet, la stratification diminue la variance et donc la qualité de l'estimation serait sous-estimée si elle n'était pas prise en compte dans le calcul de variance.

La linéarisation de Taylor

La linéarisation est utilisée dans le cadre des statistiques complexes pour lesquelles il est difficile de disposer d'une formule explicite de la variance (pour les statistiques n'étant pas des fonctions linéaires d'estimateurs de Horvitz-Thompson). Si cette difficulté est perceptible dans le cas des estimateurs complexes tels que l'indice de Gini, la difficulté existe aussi pour certains paramètres usuels tels que les ratios, voire les moyennes[3].

Statistiques simples ou paramètres explicitement définis

Dans une enquête par sondage, les paramètres couramment estimés sont les totaux, les moyennes, les proportions et les ratios. Nous appelons ici paramètre explicitement défini tout paramètre défini par une fonction explicite.

Théorème : Soit T une statistique telle que $T \xrightarrow{n\to+\infty} N\left(\theta, \frac{\sigma(\theta)}{\sqrt{n}}\right)$ et g une fonction explicite continûment dérivable, alors $g(T) \xrightarrow{n\to+\infty} N\left(g(\theta), \frac{g'(\theta).\sigma(\theta)}{\sqrt{n}}\right)$

Donc la variance asymptotique d'un paramètre explicitement défini par g est :

$$V(g(T)) = (g'(\theta))^2 . V(T)$$

Il s'agit de l'approximation de la variance par la méthode de linéarisation de Taylor. Elle repose sur un développement limité à l'ordre 1 de $g(T)$ au voisinage de θ.

Statistiques complexes et équations estimantes

Si les techniques de linéarisation sont désormais classiques pour les statistiques simples, elles ne sont pas adaptées aux statistiques plus compliquées telles que celles utilisées dans les analyses de distribution de revenu. La théorie des équations estimantes (Binder, 1993) peut en revanche s'appliquer à la grande majorité des estimateurs utilisés en pratique.

Soit *F* la fonction de répartition de la variable aléatoire d'intérêt *Y* positive.

Nous appelons ici paramètre implicitement défini tout paramètre d'intérêt défini comme solution d'une équation de la forme :

(E) $$U(\theta) = \int u(y,\theta) dF(y) = 0$$

Cette technique s'applique en particulier aux paramètres suivants :

Tableau 2 :
Quelques statistiques complexes et leur fonction estimante

	solution de *(E)*	fonction u associée
quantile d'ordre p	$Q(p) = F^{-1}(p)$	$u(y,\theta) = I(y < \theta) - p$
coefficient de régression	$b = (x'x)^{-1} x'y$	$u(y,x,\theta) = x(y - x'\theta)$
indice de Gini	$G = \frac{1}{\mu_y}\int_0^{+\infty} (2F(y)-1).y.dF(y)$	$u(y,\theta) = (2F(y)-1).y - \theta y$
ratio de X sur Y	$R = \bar{x}/\bar{y}$	$u(y,x,\theta) = y - x\theta$

Sous l'hypothèse de normalité asymptotique de θ, la variance de tels estimateurs peut être approchée par la variance de :

$$u^*(y) = -\left[\frac{\partial E[u(y,\theta)]}{\partial\theta}\bigg|_{\theta=\theta_0}\right]^{-1} u(y,\theta_0)$$

La démonstration se fait par un développement limité à l'ordre 1 de $U(\hat{\theta})$ au point $\hat{\theta} = \theta_0$, la

[3] *La linéarisation est même utile dans le cas d'estimation de moyenne. Si la taille de la population est inconnue, elle doit être estimée. La moyenne est alors le rapport de deux totaux, donc un ratio.*

vraie valeur du paramètre. Le calcul des fonctions u^* peut être très fastidieux. C'est le cas notamment pour l'indice de Gini[4], pour lequel :

$$u^* = \frac{1}{\mu}\left[\int_y^\infty J'[F(x)]xdF(x) - E\{F(y)J'[F(y)]y\} + \{J[F(y)]y - Gy\}\right]$$

avec $J(p) = 2p - 1$.

Les techniques de réplication

Le principe de ces techniques est de construire un nombre *R* d'échantillons à partir de l'échantillon initial et de calculer la statistique d'intérêt sur chacun de ces "nouveaux" échantillons. La variance de l'estimateur est alors approchée par la variance observée sur l'ensemble des échantillons :

$$V(\hat{\theta}) = c\sum_{r=1}^{R} h_r(\hat{\theta}_r - \hat{\theta})^2$$

$\hat{\theta}$ est l'estimateur sur l'échantillon initial, $\hat{\theta}_r$ est l'estimateur sur le *r*ème échantillon, R est le nombre d'échantillons crées par les réplications, c un paramètre dépendant de la méthode de réplication et h_r est le poids accordé à chaque échantillon créé.

La méthode de bootstrap est une des méthodes de réplication les plus utilisées. Elle consiste à tirer avec remise R échantillons de même taille que l'échantillon initial. La méthode du jackknife consiste à supprimer de l'échantillon initial une UP par strate, les autres étant repondérées. Dans un sondage non stratifié, il s'agit de JK1 jackknife ; dans un sondage stratifié, de JKn jackknife où n est le nombre de strate. Le nombre de réplications est donc le nombre de strates.

Le cas particulier d'un plan de sondage stratifié, avec deux UP tirées dans chaque strate, est le cas d'application de la méthode des réplications répétées équilibrées (BRR pour Balanced Repeated Replicates). A chaque réplication, un échantillon est tiré en supprimant une UP par strate (2^H échantillons pour *H* strates).

Tableau 3 :
Paramètres de la variance approchée selon la méthode de réplication

Méthode	c	h_g
réplications bootstrap	$\frac{1}{R-1}$	1
jackknife sans stratification (JK1)	$\frac{n-1}{n}$	1
jackknife avec n strates (JKn)	1	$\frac{n_h-1}{n_h}$
réplications répétées équilibrées	$\frac{1}{H}$	1

Précision des indicateurs de niveau de vie des enquêtes camerounaises de niveau de vie et optimisation des plans de sondage

Dans cette section, nous examinons la précision des indicateurs de niveaux de vie de l'enquête ECAM 2 au regard de choix de plan de sondage. Fort de ce résultat, l'optimalité des plans de sondage des enquêtes ECAM 2 et 3, effectuées respectivement en 2001 et 2007 est ensuite discutée.

Description des données

Les strates géographiques ont été définies par l'Institut National de la Statistique (INS) du Cameroun de la manière suivante : les arrondissements des villes de moins de 10 000 habitants sont considérés comme ruraux, ceux des villes de 10 000 à 50 000 habitants comme semi-urbains et pour finir ceux des villes de plus de 50 000 habitants sont considérés urbains.

Chaque province a été divisée en une strate urbaine, une strate semi-urbaine et une strate rurale. Les villes de Yaoundé et Douala constituent à elles seules deux strates urbaines. Au total, 32 strates ont été créées : 12 strates urbaines (Yaoundé, Douala et les 10 strates urbaines des provinces), 10 strates semi-urbaines et 10 strates rurales. Les strates rurales et péri-urbaines sont sondées de la même façon. Aussi seront-elles toutes appelées strates rurales dans la suite de l'article

La base de sondage utilisée est la base des zones de dénombrement (ZD) du Recensement Général de la Population et de l'Habitat (RGPH) de 1987. Une ZD est une zone géographique d'environ 200 ménages (soit environ 1 000 habitants).

Plan de sondage ECAM 2

Dans les strates urbaines le plan de sondage est à deux degrés. Au premier degré les ZD sont tirées selon un sondage aléatoire simple, puis dans chaque ZD, les ménages sont tirés à leur tour selon un sondage aléatoire simple. Pour des raisons de coût,

[4] *Le lecteur peut trouver le calcul complet dans Binder (1993).*

le tirage dans les strates rurales est à trois degrés. Au premier degré les arrondissements sont échantillonnés proportionnellement à leur taille en ménages de 1987. Au second degré, dans chaque arrondissement tiré, les ZD sont tirées selon un SAS, puis les ménages sont tirés selon un SAS dans les ZD échantillonnées. Au total, 11 553 ménages de 612 ZD ont été échantillonnés et 10 992 ont été enquêtés.

Le plan de sondage de l'ECAM2 est donc très complexe car stratifié, avec des modes de tirages différents selon les strates, à deux ou trois degrés selon les strates, et à probabilités inégales.

Plan de sondage ECAM 3

Concernant l'ECAM 3 effectuée en 2007, L'INS a opté pour un tirage à deux degrés dans toutes les strates. Disposant d'une ligne budgétaire comparable à celle de 2001, un échantillon de même taille a été envisagé (environ 12 000 ménages). Le nombre de ZD passe de 612 à 742. Compte tenu de la dispersion géographique, le fait de supprimer le troisième degré s'est traduit par l'augmentation du nombre de ZD, ce qui somme toute est de nature à améliorer la précision.

Choix méthodologiques nécessaires à l'application

La première option méthodologique choisie est de considérer le même nombre de degrés de sondage dans toutes les strates. La première façon de le faire est de considérer l'enquête comme étant à trois degrés dans toutes les strates en assimilant le premier degré dans les strates urbaines à un recensement. La deuxième façon est de considérer l'enquête comme étant à deux degrés dans toutes les strates, en regroupant les UP des zones rurales, indépendamment des arrondissements, c'est-à-dire en négligeant le premier degré en zone rurale. Nous avons opté pour la parcimonie en élimant le premier degré de la zone rurale.

Nous avons négligé le fait que dans la pratique, les tirages soient systématiques (possibilité que certaines probabilités d'inclusion double soient nulles) et non aléatoires comme dans la très grande majorité des enquêtes. Cependant si les éléments de la liste sur laquelle est opéré un tirage systématique sont dans un ordre aléatoire, le tirage systématique peut être assimilé à un tirage aléatoire.

Nous comparons ensuite les variances estimées à l'aide des méthodes de linéarisation, de bootstrap et de jackknife. En effet nos données ne se prêtent pas à la méthode des BRR (Balanced Repeated Replicates), puisque dans chaque strate, plus de 2 UP sont tirées. Le langage S (Chambers) via le logiciel R est utilisé pour implémenter les estimations[5]. Le programme que nous avons rédigé figure en annexe.

Principaux résultats

La norme retenue en matière de précision est celle de l'Institut de la Statistique du Canada (2006) présentée ci-dessous. Elle est basée sur le coefficient de variation de l'estimation, rapport de l'erreur-type sur l'estimation de la statistique d'intérêt elle-même.

Tableau 4 :
Norme en matière de précision (Statistique Canada 2006)

Coefficient de Variation (%)	Catégorie	Recommandations
inférieur à 16,5%	Acceptable	Diffusable sans aucune restriction
entre 16,6% et 33,3%	Médiocre	A utiliser avec prudence
supérieur à 33,3	Inacceptable	Préférable de ne pas diffuser

Source : Statistique Canada (2006).

Précision des indicateurs de niveau de vie de l'ECAM 2

Consommation annuelle totale moyenne par ménage.

La consommation moyenne des ménages est un paramètre simple à estimer car explicitement défini. La variance de l'estimation peut-être établie elle aussi analytiquement. La méthode analytique est basée sur les formules exactes de variance. Pour la méthode JKn le nombre de réplications est le nombre d'UP, soit 612 réplications. L'estimation par bootstrap est menée sur la base de 1 000 réplications. Les résultats des trois méthodes sont similaires. Les coefficients de variations sont tous très inférieurs au seuil de 16 %, traduisant la qualité des données.

[5] *Le logiciel R implémente le langage S développé par John Chambers. Il s'agit d'un langage de programmation dédié à la manipulation des données. L'implémentation R de S est libre : chaque utilisateur a accès au code source et peut faire bénéficier les autres utilisateurs des améliorations qu'il y apporte. Différents packages sont téléchargeables sur le site du Comprehensive R Archive Network (http://cran.r-project.org). Le package 'survey' de T. LUMLEY (Département de Biostatistique de l'Université de Washington) permet la prise en compte du plan de sondage dans le calcul d'estimateurs et de leur variance.*

Tableau 5 :
Précision de l'estimateur de consommation annuelle totale moyenne par ménage, ECAM 2

Régions	Estimation (en milliers CFA)	Méthode analytique			Bootstrap			Jackknife		
		Et	Cv (%)	deff	Et	Cv (%)	deff	Et	Cv (%)	deff
Douala	2 484	158	6,36	2,75	164	6,62	2,99	156	6,26	2,67
Yaoundé	2 475	142	5,73	1,98	145	5,88	2,08	139	5,60	1,89
Urbain	2 151	63	2,93	2,85	65	2,99	2,97	62	2,88	2,76
Rural	1 024	22	2,18	4,56	23	2,20	4,66	22	2,17	4,53
Cameroun	1 418	29,7	2,10	3,09	30	2,12	3,15	29	2,08	3,04

Source :ECAM 2, INS Cameroun, calculs des auteurs.

Et : Ecart-type, Cv : Coefficient de variation, deff : design effect

L'indice de Gini

La précision de l'indice de Gini est obtenue à l'aide de méthode des équations estimantes.

Tableau 6 :
Précision de l'indice de Gini, ECAM 2

Régions	Estimation	Bootstrap		Linéarisation		
		ET	CV (%)	ET	CV (%)	Deff
Douala	0,426	0,03	7,00	0,02	4,70	2,71
Yaoundé	0,427	0,02	4,70	0,03	7,00	2,77
Urbain	0,407	0,01	2,50	0,01	2,50	3,38
Rural	0,332	0,01	3,00	0,01	3,00	6,99
Cameroun	0,404	0,01	2,50	0,01	2,50	6,63

Source :Voir Tableau 5.

Et : Erreur-type, Cv : Coefficient de variation, deff : design effect.

La méthode analytique (linéarisation à l'aide de la fonction estimante) et le bootstrap donnent pour l'indice de Gini également, des résultats très proches. Le niveau de précision est très satisfaisant selon les normes de Statistique Canada (Tableau 2).

Pour les deux indicateurs de niveaux de vie, les effets du plan de sondage sont importants puisqu'à chaque fois bien supérieurs à 1. La stratification ayant pour effet de diminuer l'effet de sondage, l'effet de grappe pur est d'autant plus important. Cela traduit une forte ressemblance en termes de consommation des unités à l'intérieur des unités primaires. Il serait donc souhaitable à l'avenir de réduire le nombre de ménages tirés par ZD, mais d'augmenter le nombre de ZD tirées. Ce phénomène est amplifié en zone rurale.

Optimalité des plans de sondage

Taille de l'échantillon et allocation de l'échantillon

Les résultats obtenus ci-dessus permettent de déterminer la taille optimale de l'échantillon des prochaines enquêtes auprès des ménages ainsi que son allocation en termes d'unités primaires et secondaires à tirer. L'allocation entre les strates est étudiée dans Nguetse (2007).

Ardilly (1994) propose une méthode d'allocation optimale entre le nombre d'UP et le nombre d'unités secondaires interrogées par UP. Soient m le nombre d'UP tirées, et $\bar{n}$ le nombre d'unités secondaires tirées dans chaque UP. Le coût de collecte C est supposé composé du coût de déplacement pour se rendre à une unité primaire, c_1, et du coût de l'interview d'une unité secondaire, c_2. La contrainte budgétaire s'écrit :

$$C = c_1 m + c_2 m \bar{n}$$

Le problème de minimisation de la variance de l'estimateur de la moyenne de la variable d'intérêt Y sous la contrainte budgétaire a pour solution :

$$\bar{n} = \sqrt{\frac{c_1}{c_2}\frac{1-\rho}{\rho}}, m = \frac{C}{c_1 + \sqrt{c_1 c_2 \frac{1-\rho}{\rho}}}$$

où ρ est l'effet de grappe, calculé grâce à *(deff)*.

A partir des informations budgétaires de l'enquête, nous estimons c_1 à 149 346 FCFA et c_2 à 32 533 FCFA[6]. Nous obtenons les résultats ci-après :

Milieu	STRATE	Echantillon réalisé		Echantillon théorique			
		ECAM2	ECAM3	Effet grappe	ZD	Ménages	Echantillon
Urbain	Douala	1 200	1 260	0,159	135	5	667
	Yaoundé	1 200	1 248	0,089	113	7	771
	Adamaoua	270	282	0,175	39	5	180
	Centre - Yaoundé	216	282	0,181	33	5	151
	Est	270	282	0,049	26	9	241
	Extrême-nord	468	666	0,129	55	6	305
	Littoral - douala	270	390	0,072	29	8	224
	Nord	216	414	0,209	35	4	145
	Nord-ouest	324	684	0,063	32	8	265
	Ouest	414	618	0,001	7	68	470
	Sud	270	282	0,004	9	34	315
	Sud-ouest	216	672	0,108	28	6	174
Semi-urbain	Adamaoua	180	114	0,196	30	4	130
	Centre - Yaoundé	270	162	0,07	29	8	225
	Est	180	132	0,011	11	20	218
	Extrême-nord	288	132	0,088	33	7	226
	Littoral - douala	180	132	0,065	21	8	171
	Nord	270	114	0,03	21	12	260
	Nord-ouest	216	192	0,005	9	30	264
	Ouest	216	168	0,231	36	4	140
	Sud	180	114	0,007	9	26	227
	Sud-ouest	270	150	0,11	33	6	204
Rural	Adamaoua	324	204	0,126	40	6	227
	Centre - Yaoundé	396	411	0,051	35	9	323
	Est	324	204	0,06	32	8	267
	Extrême-nord	594	867	0,173	74	5	344
	Littoral - douala	324	204	0,029	24	12	301
	Nord	396	336	0,042	33	10	334
	Nord-ouest	432	699	0,15	54	5	274
	Ouest	459	576	0,03	33	12	397
	Sud	324	186	0,098	37	7	241
	Sud-ouest	396	432	0,012	20	19	391
Cameroun		**11 553**	**12 609**		**1 152**		**9 072**

Sources :ECAM 2 et ECAM 3, INS Cameroun, calculs des auteurs.

[6] *Pour estimer c1 et c2, nous avons considéré les budgets alloués à l'acquisition du matériel de terrain, la production des documents techniques et de collecte, l'acquisition et l'utilisation du matériel roulant et ceux de collecte et de saisie des données.*

On obtient un échantillon de 9 072 ménages avec 1 152 ZD. Le nombre de ZD à enquêter paraît élevé comparativement à ECAM 2 et ECAM 3 (respectivement 612 et 742 ZD). Ceci se justifie par la valeur relativement faible estimée de c_1 (car il faut noter que le nombre de ZD à tirer est une fonction décroissante de c_1). En effet, nous n'avons pas considéré la valeur d'achat des véhicules utilisés pour mener ces enquêtes dans l'estimation des coûts de transports mais seulement le quart de cette valeur (à laquelle nous avons ajouté les dépenses de carburants), dans la mesure où Ces véhicules ont été aussi utilisés pour d'autres enquêtes de l'INS. Nous avons donc privilégié leur valeur d'usage.

Les plus grands écarts entre nos résultats et ECAM 3 sont obtenus dans l'extrême-nord et celle du Nord-ouest avec 1 665 (dans 90 ZD) et 1 575 (dans 85 ZD) ménages, alors que nos calculs recommandent respectivement 875 (dans 161 ZD) et 802 (94 ZD). Ces deux provinces sont les plus pauvres, fortement rurales, et donc avec une faible dispersion des niveaux de vie. La justification de ces écarts pourrait se retrouver dans le fait que dans ces deux provinces notre procédure s'est basé sur une estimation du coût de déplacement plus faible que dans la normale, la province de l'extrême-nord étant très étendue et celle du nord-ouest très enclavée, d'où la concentration de l'échantillon dans un nombre moindre de ZD.

Des écarts importants sont aussi obtenus pour les strates de Yaoundé et Douala. Il faut noter qu'ils peuvent se justifier dans la mesure où une stratification secondaire est réalisée au sein de ses strates sur les arrondissements. Le cas de la province du Sud-ouest est aussi à souligner : au total 882 ménages avaient été tirés dans cette strate pour l'ECAM 2, et un quasi-doublement du tirage a été opéré pour l'ECAM 3. Cela peut se justifier dans la mesure où la dispersion de cette strate est élevée. Mais sans augmentation importante du nombre de ZD, une meilleure précision des estimateurs n'est pas garantie.

Dans les provinces de l'Est et du Sud, notre procédure donne des échantillons plus grands que ceux retenus dans ECAM 3, le niveau de dispersion dans ces strates est aussi important qu'ailleurs, l'argument de faible population ne devrait pas justifier à lui seul de si petits échantillons dans ces strates.

Quant à la procédure d'allocation suggérée par Nguetse (2007), les différences résident dans le fait que celle-ci ne prend pas en compte les coûts de la collecte et ne tient pas compte du fait que le tirage est à plusieurs degrés. C'est ainsi que l'échantillon utilisé pour ECAM 3 et celui suggéré par Nguetse sont plus onéreux que celui obtenu par nos calculs.

Stratification

Nous évaluons la qualité de la stratification de l'ECAM 2 à l'aide d'un test d'homogénéité des niveaux de vie entre les strates. Le test est mené séparément pour les strates urbaines (sans les strates de Yaoundé et Douala) et pour les strates rurales.

L'hypothèse nulle du test s'écrit :

$$H_0 : \hat{y}_1 = \hat{y}_2 = \cdots = \hat{y}_H \text{ contre } \exists (k,l)/\hat{y}_k \neq \hat{y}_l$$

De façon littéraire H_0 signifie que la dépense de consommation moyenne est la même dans toutes les strates. La vérification de H_0 impliquerait que la stratification n'est pas nécessaire.

Sous H_0 :

$$\Delta = \begin{pmatrix} \hat{y}_1 - \hat{y}_2 \\ \hat{y}_2 - \hat{y}_3 \\ \vdots \\ \hat{y}_{H-1} - \hat{y}_H \end{pmatrix} \approx N(0, \Sigma)$$

avec

$$\Sigma = \begin{pmatrix} v(\hat{y}_1)+v(\hat{y}_2) & -v(\hat{y}_2) & 0 & \dots & 0 \\ -v(\hat{y}_2) & v(\hat{y}_2)+v(\hat{y}_3) & -v(\hat{y}_3) & \dots & 0 \\ 0 & -v(\hat{y}_3) & v(\hat{y}_3)+v(\hat{y}_4) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & v(\hat{y}_{H-1})+v(\hat{y}_H) \end{pmatrix}$$

où $v(\hat{y}_h)$ est la variance de l'estimateur d'HT dans la h^{ème} strate. Nous retenons les estimations par la méthode analytique.

Donc sous H_0, la statistique de test définie comme $T = \Delta' \hat{\Sigma}^{-1} \Delta$, suit une loi du χ^2 à *H*-1 degrés de liberté (9 pour le test sur les 10 strates urbaines de province et 19 pour les strates rurales).

Tableau 7 :
Résultats des tests d'homogénéité

	T	p-value
Strates urbaines hors Yaoundé et Douala	14,4	0,108
Strates rurales	18,6	0,029

Sources :Voir Tableau 5.

Lecture : La chance de rejeter H_0 à tort est de 2.9 %, pour les strates rurales. Donc H_0 est rejetée au seuil de 5 %.

L'homogénéité des niveaux de vie entre les strates rurales est rejetée. Par ailleurs, les strates rurales sont homogènes en intra puisque le coefficient de variation y est faible. La stratification rurale par province est donc opportune. Les niveaux de vie des différentes strates urbaines sont davantage homogènes. Un nombre inférieur de strates urbaines aurait donc pu être retenu ou alors, une stratification combinée avec une autre variable (ce qui peut être fait par une post-stratification pour l'ECAM 2).

Nous pouvons conclure fort des résultats ci-dessus que si la stratification, géographique est pertinente par rapport à l'objet des ECAM, il n'en demeure pas moins que le plan d'échantillonnage tel que mis en œuvre est source de sous-optimalité. Dans sa conception, il n'est pas tenu compte du fait que le sondage se fait à plusieurs degrés et les effets grappe sont négligés.

De ce fait, le nombre moyen de ménages tirés par ZD reste élevé compte tenu des effets grappe en zone rurale, une augmentation du nombre de ZD permet alors de rendre les estimateurs plus précis.

Conclusion

En l'absence d'analyse de précision, il est difficile de juger de la qualité des données issues d'enquêtes par sondage. De ce fait, une perte d'information affecte la mise en œuvre des enquêtes suivantes. Cet article remédie à ces lacunes dans le cas des enquêtes camerounaises auprès des ménages ECAM 2 et 3 et permet ainsi d'établir des recommandations pour une prochaine édition de ce type d'enquête.

Recourant à différentes méthodes d'estimation de variance (linéarisation et techniques de réplications), nous avons montré que la précision dans l'ECAM 2 est très satisfaisante, selon des critères établis par l'Office de Statistique du Canada. Ces résultats nous ont permis de juger de l'optimalité des plans de sondage des enquêtes ECAM 2 et ECAM 3 et d'établir des recommandations visant à optimiser le sondage des enquêtes futures. La taille optimale de l'échantillon et le nombre optimal d'unités primaires à tirer sont calculés, en respectant les contraintes budgétaires. La précision peut être maintenue, voire augmentée, à un coût inférieur. Une mise en garde sur la stratification du milieu urbain est faite.

Cet exercice doit être réitéré dans les pays d'Afrique subsaharienne et devenir systématique. Il ne nécessite pas d'investissement informatique puisque tous les résultats présentés ici ont été obtenus à l'aide d'un logiciel libre, R. De plus le programme utilisé pour cette analyse est inclus dans l'article. Par la suite, cet effort d'estimation de précision devra s'inscrire dans un effort plus général de contrôle de qualité (défauts de couverture, non-réponse, erreurs de mesure,…).

Références Bibliographiques

Banque mondiale (2001), *Combattre la pauvreté*, Rapport sur le Développement dans le monde 2000/2001, Edition Eska pour la Banque mondiale, Paris.

Ardilly,P. (1994), *Les techniques de sondage*. Éditions Technip.

Binder, David A. (1983), « On the variances of asymptotically normal estimators from complex surveys », International Statistical Review, 51: 279–292.

Binder, David A. et al.(1993), « Estimating some measures of income inequality from survey data : An application of estimating equation approach », Proceedings of the ASA Survey Research Methods, 550-555.

Chambers, J.(1998), *Programming with data : A Guide to the S Language*. Springer.

Deville, J.C. (1998), «Estimation de variance pour des statistiques complexes: technique des résidus et de linéarisation», Document de travail INSEE, série Méthodologie Statistique, n°9802.

Direction de la Méthodologie Statistique du Canada (2006), « Lignes directrices concernant la qualité des données », Document de travail 12-539-XIF, Ministère de l'industrie, 2006.

European Communities (2002), « Variance estimation methods in the European union», Monographs of official statistics, European Communities, 2002.

Lumley, T. (2004), « Analysis of complex survey samples » Journal of Statistical Software, 9(1): 1-19,R package verson 2.2.

Lumley,T. (2006), « Survey : analysis of complex survey samples », R package version 3.6-5.

Nguetse, P. J. (2007), « Allocation optimale sous contraintes : cas de la troisième enquête camerounaise auprès des ménages », *STATECO* n°101.

R Development Core Team (2006), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. http ://www.R-project.org.

Wolter, K. (2003), *Introduction to Variance Estimation*. Springer-Verlag, New-York.

Annexe

Le programme des auteurs

Le programme utilisé pour l'estimation de la précision dans l'enquête ECAM2 est le suivant :

```
############################## © Justin BEM 2007 ##########################
# Calcul de précision                                                    #.
# Auteur : Justin BEM                                                    #.
# Date:   Avril 2007                                                     #.
##########################################################################
# Initialisation #
options(digits=3)
setwd("e:/data")
library(survey)   # Package de Thomas LUMLEY pour le calcul de précision
library(foreign)   # Package de lecture des données de SPSS, SAS, Stata,...

## Lecture de la base de données au format SPSS
EC2<-read.spss('EC2MEN.sav',to.data.frame=T)

### Implementation du plan de sondage (Stratification ; 2 ou 3 degrés)
options(survey.lonely.psu="certainty") # Gestion UPS uniques,

EC2deg3<-svydesign (ids= ~ARROND+ZD+MEN, strata=~STRATE, fpc=~FPC1+FPC2+FPC3,
weights=~COEFEXT, data=EC2)

EC2deg2<-svydesign (ids = ~ZD+MEN, strata=~STRATE, fpc=~FPC22+FPC3, weights=
~COEFEXT,data=EC2)

### Techniques de réplication.
JackEC2<-as.svrepdesign(EC2deg2)
BootEC2<-as.svrepdesign(EC2deg2,type="bootstrap",replicates=1000)

tab01<-svyby(~DEPTOT,~S00Q1 , EC2deg3, svymean, vartype=c("se","cvpct"),deff=T)
tab02<-svyby(~DEPTOT,~MILIEUA, EC2deg2, svymean, vartype=c("se","cvpct"),deff=T)
tab03<-svyby(~DEPTOT,~S00Q1 , JackEC2, svymean, vartype=c("se","cvpct"),deff=T)
tab04<-svyby(~DEPTOT,~MILIEUA, BootEC2, svymean, vartype=c("se","cvpct"),deff=T)


#Linéarisation de l'indice GINI(D'après Binder 1993)

gini<-function(x,w){                        # calcul de l'indice de Gini
    if (missing(w))
        w<-rep(1,length(x))
    else if (length(x)!=length(w))
         stop("'x' et 'w' doivent avoir même dimension")

    id_ord<-order(x)                        # Pour ordonner les données
    x1<-x[id_ord]
    w1<-w[id_ord]
    mu<-weighted.mean(x,w)
    term<-(w1/sum(w1))*x1*(cumsum(w1)/sum(w1))
    rval<-(2/mu)*sum(term)-1
    rval
}


#Calcul de u*  NB le fichier doit-être trié selon l'ordre + de la var
d'intérêt

B<-function(x,w){
    id_ord<-order(x,decreasing=T)
```

```
    x1<-x[id_ord]
    w1<-w[id_ord]
    rval<-cumsum(x1*w1)/sum(w1)
    rval[order(rval,decreasing=T) ]
}

u_star<-function(x,w) {
    rval<-((cumsum(w)/sum(w)) -(gini(x,w)+1)/2)*x
    rval<-rval+B(x,w)
    rval<-rval-(weighted.mean(x,w)*(gini(x,w)+1))/2
    rval<-2*rval/weighted.mean(x,w)
}

#ajout de u* comme variable du plan de sondage
EC2deg2<-update(EC2deg2,u=u_star(u_star(DEPTOT,COEFEXT)))

## Calcul de précision par linéarisation de l'Indice de Gini
tab05<-svyby(~u,~S00Q1    ,EC2deg2, svymean,vartype=c("se","cvpct"),deff=T)
tab06<-svyby(~u,~MILIEUA  ,EC2deg2, svymean,vartype=c("se","cvpct"),deff=T)
tab07<-svymean(~u,EC2deg2, vartype=c("se","cvpct"),deff=T)


## Calcul de précision par Jacknife et bootstrap de l'indice de Gini
tab08<-withReplicates(JackEC2,function(w,data) gini(data$DEPTOT,w))
tab081<-withReplicates(subset(JackEC2,S00Q1=="YAOUNDE"),function(w,data) gini(data$DEPTOT,w))
tab082<-withReplicates(subset(JackEC2,S00Q1=="DOUALA"),function(w,data) gini(data$DEPTOT,w))
tab083<-withReplicates(subset(JackEC2,MILIEUA=="URBAIN"),function(w,data) gini(data$DEPTOT,w))
tab084<-withReplicates(subset(JackEC2,MILIEUA=="RURAL"),function(w,data) gini(data$DEPTOT,w))

tab09<-withReplicates(BootEC2,function(w,data) gini(data$DEPTOT,w))
tab091<-withReplicates(subset(BootEC2,S00Q1=="YAOUNDE"),function(w,data) gini(data$DEPTOT,w))
tab092<-withReplicates(subset(BootEC2,S00Q1=="DOUALA"),function(w,data) gini(data$DEPTOT,w))
tab093<-withReplicates(subset(BootEC2,MILIEUA=="URBAIN"),function(w,data) gini(data$DEPTOT,w))
tab094<-withReplicates(subset(BootEC2,MILIEUA=="RURAL"),function(w,data) gini(data$DEPTOT,w))
```