

**PRISE EN COMPTE D'UN PLAN DE  
SONDAGE COMPLEXE DANS LE  
CALCUL DE LA PRECISION DES  
ESTIMATEURS DES INDICATEURS :**  
*LE CAS DES ESTIMATEURS DES CONDITIONS DE VIE ET DU  
PROFIL DE PAUVRETE AU CAMEROUN EN 2001 DANS LA  
DEUXIEME ENQUETE CAMEROUNAISE AUPRES DES  
MENAGES DE 2001 (ECAM II)*

Achille Pégoué

Septembre 2006

---

# Résumé

---

La pratique statistique à l'Institut National de la Statistique est à la réalisation des enquêtes dans le cadre de la mesure des diverses actions du Gouvernement, des opérateurs économiques et tous les acteurs de l'activité sociale, économique, culturelle ou politique au Cameroun. La masse d'information collectée alimente les statistiques de synthèses telles que la comptabilité nationale, mais aussi les politiques dans le processus de prise de décision avisée. Dans ce cadre, l'ECAM II a contribué à l'établissement d'une stratégie de réduction de la pauvreté en déterminant le seuil de pauvreté et en précisant les conditions de vie des pauvres au Cameroun. Le plan de sondage mis en oeuvre demeure cependant incomplet car sa phase ultime qui est le calcul de la précision des estimateurs des indicateurs recherchés n'a pas été effectuée.

En se situant dans le prolongement des travaux d'ECAM II, le présent travail propose des méthodologies d'approximation et d'estimation de la variance des estimateurs complexes dans le cadre d'un sondage stratifié combiné à un tirage à deux ou trois degrés. Les techniques utilisées pour atteindre cet objectif proviennent soit des approximations par linéarisation des estimateurs soit des méthodes de réplication telles que le Jackknife ou le bootstrap. De plus, une exploitation des techniques de traitement des données manquantes s'appuyant sur la simulation des chaînes de Markov par la méthode de Monte Carlo a permis de tirer profit des informations disponibles dans la base de données pour améliorer la précision des estimateurs.

Les résultats obtenus sont probants. D'abord, ils confirment la théorie en ressortant la forte contribution des unités primaires et dans une moindre mesure des unités secondaires dans l'estimation de la variance des estimateurs calculés. Ainsi, l'accroissement du nombre de ville en réduisant la variance des unités primaires améliorerait la précision des indicateurs. Ensuite, ils permettent d'estimer les effets de grappe et donc d'améliorer le calibrage des futures enquêtes. Enfin, ils montrent que la taille de l'échantillon tiré dans ECAM II peut être réduite sans affecter la précision des estimateurs des indicateurs calculés.

---

# Abstract

---

Statistical practice at National Institute of Statistics (NIS) is to carry out survey within the framework of measurement of actions taken by Government, businessmen and others actors of social, economic, cultural or politic activities in Cameroon. The huge quantity of data collected feeds synthesis statistics such as national accounts and moreover the politics in wise process of decision. In this framework, the second Cameroon Households Surveys (CHS) has contributed to establish a strategy for alleviating poverty by bringing out the poverty threshold and the living conditions of poor in Cameroon. However, the survey Design was not fully implemented for the precision of the estimators were neither determined nor computed.

Following the works did during the CHS II, this work aims at settle down an approximating methodology and estimation of standard deviation of complexes estimators of indicators into a stratified sampling with two or three steps of drawing. The techniques use to complete this aim come from analytic approximation such as linearization or from replication method such Jackknife. Moreover, we use the entire information from the data by the data missing treatment with Monte Carlo Markov chain (MCMC) to improve the precision of the estimators.

The results are evidence enough. First of all, they fit with theory by showing a great contribution of primary units and with less importance the secondary units into the whole standard deviation. Hence, increasing the number of towns will reduce the standard deviation of primary units and improve the precision of indicators. Next, the results lead to the sampling effect which can help in the following surveys by estimating an appropriate sampling size. Last, they show that the sample size of CHS II can be reducing without deteriorating the standard deviation of the estimators of indicator computed.

---

# Table des matières

---

<b>Résumé</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Introduction</b>	<b>2</b>
<b>1 Cadre Théorique</b>	<b>3</b>
1.1 Calcul de la précision des indicateurs dans le cadre des enquêtes auprès des ménages : le cas de l'EDSC III . . . . .	3
1.2 Présentation des indicateurs . . . . .	4
1.2.1 Indicateurs calculés dans ECAM II . . . . .	5
1.2.2 Autres Indicateurs . . . . .	6
1.3 Estimation de la précision dans les plans de sondages classiques . . . . .	8
1.3.1 Généralités . . . . .	8
1.3.2 Méthode analytique . . . . .	10
1.4 Estimation de la précision des estimateurs complexes par linéarisation	12
1.5 La méthode de réplification par le Jackknife . . . . .	15
1.6 traitement des données manquantes . . . . .	16
1.6.1 Méthode de repondération . . . . .	16
1.6.2 Imputation multiple par MCMC ou Data Augmentation . . . . .	17
<b>2 Données et Méthodologie</b>	<b>18</b>
2.1 Données et schéma de tirage . . . . .	18
2.1.1 Description des données . . . . .	18
2.1.2 Description du schéma de tirage et notations . . . . .	20
2.2 Méthodologie . . . . .	22
2.2.1 Forme analytique de la variance dans le plan de sondage d'ECAM II . . . . .	22
2.2.2 Méthode de réplification par le Jackknife . . . . .	28
2.2.3 Estimation de la variance avec prise en compte des données manquantes . . . . .	28
2.3 Autres indicateurs dérivées du calcul de précision des estimateurs . . . . .	30
2.3.1 Coefficient de variation ou CV . . . . .	30
2.3.2 Effet de sondage . . . . .	31
2.3.3 Effet de grappe . . . . .	31

---

2.3.4	Intervalle de confiance . . . . .	31
<b>3</b>	<b>Résultats et interprétation</b>	<b>32</b>
3.1	Capital humain . . . . .	32
3.2	Pauvreté monétaire . . . . .	33
3.3	Vulnérabilité . . . . .	33
3.4	Bonne gouvernance . . . . .	34
	<b>Conclusion</b>	<b>35</b>
	<b>Annexes</b>	<b>45</b>
	<b>Bibliographie</b>	<b>46</b>

---

# INTRODUCTION

---

La deuxième Enquête Camerounaise Auprès des Ménages (ECAM II) est une enquête sur les conditions de vie de ménages dont l'objectif majeur est de servir de base au système de suivi et d'évaluation du programme de réduction de la pauvreté du gouvernement camerounais. Elle a été réalisée en 2001 par l'Institut National de la Statistique (INS). L'ECAM II se proposait d'élaborer une méthodologie de calcul d'un indicateur de niveau et d'une ligne de pauvreté, d'étudier les différents aspects de la pauvreté et surtout de produire des analyses au niveau provincial et par type de milieu en isolant les grandes villes que sont Yaoundé et Douala. Un plan de sondage complexe a été élaboré et mis en œuvre pour les phases de collecte et d'exploitation des données. Ce plan de sondage prévoyait le tirage d'un échantillon de 10 000 ménages après stratification. S'agissant de la stratification, Douala et Yaoundé ont été définies comme des strates à part ; chacune des 10 provinces distingue une strate rurale et une strate urbaine. Ainsi, dans l'ensemble, l'enquête a utilisé 22 strates dont 10 rurales et 12 urbaines. Une seconde stratification a été faite dans les strates urbaines en urbain pour les villes de 50 000 habitants au moins et en semi urbain pour les villes de 10 000 à 50 000 habitants. La cartographie du recensement général de la population et de l'habitat de 1987 a été mise et utilisée comme base de sondage. Quant au schéma de tirage adopté, il dépendait du milieu de résidence. Dans les strates urbaines, un tirage à deux degrés a été mise en œuvre. Dans chacun des arrondissements de Yaoundé et Douala, les zones de dénombrements (ZD) ont été tirés à probabilités égales au premier degré, dans chaque zone de dénombrement tirée, 12 ménages ont ensuite été tirés à probabilités égales au second degré. Dans les strates urbaines des provinces, les ZD sont tirées à probabilité égales au premier degré et 18 ménages sont ensuite tirés à probabilités égales au second degré. Dans les strates rurales et les sous strates urbaines, un tirage à trois degrés à été conduit. Au premier degré, on tire les villes (chefs-lieu d'arrondissement) avec une probabilité proportionnelle à leur taille en ménages ; on tire ensuite à probabilités égales, les ZD au deuxième degré et, au troisième degré, 18 ménages pour le milieu semi urbain et 27 ou 36 ménages pour le milieu rural.

S'agissant des conditions de vie des ménages, l'ECAM II se proposait de calculer un indicateur de niveau de vie approché par la consommation finale des ménages et le seuil de pauvreté basé sur l'approche des besoins essentiels. En ce qui concerne le profil de pauvreté, l'ECAM II a distingué un profil de pauvreté monétaire, la pauvreté dans le marché de travail, la pauvreté selon les besoins sociaux de base, la pauvreté en termes de potentialité et de gouvernance et enfin les aspects subjectifs

de la pauvreté. Ces résultats ne permettent ni d'apprécier la qualité des indicateurs calculés, ni de mesurer statistiquement les évolutions observées avec l'enquête budget consommation de 1983, l'ECAM I de 1996 et les autres sources complémentaires d'informations. Par ailleurs, les estimations doivent refléter le plan d'échantillonnage dans la recherche des estimateurs et des intervalles de confiance sans biais. Cette tâche est ardue quand le plan d'échantillonnage est complexe comme c'est le cas dans l'ECAM II. En effet, si la plupart des logiciels statistiques peuvent produire des estimateurs sans biais, le calcul de la variance est un exercice délicat pour les plans d'échantillonnage sortant du cadre du tirage aléatoire simple .

Ce stage se situe dans le prolongement des travaux d'exploitation et d'analyse effectuées sur les données collectées par l'ECAM II. Il procède du souci de fournir aux utilisateurs des résultats sur les conditions de vie des populations et le profil de pauvreté des ménages. Il est aussi le lieu de relever l'importance des plans de sondage dans les enquêtes statistiques en montrant leur influence sur la qualité des résultats et la pertinence des hypothèses de calibration retenues. Les résultats attendus de ce stage sont :

- le calcul d'autres estimateurs que ceux proposés lors de l'exploitation d'ECAM II ;
- l'estimation des données manquantes et leur impact sur le calcul des estimateurs et leur précision ;
- la détermination, dans la mesure du possible, de la forme analytique de la variance des estimateurs de la prévalence du VIH et des facteurs associés ;
- l'estimation de la précision, soit par une méthode analytique, soit par une méthode de réplication.

La méthodologie à mettre en œuvre dans la recherche des formes analytiques des variances des estimateurs s'appuie sur le calcul de la variance d'un estimateur du total d'une population, les techniques de linéarisation des estimateurs non linéaires tels le total d'un domaine, le ratio. L'estimation de cette variance fera recours aux formes approchées des probabilités d'inclusion double afin de procéder à un calcul direct ou l'utilisation des méthodes de réplifications telles que le bootstrap et le jackknife.

# CADRE THÉORIQUE

---

## 1.1 Calcul de la précision des indicateurs dans le cadre des enquêtes auprès des ménages : le cas de l'EDSC III

Plusieurs travaux ont été conduits pour le calcul des estimateurs simples et complexes dans le cadre des enquêtes. Pourtant, au Cameroun, seules les enquêtes démographiques et de santé du Cameroun (EDSC) de l'INS font l'objet du calcul de précision de ces estimateurs. L'EDS III illustre bien cette réalité. La troisième Enquête Démographique et de Santé du Cameroun (EDSC III), réalisée au Cameroun de février à août 2004 par l'Institut National de la Statistique (INS) en collaboration avec le Comité National de Lutte contre le Sida (CNLS), devait dégager des résultats d'une portée provinciale et nationale en prenant en compte la représentation des milieux urbains et ruraux. Un plan de sondage complexe a été élaboré et mis en oeuvre pour les phases de collecte et d'exploitation des données. Ce plan de sondage prévoyait le tirage d'un échantillon de 11 556 ménages. Le tirage de l'échantillon se faisait à deux degrés. Au premier degré, les unités primaires de sondage (UPS) sont sélectionnées à partir des zones de dénombrement (ZD). Ces ZD sont fournies par la cartographie du deuxième recensement général de la population et de l'habitat réalisée de juin 2002 à avril 2003. Ces ZD servent de base de sondage pour un tirage à probabilités inégales de 466 grappes dont 222 rurales et 244 urbaines. Au second degré, un échantillon de ménages est sélectionné dans ces ZD. Les ménages sont tirés avec une probabilité inverse de façon à auto pondérer les domaines. S'agissant des individus du ménage, les femmes résidentes de 15-49 ans et, dans un ménage sur deux, les hommes résidents de 15-59 ans sont éligibles. Dans un document annexe, l'EDSC III a calculé les erreurs de sondages et les effets de sur 57 indicateurs clés de démographie et de santé. Ces indicateurs sont soit des proportions (taux d'alphabétisation, taux d'instruction), soit des moyennes (enfants nés vivants, nombre d'enfants idéal), soit des ratios (ratio de mortalité maternelle). Les résultats ont été pour le Cameroun dans son ensemble, pour les deux grandes villes Douala et Yaoundé ensembles, pour les autres villes, pour l'ensemble du milieu urbain et le milieu rural, et pour chacun des 12 domaines d'étude. Le module " erreurs de sondage " du logiciel ISSA a été utilisé pour calculer les erreurs de sondage suivant la méthodologie statistique appropriée. Ce module utilise la méthode de linéarisation (Taylor) pour

des estimations telles que les moyennes ou proportions, et la méthode de Jackknife pour des estimations plus complexes tels que l'indice synthétique de fécondité et les quotients de mortalité. Quelques limites peuvent être relevées sur le calcul de la précision des indicateurs de l'EDSC III. La première limite est la non prise en compte des erreurs de mesure. En effet, les variables d'intérêt dont la précision a été calculée ont été considérées comme des observations dont la mesure n'est entachée d'aucune erreur, l'erreur provenant uniquement des fluctuations d'échantillonnage. or, les ménages interrogés ne connaissent pas toujours précisément la réponse à la question posée ou ne souhaitent pas donner l'information exacte. La deuxième limite est la non allusion au traitement des non-réponses. En effet, les ménages peuvent refuser soit de se soumettre à l'interview (non réponse totale), soit de répondre à certaines questions (non-réponse partielle). La prise en compte des erreurs d'observation, qui transforme une variable déterministe en une variable stochastique, modifie le biais et la variance de l'estimateur.

## 1.2 Présentation des indicateurs

D'une façon générale, un indicateur est une fonction des variables qui prend une valeur fixe sur l'ensemble de la population  $U$ . Un échantillon  $S$  est un sous-ensemble de la population sur lequel des réalisations des variables des variables entrant dans le champ de l'étude sont mesurées. A partir de ces réalisations, une valeur où estimation de l'indicateur est proposée. Cette estimation est construite à partir d'une fonction des observations ou estimateur et des probabilités d'inclusion  $\pi_k$  où  $k$  désigne une observation. Comme le proposent Deville (1998) et Tillé, ces probabilités d'inclusion peuvent intégrer des variables auxiliaires corrélées à la variable d'intérêt; dans ce cas, en notant  $n$  la taille de l'échantillon  $S$  et  $X$  une variable auxiliaire, on a

$$\pi_k = \frac{nX_k}{\sum_{k' \in U} X_{k'}}.$$

Ainsi, le calcul d'une statistique sur la population  $U$  consiste à mettre un poids égal à 1 sur chaque individu de  $U$ . Un estimateur, proposé par Horvitz et Thomson, de cette même statistique sur l'échantillon consiste à affecter un poids  $w_k = \frac{1}{\pi_k}$  à chaque individu  $k$  de l'échantillon et un poids nul aux autres individus. Soit donc  $M$  la mesure qui met un poids égal à 1 sur chaque individu de  $U$ ,  $\hat{M}$  la mesure qui associé au sondage (affecte un poids  $w_k = \frac{1}{\pi_k}$  à chaque individu  $k$  de l'échantillon et un poids nul aux autres individus), une statistique  $T$  sera notée  $T(M)$  ou  $T$  sur  $U$  et son estimateur sur  $S$  sera noté  $T(\hat{M})$  ou  $\hat{T}$ . A titre d'exemple, si  $T$  désigne le total d'une variable  $Y$ , alors, on a

$$T = T(M) = \sum_{k \in U} Y_k$$

et

$$\hat{T} = T(\hat{M}) = \sum_{k \in S} \frac{Y_k}{\pi_k} = \sum_{k \in S} w_k Y_k.$$

Dans le cas d'un ratio  $R$  de deux totaux  $Y$  et  $X$ , on a

$$R = R(M) = \frac{Y(M)}{X(M)} = \frac{\sum_{k \in U} Y_k}{\sum_{k \in U} X_k}$$

et

$$\hat{R} = R(\hat{M}) = \frac{Y(\hat{M})}{X(\hat{M})} = \frac{\sum_{k \in U} w_k Y_k}{\sum_{k \in U} w_k X_k}.$$

### 1.2.1 Indicateurs calculés dans ECAM II

Plus d'une centaine d'indicateurs ont été calculés dans ECAM II soit comme des statistiques de totaux, soit comme des statistiques de ratios. L'indice de Gini et le taux de pauvreté sortent de ce lot et méritent une attention particulière.

#### Indice de Gini

Dubois (1998) affirme que, dans la pratique, l'indicateur le plus fréquemment utilisé est le coefficient de Gini. Il traduit l'écart entre une distribution hypothétique uniforme des revenus et la distribution effectivement ajustée sur les données recueillies. Il va de 0, pour l'égalité absolue, lorsque chaque individu ou ménage reçoit une part identique du revenu, à 100, lorsqu'une seule personne ou un seul ménage reçoit la totalité du revenu. Ainsi, plus l'indice de GINI est petit, plus la distribution des revenus est égalitaire dans la population. Le coefficient de Gini est fréquemment calculé à partir de la distribution de la consommation des ménage mêmes s'il tend à être sous-estimé par rapport à une distribution du revenu.

L'indice de  $GINI(G)$  s'écrit

$$G(M) = \frac{\sum_{k' \in U} (2r(k') - 1)y'_k}{N \sum_{k' \in U} y'_k} - 1,$$

où  $r(k')$  est le rang de l'individu  $k'$  dans la distribution des  $Y$  (triés par ordre croissant) et peut s'écrire

$$r(k') = \sum_{k'' \in U} 1_{y_{k''} \leq y_{k'}}.$$

Un estimateur de l'indice de Gini est donné par

$$G(\hat{M}) = \frac{\sum_{k' \in S} (2\hat{r}(k') - 1)w_{k'}y_{k'}}{\sum_{k' \in S} w_{k'} \sum_{k' \in S} w_{k'}y_{k'}} - 1, \quad (1.1)$$

où

$$r(k') = \sum_{k'' \in U} w_{k''} 1_{y_{k''} \leq y_{k'}}.$$

### Indicateur du taux de pauvreté (*poverty Headcount*)

Deux situations se distinguent, suivant que le seuil de pauvreté est connu de façon exogène (par exemple par une enquête suffisamment grande pour négliger la variance de l'estimateur) ou estimée à partir de l'enquête. Par ailleurs, le seuil peut être évalué sur une population plus large que le champ de l'étude : ainsi le taux de pauvreté des personnes âgées est calculé à partir du seuil de pauvreté définie sur la France entière.

En considérant le seuil de pauvreté exogène, le taux de pauvreté de la population s'écrit alors :

$$J(M) = F_A(M, s)$$

$F_A$  Est la fonction de répartition de  $Y$  sur la population  $A$  considérée :

$$F_A(Y) = \frac{1}{N_A} \sum_{K \in A} \mathbf{1}_{Y_K \leq Y}$$

$F_A$  Peut également être considérée comme un ratio sur  $U$  et s'estime simplement par

$$J(\hat{M}) = F_A(M, s) = \frac{1}{\sum_{k \in S_A} w_{k'}} \sum_{k' \in S} w_k \mathbf{1}_{Y_K \leq Y} \quad (1.2)$$

### 1.2.2 Autres Indicateurs

Ces indicateurs sont proposés par Dell et al, dans le cadre de l'Enquête Revenus Fiscaux en France. Il s'agit de l'indicateur d'Atkinson

#### Indicateur d'Atkinson

Tout comme le coefficient de GINI, le coefficient d'Atkinson est un indicateur d'inégalité dans la distribution du revenu. De même, plus l'indice d'Atkinson est petit, plus la distribution des revenus est égalitaire dans la population le coefficient d'Atkinson est le coût de l'inégalité. Dans l'expression de cet indicateur,  $1 - a$  est un paramètre de l'aversion à l'inégalité.

Pour  $a \neq 0$ , l'indicateur d'Atkinson  $A_a(M)$  est donné par

$$A_a(M) = 1 - \left( \frac{1}{N} \sum_{k' \in U} \left( \frac{Y_{k'}}{\bar{Y}} \right)^a \right)^{\frac{1}{a}}$$

qui peut encore s'écrire

$$1 - \frac{1}{\bar{Y}^a} \left( \frac{1}{N} \sum_{k' \in U} Y_{k'}^a \right)^{\frac{1}{a}}$$

un estimateur de cet indicateur est

$$A_a(\hat{M}) = 1 - \left( \frac{1}{\hat{N}} \sum_{k' \in S} \left( \frac{Y_{k'}}{\hat{Y}} \right)^a \right)^{\frac{1}{a}}. \quad (1.3)$$

Par prolongement par continuité de  $A_a(M)$  en  $a = 0$ , l'indicateur d'Atkinson devient

$$A_0(M) = 1 - \left( \frac{\prod_{k' \in U} Y_{k'}}{\bar{Y}} \right)^{\frac{1}{N}}$$

et un estimateur de cet indicateur est

$$A_a(\hat{M}) = 1 - \frac{1}{\hat{Y}} \left( \prod_{k' \in U} Y_{k'}^{w_{k'}} \right)^{\frac{1}{a}}. \quad (1.4)$$

### Vers la construction d'un indicateur multidimensionnel

Les indicateurs unidimensionnels présentent une seule des nombreuses facettes de la pauvreté. Un indicateur multidimensionnel a le mérite d'intégrer simultanément plusieurs aspects du même phénomène pour la caractérisation de la pauvreté. Cependant, la mise en commun des indicateurs de pauvreté soulève quelques interrogations notamment sur l'additivité et la corrélation des indicateurs. En effet, des études empiriques sur la France, le Royaume-Uni, l'Espagne et le Portugal, quelques pays issus de l'Europe dite naguère " de l'Est " (Pologne, Roumanie, la Russie), le Brésil et Madagascar montrent que les coefficients de corrélation de Pearson entre les différentes formes de pauvreté notamment la pauvreté par les conditions de vie et la pauvreté monétaire, la pauvreté par les conditions de vie la pauvreté subjective et enfin la pauvreté monétaire et la pauvreté subjective sont inférieurs à 0,3. Ces valeurs établissent que les ménages pauvres selon l'une des formes de pauvreté, le sont aussi selon les autres formes mais pas nécessairement dans la même ampleur. Verger *et al* (2001) propose un indicateur multidimensionnel de pauvreté basé sur le cumul des symptômes de pauvreté (aucun symptôme de pauvreté, un symptôme et un seulement, deux symptômes et deux seulement, trois symptômes). Il calcule ensuite les corrélations entre les différents aspects de la pauvreté à l'aide des scores. Une étude menée par Borel *et al* (2006) s'appuie sur l'analyse factorielle multiple (AFM) pour construire un indicateur composite de pauvreté (ICP) des conditions de vie au niveau du ménage. Les quatre thèmes retenus dans l'analyse de la pauvreté sont l'accès aux infrastructures publiques de base les plus proches, les conditions d'existence, le capital humain et la vulnérabilité. Une analyse de correspondance multiple permet de sélectionner les variables et modalités discriminantes par domaine. Ainsi, le temps moyen d'accès permet d'approcher l'accès aux infrastructures publiques; les sources d'énergie, le logement et ses attributs, l'évacuation des ordures ménagères appréhendent les conditions d'existence. L'alphabetisation et le niveau d'instruction évaluent le capital humain; et la possession de matériels durables estime la vulnérabilité. En définitive, les dix sept variables et quarante cinq modalités retenues sont

utilisées dans l'AFM pour la recherche d'un facteur commun résumant les disparités de conditions de vie entre les ménages. Ces variables doivent vérifier la propriété de cohérence ordinale le long du premier axe (COPA) pour que cet axe caractérise la pauvreté globale des conditions de vie des ménages.

## 1.3 Estimation de la précision dans les plans de sondages classiques

### 1.3.1 Généralités

Quatre types d'erreurs existent dans les terminologies de la théorie des sondages : l'erreur de mesure ou d'observation, le défaut de couverture, la non-réponse et l'erreur d'échantillonnage. L'erreur de mesure provient du fait que l'information collectée peut être différente de la vraie valeur attachée à l'individu. Cette différence trouve son origine dans la délicatesse du sujet, les erreurs de bonne foi de l'enquêté, erreur de remplissage, de codification ou informatique, formulation des questions, etc. En notant  $y_i$  la vraie valeur et  $y_i^*$  la valeur collectée, on modélise l'erreur d'observation par  $\epsilon_i$  en posant  $y_i = y_i^* + \epsilon_i$ .

Le défaut de couverture, quant à lui, est une situation où la base de sondage est incomplète.

La non-réponse peut être partielle ou totale. La non-réponse totale est le cas où un ménage décide de ne pas participer à l'interview. Pour tenir compte de la non-réponse totale, on considère le plus souvent que la décision de répondre est aléatoire, auquel cas l'échantillon des répondants est obtenu par un tirage en deux phases : une première phase d'échantillonnage et une deuxième d'acceptation de l'enquête. La deuxième phase est généralement modélisée par un tirage poissonnien : autrement dit, les unités sont supposées décider de répondre indépendamment les unes des autres. La non-réponse partielle est une situation où un ménage ne sait pas ou refuse de répondre à une question donnée. La valeur imputée est entachée d'une erreur qui peut être traitée comme une erreur d'observation.

L'erreur d'échantillonnage survient du fait que l'échantillon tiré ne constitue qu'un élément dans un ensemble plus grand d'échantillons probables. La théorie des sondages considère que c'est cet échantillon, ou mieux sa composition qui est aléatoire. L'effet de l'erreur d'échantillonnage sur la précision d'un estimateur est mesurée par le biais, la variance ou précision et l'erreur quadratique moyen.

#### – Calcul du biais d'un estimateur

Le biais est proche d'une caractéristique de tendance centrale et traduit l'écart moyen entre l'estimateur et la vraie valeur inconnue. Deux cas de figure sont envisagés dans le calcul du biais : "

- Le cas où le biais est nul à la suite d'un calcul théorique "
- Le cas où le biais non nul, exprimé par des formules littérales est approché numériquement par des ordres de grandeur.

#### – Calcul de la variance d'un estimateur

La variance est une mesure de l'erreur d'échantillonnage où tout écart à la moyenne contribue positivement à l'évaluation de l'imprécision. La formule de

la variance est donnée par la moyenne des carrés des écarts de l'estimateur à sa moyenne, soit

$$V(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2$$

où  $\hat{\theta}$  est un estimateur de la statistique  $\theta$  et  $E(\cdot)$  est l'espérance mathématique. La racine carrée de la variance, ou écart-type, est souvent préférée à la variance dans le but de se ramener à la même échelle que l'estimateur et permettre le calcul des intervalles de confiance.

- L'écart quadratique moyen est un indicateur synthétique de la précision qui englobe le biais et la variance. Il représente la moyenne des carrés de l'estimateur à la vraie valeur.
- Effet de grappe L'ECAM II a utilisé un plan de tirage à plusieurs degrés où les unités primaires sont soit les villes, les unités secondaires les ZD, les unités tertiaires les ménages. Dans chaque unité d'un degré de tirage, le risque est grand de rencontrer des ménages similaires. Cette similitude entre les individus vis-à-vis de la variable d'intérêt dans une unité réduit la précision des estimateurs et s'appelle effet de grappe. L'effet de grappe est mesuré par le coefficient de corrélation intra grappe et est donné, dans le cas d'un SAS, par la formule

$$\rho = \frac{\sum_{i=1}^{NB_{men}} \sum_{j=1 \text{ et } j \neq i}^{NB_{men}} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{NB_{men}} \sum_{j=1}^{NB_{men}} (Y_i - \bar{Y})(Y_j - \bar{Y})} \times \frac{1}{\bar{N} - 1}$$

où

$NB_{men}$  le nombre de ménages,

$\bar{N}$  le nombre total de ménages tirés et  $\bar{Y}$  la moyenne générale de la variable d'intérêt  $Y$ .

Cette formule de l'effet de grappe montre que si une forte similarité entre les ménages existe, les  $(Y_i - \bar{Y})(Y_j - \bar{Y})$  sont positifs et l'effet de grappe est positif.

En réécrivant la formule de la variance, une relation entre l'effet de grappe et la précision peut être exhibée. Dans le cas de l'estimation d'un total  $T$  par un tirage à deux degrés avec tirage aléatoire simple à chaque degré où toutes les unités primaires ont une même taille  $\bar{N}$ , le taux de sondage est négligeable devant 1 et les unités secondaires sont de taille constante, la variance s'écrit

$$V(\bar{T}) = N^2 \frac{S^2}{m\bar{n}(1 + \rho(\bar{n} - 1))}$$

où

$m$  est le nombre d'unités primaires tirées et  $\bar{n}$  est le nombre d'unités secondaires tirées par unité primaire.

Cette expression montre les rôles de  $\rho$ ,  $m$  et  $\bar{n}$ . Un effet de grappe positif accroît la variance. Plus le nombre d'unités primaires tirées est grand, plus la variance est petite. La dérivé de l'expression de la variance par rapport à  $\bar{n}$  conduit à

$$\frac{dV}{d\bar{n}} = N^2 S^2 \frac{\rho - 1}{\bar{n}^2}$$

, ce qui montre l'effet du nombre d'unités secondaires tirées  $\bar{n}$  dépend du rapport entre l'effet de grappe et l'unité. Si  $\rho$  est plus grand que 1, tout accroissement du nombre d'unités secondaires dégrade la précision.

Deux méthodes d'estimation de la précision des indicateurs sont mises en exergue dans la littérature. La première méthode est la recherche de la forme analytique de la variance et d'un estimateur de cette expression analytique et la seconde est l'utilisation des méthodes de réplication.

### 1.3.2 Méthode analytique

A partir de l'expression de l'estimateur, la méthode analytique consiste à rechercher par un développement mathématique une expression de la variance. Les deux grandes difficultés qu'elles rencontrent sont l'existence des probabilités d'inclusion doubles ou triples pour l'estimateur d'un total et la présence des estimateurs complexes tel un ratio. La solution apportée à l'existence des probabilités doubles est l'approximation et la solution apportée au problème des estimateurs complexes est la linéarisation.

**Estimation de la précision d'un total** Soient  $U$  une population composée de  $N$  individus,  $Y$  une variable qui prend la valeur  $Y_k$  pour l'individu  $k$ ,  $S$  un échantillon de  $n$  individus sur lesquels les  $Y_k$  sont observés,  $\pi_k$  les probabilités d'inclusion simple et  $\pi_{kl}$  les probabilités d'inclusions doubles. Le total de la variable  $Y$  sur la population s'écrit

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

et est estimé sans biais sur l'échantillon par l'estimateur de Horvitz-Thomson

$$\hat{T} = \sum_{k \in U} \frac{Y_k}{\pi_k}.$$

La variance de cette estimateur est donnée par

$$V(\hat{Y}) = \sum_{k \in U} \frac{Y_k^2}{\pi_k} (1 - \pi_k) + \sum_{k \in S} \sum_{l \in S, l \neq k} \frac{Y_k Y_l}{\pi_k \pi_l} (\pi_{kl} - \pi_k \pi_l). \quad (1.5)$$

Un estimateur sans biais de cette variance est

$$\hat{V}(\hat{Y}) = \sum_{k \in U} \frac{Y_k^2}{\pi_k} (1 - \pi_k) + \sum_{k \in U} \sum_{l \in U, l \neq k} \frac{Y_k Y_l}{\pi_k \pi_l \pi_{kl}} (\pi_{kl} - \pi_k \pi_l). \quad (1.6)$$

Cette variance et son estimateur prennent des formes diverses selon le plan de sondage classique.

#### – Sondage aléatoire simple

Dans le cas d'un sondage aléatoire simple, les probabilités d'inclusion sont

$$\pi_k = \frac{n}{N} \quad \text{et} \quad \pi_{kl} = \frac{n}{N} \frac{n-1}{N-1}$$

le total est estimé par

$$\hat{Y} = \frac{N}{n} \sum_{k \in S} Y_k,$$

et 1.5 devient

$$V(\hat{Y}) = \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{k \in U} (Y_k - \bar{Y})^2; \quad (1.7)$$

1.6 devient

$$\hat{V}(\hat{Y}) = \frac{N(N-n)}{n} \frac{1}{N-1} \sum_{k \in S} (Y_k - \hat{Y})^2; \quad (1.8)$$

Où

$$\bar{Y} = \sum_{k \in U} Y_k \quad \text{et} \quad \hat{Y} = \sum_{k \in S} Y_k.$$

– *Sondage à probabilité inégales*

Dans le cas d'un sondage à probabilités inégales, les probabilités d'inclusion sont difficiles à calculer et l'approximation de Deville n'utilisant que les probabilités d'inclusion est en général utilisé

$$\pi_k = \frac{n}{N}$$

et 1.6 devient

$$\hat{V}(\hat{Y}) = \frac{1}{1 - \sum_{k \in S} a_k^2} \sum_{k \in S} (1 - \pi_k) \left( \frac{Y_k}{\pi_k} - A \right)^2, \quad (1.9)$$

Où

$$a_k = \frac{1 - \pi_k}{\sum_{k' \in S} (1 - \pi_{k'})} \quad \text{et} \quad A = \sum_{k \in S} a_k \frac{Y_k}{\pi_k}.$$

– *Sondage Stratifié*

Dans le cas d'un sondage stratifié où la population est partitionnée en  $H$  strates et dans chaque strate  $h$  un sous-échantillon  $S_h$  est tiré de manière indépendante, le total est estimé par

$$\hat{Y} = \sum_{h=1}^H \sum_{k \in S_h} \frac{Y_k}{\pi_k}$$

et 1.5 devient

$$V(\hat{Y}) = \sum_{h=1}^H V_h \left( \sum_{k \in S_h} \frac{Y_k}{\pi_k} \right). \quad (1.10)$$

1.6 devient

$$\hat{V}(\hat{Y}) = \sum_{h=1}^H \hat{V}_h \left( \sum_{k \in S_h} \frac{Y_k}{\pi_k} \right). \quad (1.11)$$

Où  $V_h \left( \sum_{k \in S_h} \frac{Y_k}{\pi_k} \right)$  est la variance à l'intérieur de la strate  $h$  et peut être calculée selon le plan de tirage des sous-échantillons  $S_h$ .

*Sondage à deux degrés*

Dans le cas d'un sondage à deux degrés où dans un premier temps  $n$  unités primaires sont échantillonnées parmi  $M$  à l'aide d'un plan de sondage  $S_1$  et dans chaque unité primaire  $i \in S_1$  (la probabilité d'inclusion d'un individu  $i$  de  $S_1$  est  $\pi_i$ ) et indépendamment,  $n_i$  unités secondaires sont échantillonnées parmi  $N_i$  à l'aide d'un sondage  $S_{2i}$  (la probabilité d'inclusion d'un individu  $k$  sachant qu'il appartient à l' $UP_i$  est  $\pi_{k|i}$ ); le total  $Y$  est estimé par

$$\hat{Y} = \sum_{i \in S_1} \sum_{k \in S_{2i}} \frac{Y_{ik}}{\pi_i \pi_{k|i}}$$

et 1.5 devient

$$V(\hat{Y}) = V \left( \sum_{i \in S_1} \frac{\sum_{k=1}^{N_i} Y_{ik}}{\pi_k} \right) + \sum_{i=1}^M V \left( \sum_{k \in S_{2i}} \frac{Y_{ik}}{\pi_{k|i}} \middle| S_1 \right); \quad (1.12)$$

1.6 devient

$$\hat{V}(\hat{Y}) = \hat{V} \left( \sum_{i \in S_1} \frac{\sum_{k \in S_{2i}} Y_{ik}}{\pi_i} \right) + \sum_{i \in S_1} \frac{\hat{V} \left( \sum_{k \in S_{2i}} \frac{Y_{ik}}{\pi_{k|i}} \middle| S_1 \right)}{\pi_i}, \quad (1.13)$$

Où (14) est obtenu à l'aide la formule

$$V(\hat{Y}) = V \left( E \left( \hat{Y} \middle| S_1 \right) \right) + E \left( V \left( \hat{Y} \middle| S_1 \right) \right).$$

## 1.4 Estimation de la précision des estimateurs complexes par linéarisation

La détermination de la forme analytique telle que précédemment calculée n'est applicable que pour les estimations des variances de totaux. Pour un estimateur complexe, la formule exacte de l'estimateur de la variance sera inconnue. Deville montre qu'il est possible de se ramener au calcul de type précédent par approximation en procédant à un développement limité d'ordre 1 en introduisant la notion de *fonction d'influence* pour un individu ou de "linéariser l'estimateur".

La fonction d'influence d'une statistique  $T$  associée à un individu  $k$  apprécie l'effet de la variation du poids associé à cet individu sur la statistique à estimer. Soit  $M$ , une mesure qui met un poids égal à 1 sur chaque individu,  $\hat{M}$  la mesure associé au sondage,  $M + t\delta_k$  et la mesure qui met un poids égal à 1 pour tous les individus

sauf l'individu  $k$  qui à un poids égal à  $1 + t$ . Les statistiques associées sont  $T(M)$ ,  $T(\hat{M})$  et  $T(M + t\delta_k)$ . La fonction d'influence est

$$z_k = \lim_{t \rightarrow 0} \frac{T(M + t\delta_k) - T(M)}{t} = \text{lin}_k(T).$$

. Deville montre que la variance du total

$$\sum_{k \in S} \frac{z_k}{\pi_k}$$

est un estimateur de la variance de la statistique  $T(\hat{M})$ . Ainsi, si  $T$  désigne une statistique linéaire de la variable  $X$ , on a

$$T(M) = \sum_{k' \in U} \alpha_{k'} X_{k'}$$

où les  $\alpha_{k'}$  sont des réels,

$$T(M + \delta_k) = \sum_{k' \in U} \alpha_{k'} X_{k'} + t\alpha_k X_k$$

et

$$T(\hat{M}) = \sum_{k' \in S} \frac{\alpha_{k'} X_{k'}}{p_{k'}}$$

d'où

$$z_k = \text{lin}_k = \alpha_k X_k.$$

Il vient alors que

$$\text{sum}_{k \in S} \frac{z_k}{\pi_k} = T(\hat{M})$$

d'où  $\text{sum}_{k \in S} \frac{z_k}{\pi_k}$  et  $T(\hat{M})$  ont même variance. Soit  $T(M)$  et  $S(M)$  deux statistiques, les formules sur les opérations de la fonction d'influence sont semblables à celles des dérivées de fonctions. Ainsi,

$$\text{lin}_k(T + S) = \text{lin}_k(T) + \text{lin}_k(S)$$

$$\text{lin}_k(TS) = S(M)\text{lin}_k(T) + T(S)\text{lin}_k(S)$$

$$\text{lin}_k\left(\frac{T}{S}\right) = \frac{\text{lin}_k(T)}{S(M)} - \frac{T(S)\text{lin}_k(S)}{S(M)^2}$$

$$\text{lin}_k(f(T)) = f'(T(M))\text{lin}_k(T)$$

où  $f$  est une fonction dérivable

Ainsi, pour un ratio  $R(M) = \frac{Y(M)}{X(M)}$  de deux totaux  $X(M)$  et  $Y(M)$ , la fonction d'influence de  $R(M)$  ssociée à l'individu  $k$  s'écrit

$$\text{lin}_k(R) = \frac{1}{X(M)^2} (Y_k - R(M)X_k)$$

Une autre présentation de la linéarisée est le calcul direct des dérivées premières. Ainsi, si  $\theta = f(T_1, \dots, T_j, \dots, T_J)$  est une fonction des totaux  $T_1$  à  $T_J$  où  $T_j = \sum_{k' \in S} z_{jk}$

et  $z_k = z_{1k}z_{jk}, \dots, z_{JK}$  un estimateur  $\theta$  est  $\hat{\theta} = f(\hat{t}_1, \dots, \hat{t}_j, \dots, \hat{t}_J)$

où les  $\hat{t}_j$  sont par exemples des estimateurs de Horvitz et Thompson de  $T_j$  :  $\hat{t}_{j\pi} = \sum_{k \in S} \frac{z_{jk}}{\pi_k}$ . Lorsque  $f$  n'est pas une fonction linéaire, l'estimateur de la variance de  $\hat{\theta}$ , est donnée par  $\hat{V}(\hat{t}_{\pi,u})$  qui est un estimateur sans biais de la variance de l'estimateur du total d'une nouvelle variable  $u_k$ , où

$$\hat{t}_{\pi,u} = \sum_{k \in S} \frac{u_k}{\pi_k} \text{ et } u_k = \sum_{j=1}^J a_j z_{jk}$$

avec

$$a_j = \left. \frac{\partial \theta}{\partial T_j} \right|_{T_j = \hat{t}_j}$$

avec (2) Exemples de linéarisation Le premier exemple proposé par Dell et al est la linéarisation de l'indice de Gini donné par

$$G(M) = \frac{\sum_{k' \in U} (2r(k') - 1)y_{k'}}{N \sum_{k' \in U} y_{k'}} - 1,$$

. La linéarisée du dénominateur associé à un individu  $k$  est

$$\text{lin}_k(\text{deno}) = \text{lin}_k(N) \sum_{k' \in U} y_{k'} + N \text{lin}_k \left( \sum_{k' \in U} y_{k'} \right),$$

où en considérant  $N$  comme une statistique du total de la variable qui prend la valeur 1 pour tous les individus, on obtient

$$\text{lin}_k(N) = 1$$

et on tire

$$\text{lin}_k(\text{deno}) = \sum_{k' \in U} y_{k'} + N y_{k'}$$

La linéarisée du rang  $r(k)$  est  $\text{lin}_k r(k') = 1_{y_k \leq y_{k'}} y_{k'}$ . La linéarisée du numérateur se déduit en utilisant la formule de la linéarisée de la somme de statistiques soit

$$\text{lin}_k(\text{num}) = (2r(k) - 1)y_k + 2 \sum_{k' \in U} y'_k 1_{y_k \leq y_{k'}} y_{k'}.$$

Finalement, la linéarisée de l'indice de GINI est

$$\text{lin}_k(G) = \frac{2 \left( y_k r(k) + \sum_{k' \in U} 1_{y_k \leq y_{k'}} y_{k'} \right) - y_k - (G(M) + 1) \left( \sum_{k' \in U} y_{k'} + N y_k \right)}{N \sum_{k' \in U} y_{k'}} \quad (1.14)$$

Le second exemple proposé par Dell et al est la linéarisation de l'indicateur d'Atkinson donné par

$$A_a(M) = 1 - \left( \frac{1}{N} \sum_{k' \in U} \left( \frac{Y_{k'}}{\bar{Y}} \right)^a \right)^{\frac{1}{a}}$$

lorsque et lorsque  $a = 0$ . et

$$A_0(M) = 1 - \left( \frac{\prod_{k' \in U} Y_{k'}}{\bar{Y}^N} \right)^{\frac{1}{N}}$$

Pour , la linéarisée de cet indicateur est

$$\text{lin}_k(A_a) = \frac{1 - A_a(M)}{N} \left( \frac{Y_k}{\bar{Y}} - 1 - \frac{1}{a} \left( \frac{NY_k^a}{\sum_{k' \in U} Y_{k'}} - 1 \right) \right). \quad (1.15)$$

Et lorsque  $a$  prend la valeur 0 la linéarisée de 1.4 est

$$\text{lin}_k(A_0) = \frac{1 - A_0(M)}{N} \left( \frac{Y_k}{\bar{Y}} - 1 - \log(Y_k) + \frac{\sum_{k' \in U} \log Y_{k'}}{N} \right). \quad (1.16)$$

Enfin, le troisième exemple tiré de Dell et al est la linéarisation de l'estimateur du taux de pauvreté

$$J(\hat{M}) = F_A(M, s) = \frac{1}{\sum_{K \in S_A} W_K} \sum_{K \in U} W_K 1_{Y_K \leq Y}$$

$F_A$  étant un simple ratio sur  $U$ , la linéarisation de  $J(M)$  s'écrit directement :

$$\text{lin}_k(J) = \frac{1_{K \in A}}{N_A} (1_{Y_K \leq S} - J(M)) \quad (1.17)$$

## 1.5 La méthode de réplcation par le Jackknife

la méthode du Jackknife est une technique de ré échantillonnage qui vise à fournir une procédure informatique d'estimation de la variance et du biais d'un estimateur complexe. Cette technique a été proposée pour la première en 1949 par Quenouille et s'appuie sur le retrait de certains individus de la base et le recalcul de l'estimateur. Considérons l'estimateur  $\hat{T} = f(Y_1, Y_2, \dots, Y_n)$  et  $\hat{T}_{(i)} = f(Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n)$  la  $i^{me}$  réplcation du Jackknife obtenue en supprimant la  $i^{me}$  observation; la  $i^{me}$  pseudo valeur peut être estimée par

$$Y_{(i)} = N\hat{t}(n-1)\hat{T}_{(i)} \quad (1.18)$$

et joue le même rôle que  $Y_i$  dans le calcul de l'estimateur. L'estimateur Jackknife de  $T$  est donnée par

$$\hat{T}_{Jack} = \frac{1}{n} \sum_{i=1}^n Y_i = n\hat{T} - (n-1)\hat{T}_{()} \quad (1.19)$$

où  $\hat{T}_{()} = \frac{\sum_{i=1}^n \hat{T}_{(i)}}{n}$ . L'estimateur Jackknife du biais est

$$Biais_{jack} = (n-1)(\hat{T}_{()} - \hat{T}) \quad (1.20)$$

et l'estimateur Jackknife de la variance est

$$var_{jack} = var((\hat{T})_{jack}) = \frac{\sum_{i=1}^n (Y_{(i)} - \hat{T}_{Jack})^2}{n(n-1)} = \frac{n-1}{n} \sum_{i=1}^n ((\hat{T})_{(i)} - \hat{T}_{()})^2. \quad (1.21)$$

Dans notre cas, chaque sous-échantillon exclut une ZD dans les calculs des estimations. Des sous échantillons pseudo-indépendants sont créés. La variance d'un estimateur complexe est donnée par

$$var(T) = \frac{1}{k(k-1)} \sum_{i=1}^k ((\hat{T})_{(i)} - \hat{T})^2. \quad (1.22)$$

où  $k$  est le nombre de villes,  $(\hat{T})_{(i)}$  l'estimateur de Horvitz et Thomson obtenu en excluant la ville  $i$ , et  $\hat{T}$  est l'estimateur de Horvitz et Thomson sur l'ensemble des ZD.

## 1.6 traitement des données manquantes

La méthode de repondération sera utilisées pour le traitement des données manquantes sur les variables qualitatives et l'imputation par MCMC pour le traitement des données manquantes des variables quantitatives

### 1.6.1 Méthode de repondération

La méthode de repondération consiste à modifier les poids initiaux des individus pour tenir compte de la non réponse dans le calcul de l'estimateur. Pour estimer la caractéristique d'intérêt sur la population totale, la méthode de repondération procède d'abord par une inférence du sous-échantillon de répondants sur l'échantillon total en supposant que tous les individus de la population ont une probabilité de répondre non nulle. Elle infère ensuite de l'échantillon total à la population entière en supposant que tous les individus de la population ont une probabilité d'inclusion non nulle. La méthode de repondération suppose l'existence des catégories pour lesquelles les individus ont une probabilité de répondre non nulle et homogène.

### 1.6.2 Imputation multiple par MCMC ou Data Augmentation

La méthode Markov Chain Monte Carlo (MCMC) permet de simuler des tirages pseudo aléatoires à partir d'une distribution de probabilité multidimensionnelle en utilisant les chaînes de Markov. Une chaîne de Markov est une séquence de variables aléatoires dans laquelle la distribution de chaque élément dépend uniquement de l'élément précédent i.e  $P(X_t|X_{t-1}, X_{t-2}, \dots) = P(X_t|X_{t-1})$ . Dans la simulation de MCMC, on construit une chaîne de Markov suffisamment longue pour que la distribution des éléments se stabilise en un processus stationnaire. On utilise pour cela le théorème de Bayes sur la distribution *a posteriori* d'un paramètre inconnu  $\theta$  qui est donnée par

$$P(\theta|Y) = \frac{P(Y|\theta) P(\theta)}{\int P(Y|\theta) P(\theta) d\theta}. \quad (1.23)$$

La simulation MCMC permet alors de simuler la distribution jointe a posteriori des quantités inconnues et de dégager des estimations des paramètres sur la base. En supposant que le tableau des données est tiré d'une distribution normale multivariée, la Data Augmentation consiste à itérer plusieurs fois les étapes *I* et *P* suivantes.

- Etape *I* : Cette étape permet de simuler de façon indépendante les valeurs manquantes pour chaque observation  $i$ . Soient  $\theta$  le paramètre d'intérêt,  $Y_{manq}$  l'ensemble des variables ayant des données manquantes et  $Y_{obs}$  l'ensemble des variables ayant des données observées, à l'itération  $t$ , on tire  $Y_{manq}^{t+1}$  dans la distribution donnée par  $p(Y_{manq}, \theta^t)$  où  $\theta^t$  est la valeur du paramètre d'intérêt calculée avant l'itération  $t$
- Etape *P* : Cette étape simule le paramètre  $\theta$  à partir d'un tableau complet de données (i.e tableau sans valeurs manquantes). A l'itération  $t$ , on tire  $\theta^t$  dans la distribution donnée par  $p(\theta|Y_{obs}, Y_{manq})$ .

La chaîne de Markov obtenue est

$$(Y_{manq}^1, \theta^1), (Y_{manq}^2, \theta^2), \dots$$

Cette chaîne converge en probabilité vers  $p(Y_{manq}, \theta)$ . La stationnarité est atteinte après  $t$  itérations si

- $\theta^t$  est indépendant de  $\theta^0$
- $\theta^{2t}$  est indépendant de  $\theta^t$ , etc.

Le test de stationnarité peut-être graphique ou à travers les fonctions d'auto-corrélation.

# DONNÉES ET MÉTHODOLOGIE

---

## 2.1 Données et schéma de tirage

### 2.1.1 Description des données

Plusieurs outils de collecte ont été utilisés lors de la phase d'exécution d'ECAM II.

- Un questionnaire principal ;
- Un formulaire d'enregistrement des dépenses rétrospectives du ménage ;
- Un formulaire d'enregistrement des dépenses et acquisitions quotidiennes du ménage.

Chaque questionnaire est subdivisé en sections. Les numéros des sections sont consécutifs. Chaque section a été saisie séparément et stockée dans un fichier identifié par le numéro de la section. La saisie des questionnaires a été faite à l'aide d'un masque et des contrôles de saisie dans l'environnement CsPro. L'apurement des données a été faite sous SPSS. L'analyse des données a été faite à l'aide de SPSS et Stata. Les fichiers sont disponibles sous format SPSS.

La gestion des informations sous SPSS se fait par l'intermédiaire

- D' un dictionnaire de variables qui contient les informations sur le nom, leur libellé, le type et la longueur des variables ;
- D'un tableau de données qui présente en ligne les observations et en colonne les variables. En plus des variables correspondant aux questions des outils de collecte, une variable représentant le poids des individus figure dans le tableau des données.

La description des différentes sections est faite comme suit :

- La section00 porte sur les renseignements généraux, à savoir l'identification du ménage, les renseignements sur le ménage et la collecte. Cette section permet d'affecter un identifiant unique au ménage qui est une concaténation du code la région, du numéro de la ZD et du numéro du ménage dans la ZD (21 variables) ;
- La section01 décrit la composition et les caractéristiques des membres du ménage. (10 variables en plus des noms des membres du ménage) ;
- La section02 est relative à la santé des membres (12 variables en plus des noms).
- La section03 porte sur le niveau d'instruction du membres de ménages âgés de plus de 5 ans (12 variables en plus du nom).
- La section04 retrace l'activité des membres du ménage (30 variables).

- La section05 porte sur la natalité, la mortalité et la fécondité au sein du ménage (27 variables) ;
- La section06 est réservée à l’anthropométrie et la couverture vaccinale des enfants de 0 à 35 mois (15 variables).
- La section07 concerne les logements et équipements du ménage (11 variables) ;
- La section08 porte sur les migrations du ménage (11 variables) ;
- La section09 décrit l’accessibilité aux infrastructures de base (7 variables) ;
- La section10 retrace la perception des conditions de vie du ménage (18 variables) ;
- La section11 est consacrée aux entreprises familiales non agricoles (30 variables) ;
- La section12 est relative au patrimoine matériel et financier ainsi qu’à l’épargne et au capital social du ménage (32 variables) ; La section13 porte sur l’agriculture et les activités du monde rural (56 variables) ;
- La section14 retrace les dépenses rétrospectives non alimentaires du ménage (14 sous sections de 4 variables chacune) ;
- La section15 retrace les dépenses et acquisitions quotidiennes des ménages (12 variables).

Les variables retenues dans le cadre de cette étude peuvent être regroupées en quatre thèmes en plus des variables démographiques qui sont la région (s00q1), le numéro de la ZD (s00q2), le numéro du ménage (s00q3), le code du département (s00q4), le code de l’arrondissement (s00q5), et le milieu (s00q7). Ces variables définissent les 32 strates (combinaison de 12 régions et 3 milieux) et les trois degrés de tirage (arrondissement, ZD et ménage). Ce regroupement s’inspire des variables retenues par Borel et al (2006).

Le premier thème est relatif au capital humain. Il regroupe les variables relatives à l’éducation et à la santé. L’éducation est appréhendée par l’alphabétisation (s03q2) et le niveau d’instruction le plus élevé du ménage (s03q10). La santé est caractérisée par la prévalence du paludisme (s02q11a), les dépenses en médicaments (pharma) et dépenses en consultation (consult).

Le second thème est la pauvreté monétaire caractérisée par les dépenses alimentaires (depalim), les dépenses non alimentaires (depnalim) et les dépenses totales (deptot)

Le troisième thème est la vulnérabilité à travers :

- les commodités : de télécommunication (possession d’un téléphone fixe (equip1), d’un téléphone mobile (equip2), d’un poste radio (equip3), d’une télévision (equip14) ou d’une chaîne musicale (equip16)) ; de communication (bicyclette (equip7), motocyclette (equip9), véhicule ( equip13)) ; d’un matériel moderne de cuisson (cuisinière (equip10), réchaud à gaz (equip11) , réchaud à pétrole (equip12), bouteille à gaz (equip17)) ; ou autres bien, de luxe (réfrigérateur (equip4), congélateur (equip5), climatiseur (equip6), ventilateur (equip8), fer à repasser (equip15)) ;
- la possession de terre (s12q1) ;
- la vie associative (s12q26).

Le quatrième thème est la bonne gouvernance mesurée par le paiement des frais non réglementaires pour la scolarisation (s10q15), pour les soins médicaux (s10Q16),

pour autres services (s10q17) ; le paiement volontaire de frais à un agent des forces de l'ordre.

### 2.1.2 Description du schéma de tirage et notations

La population totale est constituée de l'ensemble des ménages résidant sur le territoire camerounais pendant la période de collecte qui va de septembre à décembre 2001. Le ménage est à la fois unité d'observation et unité d'échantillonnage. Le nombre total de ménage ayant été utilisé pour le tirage de l'échantillon est de  $N = 2865265$  et la taille de l'échantillon désiré est de 11553. Deux grandes étapes sont considérées dans la constitution de l'échantillon. La première étape est la définition des strates et la seconde étape est le tirage dans les strates constituées.

#### Phase d'allocation

Dans cette première étape de constitution de l'échantillon, deux niveaux de stratification sont mis en œuvre. A l'issue de cette phase, un nombre de ménages à tirer dans chaque milieu est déterminé.

Le premier niveau est une stratification en 12 régions constituées des 12 provinces et des grandes métropoles que sont Yaoundé et Douala ; une allocation proportionnelle à la taille des régions en ménages est utilisée pour la répartition de l'échantillon entre les 12 régions. Le poids d'une région est  $P_{reg} = N_{reg}/N = n_{reg}/n$  où  $N_{reg}$  est le nombre de ménages dans la région  $reg$  et  $n_{reg}$  le nombre de ménages tirés dans la région  $reg$ .

Le second niveau est la stratification des régions en milieu urbain, semi-urbain et rural. Une première allocation proportionnelle est faite entre les milieux urbain et semi urbain d'une part et le milieu rural d'autre part sur la base d'un poids de  $4/7$  pour les premiers et de  $3/7$  pour le second. Ces pondérations sont tirées de la phase de correction des ZD du recensement de 1997 avant le tirage de l'échantillon où il apparaît que le milieu urbain et semi-urbain est plus homogène que le milieu rural. Une seconde allocation proportionnelle à la taille des ménages est réalisée entre le milieu urbain et le milieu semi urbain. Le milieu urbain est constitué des grandes villes (plus de 50 000 habitants) et le milieu semi urbain est constitué des villes de 10 000 à 50 000 habitants. Les deux grandes métropoles sont entièrement dans le milieu urbain. Le poids d'un milieu urbain ou semi urbain sachant sa région  $reg$  hors de l'une des deux métropoles est  $P_{mil/reg} = (4/7)N_{mil}/N_{reg} = (4/7)n_{mil}/n_{reg}$  où  $n_{mil}$  est le nombre de ménages dans le milieu  $mil$  et  $n_{mil}$  le nombre de ménages tirés dans le milieu  $mil$  ; le poids du milieu rural est  $P_{mil/reg} = 3/7$  ; le poids du milieu urbain sachant que la région  $reg$  est une métropole est  $P_{mil/reg} = 1$ .

En définitive, le poids du milieu  $mil$  est  $P_{mil} = P_{reg}P_{mil/reg}$ .

#### Phase de sélection

Les strates précédentes permettent de délimiter quatre milieux qui sont (i) les grandes métropoles que sont Yaoundé et Douala (ii) les grandes villes des provinces (iii) les petites villes des provinces et (iv) le milieu rural. A chaque milieu, un schéma de tirage tenant compte de ses spécificités a été mis en œuvre.

**Tirage à Yaoundé et Douala** Chaque ville est découpée en quatre arrondissements. Le nombre de zones de dénombrement <sup>1</sup> (ZD) à tirer par ville est fixé à 100 ( $nb_{ZD=100}$ ). La répartition des 100 ZD par arrondissement est faite proportionnellement à l'effectif des ménages de l'arrondissement de 1987. Dans chaque arrondissement, un tirage à deux degrés est conduit. Au premier degré, les ZD sont tirées à probabilités égales dans un arrondissement. Au deuxième degré, 12 ménages sont tirés à probabilités égales dans chaque ZD.

Notations :

$NB_{arr}$  : nombre d'arrondissement ( $NB_{arr} = 4$ );

$nb_{arr}$  : nombre d'arrondissement tirés ( $nb_{arr} = 4$ );

$NB_{ZD}$  : nombre de ZD d'un arrondissement (cartographie de 1987;)

$nb_{ZD}$  : nombre de ZD à tirer d'un arrondissement (issue d'une répartition proportionnellement à l'effectif des ménages de 1987;)

$NB_{men}$  : nombre de ménages d'une ZD (obtenu après dénombrement en 2001;)

$nb_{men}$  : nombre de ménages à tirer dans une ZD ( $nb_{ZD} = 12$ ).

La probabilité d'inclusion d'un ménage dans la ville de Yaoundé ou Douala est donnée par

$$P_1 = \frac{nb_{arr}}{NB_{arr}} \times \frac{nb_{ZD}}{NB_{ZD}} \times \frac{nb_{men}}{NB_{men}}$$

**Tirage dans les grandes villes** Les grandes villes sont les villes de 50 000 habitants et plus autres que Yaoundé et Douala. Elles constituent le milieu urbain de la province. Un tirage à deux degrés est adopté dans chaque grande ville comme à Yaoundé et à Douala, à l'exception qu'au deuxième degré, 18 ménages sont tirés.

Avec les notations de Yaoundé et Douala, la probabilité d'inclusion d'un ménage dans une grande ville est donnée par

$$P_2 = \frac{nb_{ZD}}{NB_{ZD}} \times \frac{nb_{men}}{NB_{men}}$$

**Tirage dans les petites villes** Les petites villes sont les villes de 10000 à moins de 50000 habitants. Elles constituent la strate semi-urbaine des provinces et sont les chefs-lieux d'arrondissement autres que les grandes villes, Yaoundé et Douala. Un tirage à trois degrés est adopté dans chaque petite ville. Au premier degré, les villes sont tirées avec une probabilité proportionnelle à leur taille en ménage en 1987. Les degrés suivants sont identiques au tirage dans les grandes villes.

Avec les notations de Yaoundé et Douala complétées par :

$NB_{men87}$  : nombre de ménages dans la ville en 1987

$NBTOT_{men87}$  : nombre total de ménages dans toutes les petites villes de la province en 1987. La probabilité d'inclusion d'un ménage dans une grande ville est donnée par

$$P_3 = nb_{arr} \times \frac{NB_{men87}}{NBTOT_{men87}} \times \frac{nb_{ZD}}{NB_{ZD}} \times \frac{nb_{men}}{NB_{men}}$$

<sup>1</sup>Regroupement des ménages selon la cartographie de 1987 mise à jour en 2001 pour les besoins d'ECAM II.

où le terme  $nb_{arr} \times \frac{NB_{men87}}{NBTOT_{men87}}$  représente la probabilité d'inclusion d'une ville. Lorsque cette dernière est supérieure à  $1/nb_{arr}$  le chef-lieu d'arrondissement est tiré et le calcul est refait pour la sélection des autres petites villes.

**Tirage dans les zones rurales** Il est semblable au tirage dans les petites villes, à l'exception qu'au dernier degré, 27 ménages ou 36 ménages sont tirés à probabilités égales.

## 2.2 Méthodologie

### 2.2.1 Forme analytique de la variance dans le plan de sondage d'ECAM II

#### Cas d'un total

Le plan de sondage d'ECAM II peut être modélisé comme un sondage ayant 32 strates. Dans chaque strate, la sélection des ménages se fait par un sondage à trois degrés : au premier degré,  $n$  villes ou arrondissements (unités primaires) sont tirés suivant un plan de sondage  $S_1$  dans un ensemble de  $N$  villes ou arrondissements de la strate;  $S_1$  est un tirage sans remise à probabilités  $\pi_i$  proportionnelles à la taille de la ville en ménages. Dans chaque ville ou arrondissement  $i$ ,  $n_i$  zones de dénombrement (unités secondaires) sont tirées suivant un plan de sondage  $S_{2i}$  dans un ensemble de  $N_i$  ZD de la ville  $i$ ;  $S_{2i}$  est tirage sans remise à probabilités égales  $\pi_{j|i}$ . Au troisième degré,  $n_{jB}$  ménages (unités tertiaires) sont tirées suivant un plan de sondage  $S_{2i}$  dans un ensemble de  $N_{j|i}$  de ménages de la ZD  $j$  de la ville  $i$ ;  $S_{2i}$  est un tirage sans remise à probabilités égales  $\pi_{k|i,j}$

Dans une strate, le total d'une variable  $Y$  qui prend la valeur  $Y_{ijk}$  sur le ménage  $k$  de la ZD  $j$  de la ville  $i$  est  $Y = \sum_{i=1}^N \sum_{j=1}^{N_i} \sum_{k=1}^{N_{j|i}} Y_{ijk}$  et son estimateur de Horvitz et Thompson est

$$\hat{Y} = \sum_{i=1}^N \sum_{j=1}^{N_i} \sum_{k=1}^{N_{j|i}} \frac{Y_{ijk}}{\pi_i \pi_{j|i} \pi_{k|i,j}}$$

La variance de l'estimateur  $\hat{Y}$  est

$$V(\hat{Y}) = V(E(\hat{Y}|S_1)) + E(V(\hat{Y}|S_1)) \quad (2.1)$$

obtenue en conditionnant par rapport à  $S_1$ . Le premier terme de cette somme

est

$$\begin{aligned}
V\left(E(\hat{Y}|S_1)\right) &= V\left(\sum_{i \in S_1} \frac{1}{\pi_i} E\left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_{j|i}} \frac{Y_{ijk}}{\pi_{j|i}\pi_{k|i,j}} \middle| S_1\right)\right) \\
&= V\left(\sum_{i \in S_1} \frac{1}{\pi_i} E\left(\sum_{j=1}^{N_i} \sum_{k=1}^{N_{j|i}} \frac{Y_{ijk}}{\pi_{j|i}\pi_{k|i,j}} \delta_{jk|i} \middle| S_1\right)\right) \\
&= V\left(\sum_{i \in S_1} \frac{1}{\pi_i} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{j|i}} \frac{Y_{ijk}}{\pi_{j|i}\pi_{k|i,j}} E(\delta_{jk|i}|S_1)\right) \\
&= V\left(\sum_{i \in S_1} \frac{\sum_{j=1}^{N_i} \sum_{k=1}^{N_{j|i}} Y_{ijk}}{\pi_i}\right) \text{ car } E(\delta_{jk|i}|S_1) = \pi_{j|i}\pi_{k|i,j}
\end{aligned}$$

Avec  $\delta_{jk|i} = 1$  si le ménage  $k$  de la ZD  $j$  a été sélectionné sachant que la ville  $i$  est tirée, et  $\delta_{jk|i} = 0$  sinon

$$\text{D'où } V\left(E(\hat{Y}|S_1)\right) = V\left(\sum_{i \in S_1} \frac{Y_i}{\pi_i}\right)$$

Avec  $Y_i = \sum_{j=1}^{N_i} \sum_{k=1}^{N_{j|i}} Y_{ijk}$  est le total de la ville  $i$ ) qui est la variance d'un total dans un tirage à probabilités inégales. Selon les formules classiques développées dans le cadre théorique d'un sondage à un degré à probabilités inégales, cette variance vaut :

$$V\left(E(\hat{Y}|S_1)\right) = \sum_{i=1}^N \frac{Y_i^2}{\pi_i} \pi_i (1 - \pi_i) + \sum_{i=1}^N \sum_{\substack{l=1 \\ l \neq i}}^N \frac{Y_i Y_l}{\pi_i \pi_l} (\pi_{il} - \pi_i \pi_l) \quad (2.2)$$

Un estimateur de cette variance est donné par

$$\hat{V}\left(E(\hat{Y}|S_1)\right) = \sum_{i \in S_1} \frac{Y_i^2}{\pi_i} \pi_i (1 - \pi_i) + \sum_{i \in S_1} \sum_{\substack{l \in S_1 \\ l \neq i}} \frac{\hat{Y}_i \hat{Y}_l}{\pi_i \pi_l \pi_{il}} (\pi_{il} - \pi_i \pi_l) \quad (2.3)$$

Où  $\hat{Y}_i = \sum_{j=1}^{n_i} \sum_{k=1}^{n_{j|i}} \frac{Y_{ijk}}{\pi_{j|i}\pi_{k|i,j}}$  est l'estimateur du total  $Y_i$  de la ville  $i$ .

Le second terme de la somme (2.1) vaut

$$\begin{aligned}
E\left(V(\hat{Y}|S_1)\right) &= E\left(V\left(\sum_{i \in S_1} \frac{\hat{Y}_i}{\pi_i} \middle| S_1\right)\right) \\
&= E\left(\sum_{i \in S_1} \frac{1}{\pi_i^2} V(\hat{Y}_i|S_1)\right)
\end{aligned} \quad (2.4)$$

or

$$V(\hat{Y}|S_1) = V\left(E(\hat{Y}|S_1, S_2)\right) + E\left(V(\hat{Y}|S_1, S_2)\right) \quad (2.5)$$

$$\begin{aligned}
\text{où } V\left(E(\hat{Y}|S_1, S_2)\right) &= V\left(E\left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_{j|i}} \frac{Y_{ijk}}{\pi_{j|i}\pi_{k|i,j}} |S_1, S_2\right)\right) \\
&= V\left(\sum_{j=1}^{n_i} \frac{1}{\pi_{j|i}} \left(E\left(\sum_{k=1}^{N_{j|i}} \frac{Y_{ijk}}{\pi_{k|i,j}} \delta_{jk|i} |S_1, S_2\right)\right)\right) \\
&= V\left(\sum_{j=1}^{n_i} \frac{1}{\pi_i} \sum_{k=1}^{N_{j|i}} \frac{Y_{ijk}}{\pi_{k|i,j}} E(\delta_{k|ij} |S_1, S_2)\right) \\
&= V\left(\sum_{j=1}^{n_i} \frac{\sum_{k=1}^{N_{j|i}} Y_{ijk}}{\pi_{j|i}}\right) \text{ car } E(\delta_{k|ij} |S_1, S_2) = \pi_{k|ij} \\
&= V\left(\sum_{j=1}^{n_i} \frac{Y_{j|i}}{\pi_{j|i}}\right) \text{ où } Y_{j|i} = \sum_{k=1}^{N_{j|i}} Y_{ijk} \\
&= \frac{N_i(N_i - n_i)}{n_i} \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{j|i} - \bar{Y}_i)^2
\end{aligned} \tag{2.6}$$

Avec  $\delta_{k|ij} = 1$  si le ménage  $k$  a été sélectionné sachant que la ZD  $j$  de la ville  $i$  est tirée, et  $\delta_{k|ij} = 0$  sinon.

En utilisant la variance de l'estimateur d'un total dans le cas d'un sondage aléatoire simple qu'est le plan  $S_{2i}$  et où  $\bar{Y} = \frac{1}{N_i} \sum_{k=1}^{N_i} Y_{j|i}$ .

Les  $Y_{j|i}$  n'étant pas connus et encore moins pour tous les  $N_i$  éléments de la population où sont tirés les  $n_i$ , un estimateur de cette variance est donnée par

$$\hat{V}\left(E(\hat{Y}_i |S_1, S_2)\right) = \frac{N_i(N_i - n_i)}{n_i} \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\hat{Y}_{j|i} - \bar{Y}_i)^2 \tag{2.7}$$

$$\text{Où } \hat{Y}_{j|i} = \sum_{k=1}^{n_{j|i}} \frac{Y_{ijk}}{\pi_{j|i}\pi_{k|i,j}} \text{ et } \bar{Y} = \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{Y}_{j|i}.$$

Le second terme de (2.5) est

$$\begin{aligned}
E\left(V(\hat{Y}_i |S_1, S_{2i})\right) &= E\left(V\left(\sum_{j \in S_{2i}} \sum_{k \in S_{3ij}} \frac{Y_{ijk}}{\pi_{j|i}\pi_{k|i,j}} |S_1, S_{2i}\right)\right) \\
&= E\left(\sum_{j \in S_{2i}} \frac{1}{\pi_{j|i}^2} V\left(\sum_{k \in S_{3ij}} \frac{Y_{ijk}}{\pi_{k|i,j}} |S_1, S_{2i}\right)\right) \\
&= E\left(\sum_{j \in S_{2i}} \frac{1}{\pi_{j|i}^2} \left(\frac{N_{j|i}(N_{j|i} - n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i} - 1} \sum_{k=1}^{N_{j|i}} (Y_{ijk} - \bar{Y}_{j|i})^2\right)\right)
\end{aligned}$$

En utilisant la variance de l'estimateur d'un total dans le cas d'un sondage aléatoire simple qu'est le plan  $S_{3ij}$  et où  $\bar{Y}_{j|i} = \frac{1}{N_{j|i}} \sum_{k=1}^{n_{j|i}} Y_{ijk}$

$$\begin{aligned}
E\left(V(\hat{Y}_i|S_1, S_{2i})\right) &= E\left(\sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}} \left(\frac{N_{j|i}(N_{j|i}-n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i}-1} \sum_{k=1}^{N_{j|i}} (Y_{ijk} - \bar{Y}_{j|i})^2\right) \delta_{j|i}\right) \\
&= \sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}^2} \left(\frac{N_{j|i}(N_{j|i}-n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i}-1} \sum_{k=1}^{N_{j|i}} (Y_{ijk} - \bar{Y}_{j|i})^2\right) E(\delta_{j|i}) \\
&= \sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}} \left(\frac{N_{j|i}(N_{j|i}-n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i}-1} \sum_{k=1}^{N_{j|i}} (Y_{ijk} - \bar{Y}_{j|i})^2\right)
\end{aligned} \tag{2.8}$$

Car  $E(\delta_{j|i}) = E(E(\delta_{j|i}|S_1)) = \pi_{j|i}$ .

Les  $Y_{ijk}$  n'étant pas connus pour tous les  $N_{j|i}$  éléments de la population où sont tirés les  $n_{j|i}$ , un estimateur de cette variance est donnée par

$$\hat{E}\left(V(\hat{Y}_i|S_1, S_{2i})\right) = \sum_{j \in S_{2i}} \frac{1}{\pi_{j|i}^2} \left(\frac{N_{j|i}(N_{j|i}-n_{j|i})}{n_{j|i}} \frac{1}{n_{j|i}-1} \sum_{k \in S_{3ij}} (Y_{ijk} - \hat{Y}_{j|i})^2\right) \tag{2.9}$$

où  $\hat{Y}_{j|i} = \frac{1}{n_{j|i}} \sum_{k \in S_{3ij}} Y_{ijk}$

En remplaçant (2.6) et (2.8) dans (2.5), et (2.5) dans (2.4), il vient

$$\begin{aligned}
E\left(V(\hat{Y}_i|S_1)\right) &= E\left(\sum_{j \in S_{2i}} \frac{1}{\pi_i^2} \left(\frac{N_i(N_i-n_i)}{n_i} \frac{1}{N_i-1} \sum_{j=1}^{N_i} (Y_{j|i} - \bar{Y}_i)^2 + \right.\right. \\
&\quad \left.\left. + \sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}} \left(\frac{N_{j|i}(N_{j|i}-n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i}-1} \sum_{k=1}^{N_{k|ij}} (Y_{ijk} - \hat{Y}_{j|i})^2\right)\right)\right) \\
&= E\left(\sum_{i=1}^N \frac{1}{\pi_i^2} \left(\frac{N_i(N_i-n_i)}{n_i} \frac{1}{N_i-1} \sum_{j=1}^{N_i} (Y_{j|i} - \bar{Y}_i)^2 + \right.\right. \\
&\quad \left.\left. + \sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}} \left(\frac{N_{j|i}(N_{j|i}-n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i}-1} \sum_{k=1}^{N_{k|ij}} (Y_{ijk} - \hat{Y}_{j|i})^2\right)\right)\right) \delta_i
\end{aligned}$$

$$\begin{aligned}
E\left(V(\hat{Y}|S_1)\right) &= \sum_{i=1}^N \frac{1}{\pi_i^2} \left( \frac{N_i(N_i - n_i)}{n_i} \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{j|i} - \bar{Y}_i)^2 + \right. \\
&\quad \left. + \sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}} \left( \frac{N_{j|i}(N_{j|i} - n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i} - 1} \sum_{k=1}^{N_{k|ij}} (Y_{ijk} - \hat{Y}_{j|i})^2 \right) \right) E(\delta_i) \\
&= \sum_{i=1}^N \frac{1}{\pi_i} \left( \frac{N_i(N_i - n_i)}{n_i} \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{j|i} - \bar{Y}_i)^2 + \right. \\
&\quad \left. + \sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}} \left( \frac{N_{j|i}(N_{j|i} - n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i} - 1} \sum_{k=1}^{N_{k|ij}} (Y_{ijk} - \hat{Y}_{j|i})^2 \right) \right)
\end{aligned} \tag{2.10}$$

Car  $E(\delta_i) = \pi_i$  où  $\delta_i = 1$  si la ville  $i$  est tirée et  $\delta_i = 0$  sinon.

En définitive, la variance de l'estimateur  $\hat{Y}$  du total  $Y$  est obtenue en sommant (2.2) et (2.10); ainsi, (2.1) devient

$$\begin{aligned}
V(\hat{Y}) &= \sum_{i=1}^N \frac{Y_i^2}{\pi_i} \pi_i (1 - \pi_i) + \sum_{i=1}^N \sum_{\substack{l=1 \\ l \neq i}}^N \frac{Y_i Y_l}{\pi_i \pi_l} (\pi_{il} - \pi_i \pi_l) \\
&\quad + \sum_{i=1}^N \frac{1}{\pi_i} \left( \frac{N_i(N_i - n_i)}{n_i} \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{j|i} - \bar{Y}_i)^2 \right) \\
&\quad + \sum_{i=1}^N \frac{1}{\pi_i} \left( \sum_{j=1}^{N_i} \frac{1}{\pi_{j|i}} \left( \frac{N_{j|i}(N_{j|i} - n_{j|i})}{n_{j|i}} \frac{1}{N_{j|i} - 1} \sum_{k=1}^{N_{k|ij}} (Y_{ijk} - \bar{Y}_{j|i})^2 \right) \right)
\end{aligned} \tag{2.11}$$

En utilisant l'approximation proposée par Deville en (1.9), les équations (2.3), (2.7) et (2.9), (2.11) peut être estimé par

$$\begin{aligned}
\hat{V}(\hat{Y}) &= \frac{1}{1 - \sum_{i \in S_1} a_i^2} \sum_{i \in S_1} ((1 - \pi_i) \left( \frac{\hat{Y}_i}{\pi_i} - A \right)^2 \\
&\quad + \sum_{i \in S_1} \frac{1}{\pi_i^2} \left( \frac{N_i(N_i - n_i)}{n_i} \frac{1}{n_i - 1} \sum_{j \in S_{2i}} (\hat{Y}_{j|i} - \bar{Y}_i)^2 \right) \\
&\quad + \sum_{i \in S_1} \frac{1}{\pi_i^2} \left( \sum_{j \in S_{2i}} \frac{1}{\pi_{j|i}^2} \left( \frac{N_{j|i}(N_{j|i} - n_{j|i})}{n_{j|i}} \frac{1}{n_{j|i} - 1} \sum_{k \in S_{3ij}} (Y_{ijk} - \hat{Y}_{j|i})^2 \right) \right)
\end{aligned} \tag{2.12}$$

$$\begin{aligned}
\text{Où } \hat{Y}_i &= \sum_{j=1}^{n_i} \sum_{k=1}^{n_{j|i}} \frac{Y_{ijk}}{\pi_{j|i} \pi_{k|i,j}}, & a_i &= \frac{1 - \pi_i}{\sum_{i' \in S_1} (1 - \pi_{i'})}, & A &= \sum_{i \in S_1} a_i \frac{\hat{Y}_i}{\pi_i}, & \hat{Y}_{j|i} &= \\
\sum_{k=1}^{n_{j|i}} \frac{Y_{ijk}}{\pi_{k|i,j}}, & \bar{Y}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{Y}_{j|i}, & \bar{Y}_{j|i} &= \frac{1}{n_{j|i}} \sum_{k \in S_{3ij}} Y_{ijk}.
\end{aligned}$$

L'expression (2.12) de l'estimateur de la variance d'un estimateur s'interprète en termes de contribution de chaque degré de tirage à la variance. Le premier terme de (2.12) est l'estimateur de la variance du total calculé sur les UP ; le deuxième terme est l'estimateur de la variance du total calculé sur les US ; et le troisième terme est l'estimateur de la variance du total calculé sur les UT.

Bien que ne faisant pas partie des résultats attendus d'ECAM II, l'estimation des totaux a été faite dans le cadre d'une meilleure lisibilité des indicateurs produits par ECAM II. De plus, ces totaux interviennent généralement dans l'estimation des indicateurs complexes.

Un total estimé par ECAM II est la population par strate et la population totale en 2001. En considérant  $Y$  comme une variable indiquant le nombre d'individus résidant dans un ménage, une estimation de la population par strate peut être faite selon l'expression obtenue en (2.12). La population totale est estimée en combinant (2.12) et (1.11) où la variance à l'intérieur des strates est calculée à partir de l'expression (2.12).

### Précision des estimateurs complexes

Dans le cas des estimateurs complexes, la méthode de linéarisation est utilisée et l'estimateur de la variance de l'estimateur est approché par l'expression (2.12) appliquée sur la linéarisée de l'estimateur.

Ainsi, en considérant la moyenne d'un domaine, par exemple la consommation moyenne des hommes chefs de ménage, qui s'écrit comme un ratio  $R = \frac{T_Y}{T_X}$ ,  $T_Y$  et  $T_X$  sont des totaux des variables  $Y$  pour la consommation totale des hommes et  $X$  pour le nombre d'hommes et dont l'estimateur de Horvitz et Thomson est  $\hat{R}_\pi = \frac{\hat{t}_{\pi,Y}}{\hat{t}_{\pi,X}}$  et sa linéarisée  $\hat{z}_k = \frac{1}{\hat{t}_{\pi,X}} (Y_k - \hat{R}_\pi X_k)$ , un estimateur de la variance dans une strate est donnée par (2.12). Cette linéarisée est estimée par application sur le total de  $\hat{z}$ .

Dans le cas de l'indice de Gini, le revenu est représenté par la dépense de consommation totale du ménage. Cette dépense de consommation comprend les dépenses alimentaires, non alimentaires et le solde des transferts. Le rang de chaque ménage est approché par un estimateur de type Horvitz et Thomson, soit  $\hat{r}(k') = \sum_{k'' \in S} w_{k''} 1_{Y_{k''} < Y_{k'}}$ . Chaque terme de l'expression (2.12) correspond ainsi à un estimateur de Horvitz et Thomson de la vraie valeur dans l'indice de Gini. La linéarisée de cette expression est calculée en (1.14). Son estimateur est

$$\hat{lin}_k(G) = \frac{2 \left( Y_k r(k) + \sum_{k' \in S} w_{k'} 1_{Y_{k'} < Y_k} Y_{k'} \right) - Y_k - \left( G(\hat{M}) + 1 \right) \left( \sum_{k' \in S} w_{k'} Y_{k'} + \left( \sum_{k' \in S} w_{k'} \right) Y_k \right)}{\left( \sum_{k' \in S} w_{k'} \right) \left( \sum_{k' \in S} w_{k'} Y_{k'} \right)} \quad (2.13)$$

où les  $w_k$  sont les inverses des probabilités d'inclusion et

$$G(\hat{M}) = \frac{\sum_{k' \in S} (2\hat{r}(k') - 1) w_{k'} Y_{k'}}{\sum_{k' \in S} w_{k'} \sum_{k' \in S} w_{k'} Y_{k'}} - 1. * \quad (2.14)$$

En ce qui concerne l'indicateur d'Atkinson, dans le cas où  $a = 0$ , l'estimateur utilisé introduit la fonction logarithme pour éviter l'apparition des termes très grands lors du produit des dépenses des ménages. Cette variante s'écrit

$$A_0(\hat{M}) = 1 - \exp(\log(\hat{Y}) + \frac{1}{a} \sum_{k'l \in S} \log(Y_{k'l})). \quad (2.15)$$

Les estimateurs des linéarisées d'Atkinson devant permettre le calcul de la variance sont pour  $a \neq 0$  :

$$\hat{lin}_k(A_a) = \frac{1 - a_0(\hat{M})}{N} \left( \frac{Y_k}{\hat{Y}} - 1 - \frac{1}{a} \left( \frac{NY_k^a}{\sum_{k'l \in S} w_{k'l} Y_{k'l}^a} - 1 \right) \right) \quad (2.16)$$

et pour  $a = 0$  :

$$\hat{lin}_k(A_0) = \frac{1 - A_0(\hat{M})}{N} \left( \frac{Y_k}{\hat{Y}} - 1 - \log(Y_k) + \frac{\sum_{k'l \in S} w_{k'l} \log(Y_{k'l})}{N} \right) \quad (2.17)$$

### 2.2.2 Méthode de réplification par le Jackknife

Pour calculer le Jackknife de chaque estimateur, la méthodologie utilisée est proche de celle mise en œuvre dans l'EDSCIII. Les grappes sont ici remplacées par les ZD. Ainsi, le Jackknife est calculé sur 612 zones de dénombrement. Les expressions des estimateurs, du biais et de la variance sont celles tirées de Bontempi(?) et présentées dans le cadre théorique

### 2.2.3 Estimation de la variance avec prise en compte des données manquantes

Les cas de non-réponses traités ici sont des non-réponses partielles.

S'agissant des non réponses totales et des questionnaires rejetés ou incomplets (l'enquête est complète pour un ménage si les sections 00 à 15 sont toutes fournies et si la section 15 contient les dépenses quotidiennes pour au moins 10 jours sur les 15 requis en milieu urbain et semi-urbain), leur taux est de 4,9%. La répartition par montre que les taux les plus élevés sont dans les provinces du Sud-Ouest (12,9La bonne qualité de l'enquête (le taux de non-réponse prévue était de 5% s'explique par le succès des campagnes de sensibilisation auprès des ménages, la flexibilité de la procédure de remplacement, et la forte motivation des agents enquêteurs dont la prime de fin de contrat était liée à leur rendement.

#### Pour des variables quantitatives

La méthodologie qui est mise en œuvre est proche de celle proposée par Münnich et Rässler. Dans un premier temps, nous simulons  $m$  tableaux de données à partir de la méthode MCMC ; pour chaque tableau de données, un estimateur de Horvitz et Thomson est calculé ; une moyenne des estimateurs permet d'estimer le paramètre d'intérêt ; la variance est décomposée en variance intra-tableaux qui est la moyenne des variances obtenues dans chaque tableau par une l'estimateur analytique ou une

méthode de réplication et la variance inter-tableau qui variance des estimateurs du paramètre d'intérêt sur chaque tableau.

Formellement, soient

$m$  le nombre de simulations indépendantes sont réalisées

$\hat{t}_{MCMC}^k$  : l'estimateur Horwitz-Thompson

$Y_{obs}$  : ensemble des variables ayant des données observés

$Y_{men}$  : ensemble des variables ayant des données manquants

$T$  : une statistique et  $\hat{T}$  un estimateur

$$\hat{T}_{MCMC}^{(k)} = \hat{T}(Y_{obs}, Y_m^{(k)} an)$$

et sa variance

$$V(\hat{T}_{MCMC}^{(k)}) = V(\hat{T}(Y_{obs}, Y_m^{(k)} an))$$

où  $k = 1$  à  $m$ .

L'estimateur

$$\hat{T}_{MCMC} = \frac{1}{m} \sum_{k=1}^m \hat{T}_{MCMC}^{(k)}$$

et sa variance

$$\hat{V}(\hat{T}_{MCMC}) = \hat{V}_{intra}(\hat{T}) + \frac{m+1}{m} \hat{V}_{inter}(\hat{T})$$

où

$$\hat{V}_{intra}(\hat{T}) = \frac{1}{m-1} \sum_{k=1}^m (\hat{T}_{MCMC}^{(k)} - \hat{T}_{MCMC})^2$$

et

$$\hat{V}_{inter}(\hat{T}) = \frac{1}{m} \sum_{k=1}^m V(\hat{T}_{MCMC}^{(k)})$$

. Seules deux variables quantitatives retenues dans l'étude ont été identifiées comme contenant des valeurs manquantes. Ceci provient du fait que par défaut dans les masques de saisie, les variables quantitatives sont initialisées à 0. Ces taux semblent

TAB. 2.1 – Variables quantitatives et leurs valeurs initialisées

Libellé variable	Effectif manquant	pourcentage
Consultation	4378	39,92%

être élevés pour être assimilés à des valeurs manquantes dans l'enquête. Même si le logiciel a été utilisé pour l'estimation des données manquantes par MCMC à titre d'exercice (voir le listing en annexe), l'estimation obtenu sur les données corrigées à conduits à des résultats qui ne pouvaient pas être validés et ne sont donc pas présentés. Etant donné le très faible

TAB. 2.2 – Variables qualitatives et leurs données manquantes

Libellé variable	Effectif manquant	pourcentage
Possession d'une terre exploitée	0	0%
Membre du ménage membre d'une association	1	0,01%
Paiement des frais non réglementaire pour la scolarisation	25	0,23%
Peiement des frais non réglémentataire pour les soins médicaux	18	0,16%
Paiement des frais non réglementaires pour autres services	17	0,15%
Atteint du paludisme	26	0,24%
Savoir lire et écrire	7	0,06%
Membre du ménage membre d'une association	1	0,01%

### Pour des variables qualitatives

La méthode de repondération est utilisé pour les variable qualitative. Les taux de non-réponse enregistrés se présentent comme suit

Pour une variable, le taux  $t$  de non réponse permet de corriger la propobilité d'inclusion du ménage dans la ZD. Ainsi,on obtient une nouvelle probabilité d'inclusion

$$\pi'_{k||i,j} = p_{i_k||i,j} * t$$

qui permet d'écrire l'estimateur de Horvitz et Thomson et le calcul de sa précision.

## 2.3 Autres indicateurs dérivées du calcal de précision des estimateurs

A partir d'un estimateur et de l'estimation de sa variance, plusieurs autres indicateurs peuvent être calculés pour apprécier la qualité de la précision.En considérant  $\theta$  un paramètre,  $\hat{\theta}$  son estimateur et  $\hat{\sigma}$  la racine carrée de l'estimateur de la variance  $\hat{\theta}^2$ , on peut définir le coefficient de variation, l'effet de sondage, l'effet de grappe et l'intervalle de confiance.

### 2.3.1 Coefficient de variation ou CV

Le CV est le rapport de l'estimateur de l'écart type d'un estimateur sur cet estimateur ; il est donné par la formule

$$CV = \frac{\hat{\sigma}}{\hat{\theta}}$$

Plus le coefficient de variation est petit, plus les différentes valeurs prises par l'estimateur sont proches.

Les taux de non-réponses obtenus ont conduit à de très faible correction qui n'ont pas modifié les résultats.

### 2.3.2 Effet de sondage

Il est calculé comme le rapport entre l'estimation de la variance dans le cas d'un sondage complexe et l'estimation de la variance calculée comme si l'échantillon avait été tiré par un SAS. La racine carrée de l'effet de sondage révèle dans quelle mesure le plan de sondage mis en œuvre se rapproche d'un échantillon aléatoire simple alors qu'une valeur supérieure à 1 indique un accroissement de l'erreur de sondage dû à un plan de sondage complexe et moins efficace au point de vue statistique. Il est encore appelé coefficient multiplicateur de l'intervalle de confiance d'un SAS

### 2.3.3 Effet de grappe

L'effet de grappe est mesuré par le coefficient de corrélation inter-grappe qui a été développé dans le cadre théorique de ce travail.

### 2.3.4 Intervalle de confiance

L'intervalle de confiance est calculé pour un niveau de maximal de 0,95 en utilisant le fractile  $z$  de la loi normale. Comme cette loi est symétrique, le minimum de l'intervalle de confiance est donné par

$$Min.IC = \hat{\theta} + z\hat{\sigma}$$

et son maximum est donné par

$$Max.IC = \hat{\theta} + z\hat{\sigma}$$

## RÉSULTATS ET INTERPRÉTATION

---

et le Jackknife. Les codes nécessaires aux différents calculs ont été pour l'essentiel écrits à l'aide du logiciel statistique R.

Ne disposant pas des probabilités de tirage aux trois degrés, nous avons en nous calant sur le coefficient d'extrapolation disponible par individu simulé des probabilités d'inclusion par degré en nous procédant comme suit pour chaque strate :

- Chaque ZD est supposée contenir 300 ménages ; En calculant extrayant le nombre de ménages tirés par ZD, il est facile de déduire la probabilité de tirage d'un ménage connaissant la ZD et la ville ;
- Les villes ayant été tirées proportionnellement à leur effectif en ménage ; la probabilité d'inclusion d'une ville est égale à l'effectif de ménage de la ville divisé par le nombre total de ménages de la strate
- La probabilité d'une ZD est calculée comme solde en divisant la probabilité d'inclusion du ménage par le produit de la probabilité d'inclusion du ménage sachant la ville et la ZD et la probabilité d'inclusion de la ville.

Le jackknife a été testé sur l'estimation de la variance du nombre de ménage par construction d'une variable unité. Les résultats montrent que l'estimation du total des ménages appartient bien à l'intervalle de confiance calculé. La présentation des résultats et leur interprétation s'articulent autour des quatre thèmes retenus dans le calcul de la précision des indicateurs d'ECAM II. Les variables utilisées sont d'abord explorées pour apprécier la qualité des données collectées, leur précision est ensuite calculée, utilisant les expressions analytiques

### 3.1 Capital humain

Il regroupe les variables relatives à l'éducation et à la santé. L'éducation est appréhendée par l'alphabétisation (s02q11a) et le niveau d'instruction le plus élevé du ménage (s03q10). La santé est caractérisée par la prévalence du paludisme (s02q11a), les dépenses en médicament (pharma) et dépenses en consultation (consult). Les résultats des estimations figurent en annexes dans le tableau 1. Ils proviennent d'une estimation du ratio de deux totaux par la méthode de linéarisation du total et le calcul de la variance de la linéarisée. Les estimateurs des ratios sont les rapports des estimateurs de Horvitz et Thomson des totaux. Comme prévue par les expressions analytiques, la contribution des degrés de tirage va décroissant des unités primaires vers les unités tertiaires. Dans l'ensemble, les coefficients de variation sont faibles

(moins de 0,5%) pour toutes les variables, traduisant une faible dispersion des ratios. L'effet de grappe est positif ce qui reflète une similitude entre les ménages. Cet effet de grappe est plus élevé dans le calcul de la précision des domaines, ce qui laisse suggérer un effet taille positif. Les intervalles de confiance obtenus en mettant en œuvre le plan de sondage d'ECAM II sont au moins trois fois plus grand que ceux obtenus dans le cas d'un tirage du même échantillon avec un sondage aléatoire simple. Les écarts types de l'estimation du ratio dans les domaines sont en général plus grand corroborant ainsi l'effet de grappe plus élevé que dans l'ensemble déjà observé. Les variances dans le domaine défini par les femmes chefs de ménages sont en général plus petites que celles calculées dans le domaine défini par les hommes chefs de ménages

## 3.2 Pauvreté monétaire

La pauvreté, vue sous l'angle monétaire, est surtout appréhendée par la fraction des ménages vivant en dessous d'un certain seuil. En 2001, L'ECAM II a ainsi calculé un seuil de pauvreté alimentaire (151 398 F CFA), un seuil de pauvreté non alimentaire (81 149 F CFA) et un seuil de pauvreté globale (232 547 F CFA). En construisant une indicatrice qui prend la valeur 1 si le ménage a une consommation en dessous du seuil correspondant, les précisions des taux de pauvreté alimentaire, non alimentaire et globale peuvent être calculées. De nouveau, ces calculs aboutissent à l'estimation de la précision d'un ratio par linéarisation. Les coefficients de variation sont d'un ordre de grandeur inférieur à l'unité, donc une faible dispersion de la distribution des totaux des variables de pauvreté monétaire; la pauvreté globale est la variable la plus dispersée avec un coefficient de variation de 0,2% et un coefficient multiplicateur de l'intervalle de confiance d'un SAS de 5,6.

cette analyse de la pauvreté monétaire est complétée par le calcul de la précision de l'estimateur du Coefficient de Gini et d'Atkinson. La contribution du troisième degré est nulle dans le coefficient d'Atkinson. Le coefficient de variation de ces deux indices sont tous inférieurs à l'unité. Leur variance calculée dans le cas d'un SAS est nulle ce qui ne permet pas de disposer de l'effet de grappe et du coefficient de multiplication de l'intervalle de confiance.

## 3.3 Vulnérabilité

Trois de variables ont permis d'approcher la vulnérabilité des ménages. Ces groupes sont la possession des commodités, la possession/exploitation d'une terre et l'appartenance à un groupe associatif.

Les items relatifs à la possession de commodité ont été regroupés en possession de moyens de télécommunication, possession de moyen de communication, possession de matériel moderne de cuisson, possession d'autres biens de luxe tels que ventilateur, fer à repasser, etc. Chaque groupe est représenté par une variable indicatrice qui prend la valeur 1 si le ménage possède au moins un item du groupe et 0 sinon. Ce procédé ramène donc l'estimateur des taux de possession à l'estimation d'un ratio comportant au numérateur le total de ménages possédant l'un au moins des items

et au dénominateur le nombre total des ménages. Ce calcul est prolongé dans les domaines en se restreignant aux individus du domaine. Dans l'ensemble, l'appartenance à un groupe associatif, la possession d'un moyen de télécommunication et la possession d'un moyen de télécommunication sont les variables de vulnérabilité les moins dispersées.

### **3.4 Bonne gouvernance**

La bonne gouvernance est mesurée par le paiement des frais non réglementaire pour la scolarisation, pour les soins médicaux, pour autres services et le paiement volontaire de frais à un agent des forces de l'ordre. Les coefficients de variation sont tous inférieurs 1%. ; les effets de grappe se situent autour de 0,3 et les intervalles de confiance obtenus sont tous plus de trois fois supérieurs à l'intervalle de confiance obtenu dans le cas d'un SAS.

---

---

# Conclusion

---

En plus des objectifs pédagogiques visant entre autres l'application des enseignements reçus pendant la formation en mastère de de statistique et l'impregnation dans le monde professionnel, le présent stage s'est proposé d'une part de parachever le plan de sondage retenu dans la deuxième Enquête Camerounaise en proposant une méthodologie de calcul des estimateurs complexes retenus dans cette enquête et, d'autre part d'examiner la portée des données manquantes dans la précision de ces indicateurs. Par ailleurs, deux nouveaux indicateurs d'inégalité à savoir le Theil et l'Atkinson ont été estimés et leur précision calculée.

La méthodologie mise en œuvre s'appuie sur les travaux antérieurs dans le domaine des sondages qui pour la plupart soit s'arrêtent ou se ramènent à un plan de sondage à deux degrés au plus, soit alors proposent des algorithmes, pour dériver la variance à trois degrés avec probabilités inégales dans le tirage des unités primaires de l'estimateur d'un total. Pour les estimateurs complexes tels que l'estimateur d'un ratio ou l'estimateur de l'indice de Gini, les techniques de linéarisation permettent de se ramener approximativement à l'estimation de la variance d'un total. Dans les cas des estimateurs de total et des estimateurs complexes, les méthodes de réplification à savoir le Jackknife permettent de conforter les résultats obtenus par les approximations analytiques.

La méthodologie proposée ici peut facilement être adaptée à d'autres types d'enquête réalisée à l'Institut National de la Statistique notamment l'Enquête Emploi et Secteur Informel réalisé en 2005. Les résultats ont été obtenus à l'aide d'une simulation des probabilités d'inclusion faute de données sources ayant permis le tirage. Ceci n'a pas permis la comparaison des résultats et surtout l'appartenance à l'intervalle de confiance des estimateurs des indicateurs d'ECAM II et limite leur portée. Les programmes informatiques sont disponibles pour une nouvelle compilation des données pour disposer des estimations plus robustes. En termes de perspectives, il peut être envisager de prolonger le calcul de la variance pour les taux de pauvreté avec un seuil de pauvreté endogénéisé. Le seuil de pauvreté en lui même peut constituer un pan de la recherche pour le calcul de la précision ; en effet, il est calculé à la fois à partir des données collectées dans ECAM II mais aussi à partir des prix relevés dans les différents points de vente de certaines localités du pays. L'enjeu serait alors de consolider les deux plans de sondage pour le calcul de la variance du seuil de pauvreté.

---

# Annexes

---

TAB. 3.1 – Description des abréviations des lignes utilisés dans les tableaux

Code ligne	Libellé
est.HT	: Estimateur de Hurvitz et Thomson
SE.deg1	: Écart type de la Contribution du premier degré
SE.deg2	: Écart type de la Contribution du deuxième degré
SE.deg3	: Écart type de la Contribution du troisième degré
SE	: Écart type du total de la variable
CV(%)	: Coefficient de variation
SE.SAS	: Écart type dans le cas d'un sondage aléatoire simple
DEFF	: Effet de sondage
Effet.grap(rho)	: Effet de grappe
Min.IC	: Minimum de l'intervalle de confiance au risque 5%
Max.IC	: Maximum de l'intervalle de confiance au risque 5%
Coef.IC	: Coefficient multiplicateur de l'intervalle de confiance

TAB. 3.2 – Précision des indicateurs du capital humain pour l'ensemble des ménages

	palu	alpha	sans_niv	prim	second	sup	pharma	consult
est.HT	0,141	0,643	0,322	0,328	0,286	0,064	52244,41	23498,817
SE.deg1	0,008	0,027	0,025	0,012	0,018	0,01	4477,181	3368,618
SE.deg2	0,007	0,013	0,013	0,012	0,013	0,009	4901,028	5790,115
SE.deg3	0,004	0,005	0,005	0,006	0,005	0,002	1383,831	1248,442
SE	0,01	0,028	0,026	0,016	0,022	0,014	6652,199	6770,017
CV(%)	0,074	0,043	0,08	0,05	0,076	0,212	0,127	0,288
SE.SAS	0,003	0,004	0,004	0,004	0,004	0,003	1753,031	1219,498
DEFF	9,82	39,774	36,28	13,907	23,078	28,284	14,4	30,819
Effet.grap(rho)	0,107	0,471	0,429	0,157	0,268	0,332	0,163	0,362
Min.IC	0,121	0,588	0,271	0,297	0,243	0,037	39206,34	10229,828
Max.IC	0,161	0,698	0,373	0,359	0,329	0,091	65282,48	36767,806
Coef.IC	3,134	6,307	6,023	3,729	4,804	5,318	3,795	5,551

TAB. 3.3 – Précision des indicateurs du capital humain pour le domaine des hommes chefs de ménage

	palu	alpha	sans_niv	prim	second	sup	pharma	consult
est.HT	0,127	0,688	0,281	0,342	0,303	0,075	54414,397	24602,271
SE.deg1	0,009	0,028	0,026	0,014	0,019	0,011	4843,618	3559,794
SE.deg2	0,009	0,017	0,016	0,015	0,016	0,011	5383,583	7119,773
SE.deg3	0,004	0,005	0,005	0,006	0,006	0,003	1631,238	1486,369
SE	0,012	0,03	0,028	0,02	0,024	0,015	7263,711	8053,642
CV(%)	0,092	0,044	0,099	0,058	0,079	0,207	0,133	0,327
SE.SAS	0,004	0,005	0,005	0,005	0,005	0,003	1736,775	1530,311
DEFF	10,084	39,682	35,589	15,081	21,253	23,66	17,492	27,697
Effet.grap(rho)	0,147	0,624	0,558	0,227	0,327	0,366	0,266	0,431
Min.IC	0,103	0,629	0,226	0,303	0,256	0,046	40177,785	8817,423
Max.IC	0,151	0,747	0,336	0,381	0,35	0,104	68651,009	40387,119
Coef.IC	3,175	6,299	5,966	3,883	4,61	4,864	4,182	5,263

TAB. 3.4 – Précision des indicateurs du capital humain pour le domaine des hommes chefs de ménage

	palu	alpha	sans_niv	prim	second	sup	pharma	consult
est.HT	0,19	0,5	0,45	0,28	0,23	0	45306,79	19970,99
SE.deg1	0,01	0,03	0,03	0,01	0,02	0	4150,85	3209,1
SE.deg2	0,02	0,02	0,02	0,02	0,02	0	8552,64	5113,57
SE.deg3	0,01	0,01	0,01	0,01	0,01	0	2140,07	1661,2
SE	0,02	0,04	0,04	0,03	0,03	0	9671,91	6155,44
CV(%)	0,1	0,08	0,08	0,09	0,13	0	0,21	0,31
SE.SAS	0,01	0,01	0,01	0,01	0,01	0	4761,73	1577,22
DEFF	6,1	15,41	14,69	9,45	11,99	0	4,13	15,23
Effet.grap(rho)	0,26	0,74	0,7	0,43	0,56	0	0,16	0,73
Min.IC	0,151	0,422	0,372	0,221	0,171	0	26350,195	7906,549
Max.IC	0,229	0,578	0,528	0,339	0,289	0	64263,385	32035,431
Coef.IC	2,47	3,93	3,83	3,07	3,46	0	2,03	3,9

palu= part de ménage ayant été atteint par le paludisme ;

alpha : part des ménages sachant lire et écrire ;

sans\_niv : part des ménages sans niveau d'instruction ;

prim : part des ménages ayant effectué le cycle primaire ;

second : part des ménages ayant effectué le cycle secondaire ;

sup : part des ménages ayant effectué le cycle supérieur ;

pharma : dépense moyenne en médicament par ménage ;

consult : dépense moyenne en frais de consultation par ménage

TAB. 3.5 – Précision des indicateurs de la pauvreté monétaire pour l'ensemble des ménages

	depalim	depnalim	deptot
est.HT	0,081	0,039	0,037
SE.deg1	0,009	0,006	0,007
SE.deg2	0,006	0,006	0,006
SE.deg3	0,003	0,002	0,002
SE	0,01	0,008	0,008
CV(%)	0,121	0,202	0,228
Min.IC	0,002	0,002	0,002
Max.IC	18,457	27,591	31,653
SE.SAS	0,212	0,323	0,373
REPS	0,061	0,023	0,021
Effet.grap(rho)	0,101	0,055	0,053
Coef.IC	4,296	5,253	5,626

TAB. 3.6 – Précision des indicateurs de la pauvreté monétaire pour le domaine des femmes chefs de ménage

	depalim	depnalim	deptot
est.HT	0,146	0,088	0,075
SE.deg1	0,016	0,013	0,013
SE.deg2	0,024	0,011	0,025
SE.deg3	0,008	0,006	0,006
SE	0,029	0,016	0,027
CV(%)	0,197	0,185	0,365
Min.IC	0,006	0,004	0,004
Max.IC	24,43	14,113	45,288
SE.SAS	1,204	0,674	2,275
REPS	0,089	0,057	0,022
Effet.grap(rho)	0,203	0,119	0,128
Coef.IC	4,943	3,757	6,73

TAB. 3.7 – Précision des indicateurs de la pauvreté monétaire pour le domaine des hommes chefs de ménage

	depalim	depnalim	deptot
est.HT	0,061	0,024	0,025
SE.deg1	0,007	0,004	0,005
SE.deg2	0,005	0,005	0,004
SE.deg3	0,003	0,002	0,002
SE	0,008	0,007	0,006
CV(%)	0,134	0,293	0,252
Min.IC	0,002	0,001	0,001
Max.IC	12,037	24,932	18,426
SE.SAS	0,178	0,386	0,281
REPS	0,045	0,01	0,013
Effet.grap(rho)	0,077	0,038	0,037
Coef.IC	3,469	4,993	4,293

TAB. 3.8 – Précision de l'estimateur de l'indicateur de GINI

	depalim	depnalim	deptot
est.HT.GINI	0.397	0.564	0.46
SE.deg1	0.141	0.126	0.14
SE.deg2	0.018	0.069	0.089
SE.deg3	0.005	0.008	0.012
SE	0.127	0.129	0.149
CV(%)	0.321	0.229	0.324
Min.IC	0.147	0.311	0.168
Max.IC	0.647	0.817	0.752

TAB. 3.9 – Précision de l'estimateur de l'indicateurs d'Atkinson

	depalim	depnalim	deptot
est.HT.ATKI		0.431	0.304
SE.deg1		0.03	0.017
SE.deg2		0.051	0.03
SE.deg3		0	0
SE		0.059	0.034
CV(%)		0.136	0.112
Min.IC		0.316	0.237
Max.IC		0.546	0.37

depalim= taux de pauvreté alimentaires ; depnalim= taux de pauvreté non alimentaires ; deptot = taux de pauvreté globale

TAB. 3.10 – Précision des indicateurs de vulnérabilité pour l'ensemble des ménages

	telecom	commu	cuisson	luxe	terre	asso
est.HT	0,573	0,188	0,368	0,41	0,59	0,574
SE.deg1	0,022	0,013	0,033	0,029	0,036	0,018
SE.deg2	0,011	0,012	0,015	0,012	0,009	0,017
SE.deg3	0,006	0,005	0,004	0,005	0,004	0,005
SE	0,023	0,016	0,035	0,03	0,036	0,023
CV(%)	0,041	0,084	0,095	0,072	0,061	0,04
SE.SAS	0,005	0,004	0,005	0,005	0,005	0,005
DEFF	26,255	18,903	54,18	39,04	56,706	23,524
Effet.grap(rho)	0,307	0,218	0,646	0,462	0,677	0,274
Min.IC	0,528	0,157	0,299	0,351	0,519	0,529
Max.IC	0,618	0,219	0,437	0,469	0,661	0,619
Coef.IC	5,124	4,348	7,361	6,248	7,53	4,85

TAB. 3.11 – Précision des indicateurs de vulnérabilité pour le domaine des femmes chefs de ménage

	telecom	commu	cuisson	luxe	terre	asso
est.HT	0,403	0,04	0,343	0,333	0,531	0,598
SE.deg1	0,029	0,005	0,036	0,032	0,038	0,02
SE.deg2	0,025	0,012	0,024	0,026	0,026	0,026
SE.deg3	0,01	0,004	0,007	0,008	0,008	0,01
SE	0,038	0,013	0,043	0,041	0,045	0,032
CV(%)	0,095	0,332	0,124	0,123	0,084	0,053
SE.SAS	0,01	0,004	0,01	0,01	0,01	0,009
DEFF	15,739	9,755	19,905	18,473	21,908	11,203
Effet.grap(rho)	0,757	0,45	0,971	0,898	1,074	0,524
Min.IC	0,329	0,015	0,259	0,253	0,443	0,535
Max.IC	0,477	0,065	0,427	0,413	0,619	0,661
Coef.IC	3,967	3,123	4,462	4,298	4,681	3,347

TAB. 3.12 – Précision des indicateurs de vulnérabilité pour le domaine des hommes chefs de ménage

	telecom	commu	cuisson	luxe	terre	asso
est.HT	0,626	0,234	0,376	0,434	0,609	0,566
SE.deg1	0,023	0,015	0,033	0,029	0,037	0,019
SE.deg2	0,011	0,015	0,018	0,013	0,013	0,018
SE.deg3	0,006	0,006	0,005	0,006	0,004	0,006
SE	0,023	0,02	0,036	0,03	0,037	0,024
CV(%)	0,037	0,084	0,096	0,07	0,061	0,043
SE.SAS	0,005	0,005	0,005	0,005	0,005	0,005
DEFF	21,127	18,954	43,725	30,723	45,881	19,982
Effet.grap(rho)	0,325	0,29	0,69	0,48	0,724	0,306
Min.IC	0,581	0,195	0,305	0,375	0,536	0,519
Max.IC	0,671	0,273	0,447	0,493	0,682	0,613
Coef.IC	4,596	4,354	6,613	5,543	6,774	4,47

**telecom** : part de ménages possédant un téléphone ; **commu** : part de ménages possédant un outil de communication ; **cuisson** : part de ménage possédant un outil moderne de cuisson ; **luxe** : part de ménage possédant d'autres bien de luxe ; **terre** : part de ménages exploitant une terre ; **asso** : part de ménage appartenant à une association

TAB. 3.13 – Précision des indicateurs de bonne gouvernance pour l'ensemble des ménages

	scola	medi	autres	volon
est.HT	0,149	0,211	0,261	0,174
SE.deg1	0,011	0,017	0,015	0,01
SE.deg2	0,01	0,012	0,014	0,013
SE.deg3	0,004	0,004	0,005	0,005
SE	0,014	0,02	0,02	0,015
CV(%)	0,096	0,096	0,076	0,089
SE.SAS	0,004	0,004	0,004	0,004
DEFF	16,519	26,454	21,652	19,173
Effet.grap(rho)	0,189	0,309	0,251	0,221
Min.IC	0,122	0,172	0,222	0,145
Max.IC	0,176	0,25	0,3	0,203
Coef.IC	4,064	5,143	4,653	4,379

TAB. 3.14 – Précision des indicateurs de bonne gouvernance pour le domaine des hommes chefs de ménage

	scola	medi	autres	volon
est.HT	0,14	0,175	0,163	0,092
SE.deg1	0,012	0,017	0,013	0,008
SE.deg2	0,019	0,022	0,022	0,015
SE.deg3	0,007	0,007	0,007	0,007
SE	0,022	0,028	0,026	0,018
CV(%)	0,16	0,158	0,158	0,189
SE.SAS	0,007	0,007	0,007	0,005
DEFF	10,366	13,735	12,202	10,638
Effet.grap(rho)	0,481	0,654	0,575	0,495
Min.IC	0,097	0,12	0,112	0,057
Max.IC	0,183	0,23	0,214	0,127
Coef.IC	3,22	3,706	3,493	3,262

TAB. 3.15 – Précision des indicateurs de bonne gouvernance pour le domaine des femmes chefs de ménage

	scola	medi	autres	volon
est.HT	0,151	0,222	0,291	0,199
SE.deg1	0,012	0,019	0,016	0,012
SE.deg2	0,011	0,013	0,015	0,015
SE.deg3	0,004	0,005	0,006	0,005
SE	0,016	0,021	0,021	0,018
CV(%)	0,103	0,096	0,072	0,089
SE.SAS	0,004	0,005	0,005	0,004
DEFF	14,568	21,704	17,532	17,171
Effet.grap(rho)	0,219	0,334	0,267	0,261
Min.IC	0,12	0,181	0,25	0,164
Max.IC	0,182	0,263	0,332	0,234
Coef.IC	3,817	4,659	4,187	4,144

**scola** : part de ménages ayant payé des frais non réglementaire pour la scolarisation ;

**medi** : part de ménages ayant payé des frais non réglementaire pour les soins médicaux ;

**autres** : part de ménages ayant payé d'autre frais non réglementaire ;

**volon** : part des ménages ayant volontairement payé des frais à un agent des forces de l'ordre

TAB. 3.16 – Précision de l'estimateur de l'effectif total des ménages par le Jackknife

	unite
jack.tot.HT	3115836,089
se.jack.tot.HT	6465,436
Min.IC.jack.tot	3103164,067
Max.IC.jack.tot	3128508,112

The SAS System 09:18 Tuesday, July 2, 2002 3

The MI Procedure

Model Information

```

Data Set          ZACH.ECAM_MENO
Method           MCMC
Multiple Imputation Chain  Single Chain
Initial Estimates for MCMC  EM Posterior Mode
Start            Starting Value
Prior            Jeffreys
Number of Imputations      5
Number of Burn-in Iterations 200
Number of Iterations       100
Seed for random number generator 37851
    
```

Missing Data Patterns

Group	depalim	deplim	quint	deptot	consult	pharma	Freq	Percent
1	X	X	X	X	X	X	6145	55.90
2	X	X	X	X	X	.	469	4.27
3	X	X	X	X	.	X	3522	32.04
4	X	X	X	X	.	.	856	7.79

Missing Data Patterns

-----Group Means-----						
Group	depalim	deplim	quint	deptot	consult	pharma
1	713389	1004839	3.753133	1855069	42585	94256
2	533495	741353	3.132196	1336961	62113	.
3	508459	514169	3.448609	1053241	.	30613
4	428360	457794	3.364486	886154	.	.

EM (Posterior Mode) Estimates

_TYPE_	_NAME_	depalim	deplim	quint	deptot	consult
MEAN		617854	793778	3.598799	1500591	25937
COV	depalim	337682612246	513351888202	214468	894733787611	20766914975
COV	deplim	513351888202	2.3582371E12	548566	3.0313069E12	89206204889

EM (Posterior Mode) Estimates

pharma

```

63022
22932455953
70511835792
    
```

The MI Procedure

EM (Posterior Mode) Estimates

_TYPE_	_NAME_	depalim	deplim	quint	deptot	consult
COV	quint	214468	548566	1.850414	826798	20844
COV	deptot	894733787611	3.0313069E12	826798	4.1911546E12	132126914252
COV	consult	20766914975	89206204889	20844	132126914252	17117835212
COV	pharma	22932455953	70511835792	42920	132987195053	5035988647

EM (Posterior Mode) Estimates

pharma

```

42920
132987195053
5035988647
34506910685
    
```

Multiple Imputation Variance Information

Variable	-----Variance-----			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
consult	9348.414770	1559662	1570880	9578.2	0.007193	0.007167
pharma	9348.427206	3142368	3153586	10583	0.003570	0.003564

Multiple Imputation Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits	DF
consult	25911	1253.347561	23453.73 28367.38	9578.2
pharma	63049	1775.833800	59567.57 66529.50	10583

Multiple Imputation Parameter Estimates

Variable	Minimum	Maximum	Mu0	t for H0:	
				Mean=Mu0	Pr >  t
consult	25754	26009	0	20.67	<.0001
pharma	62950	63205	0	35.50	<.0001

---

# Bibliographie

---

- Ardilly, P. (1994) *Les techniques de sondages*. Technip, Paris.
- Ardilly P. OSIER G. (2005) *Calcul de précision transversale dans l'enquête emploi en France*. Insee - Actes des Journées de Méthodologie Statistique 2005.
- Bontempi G. (?) *Resampling techniques for statistical modelling*. Département d'Informatique Boulevard de Triomphe - CP 212. <http://www.ulb.ac.be/di>.
- Carpenter J. R. Kenward, M. G. (?) *A comparison of multiple imputation and inverse probability weighting for analyses with missing data*. Medical Statistics Unit, London School of Hygiene & Tropical Medicine.
- Dell, F. D'haulfoeuille, X. Février, P. Massé, e. (2005) *Mise en œuvre du calcul de variance par linéarisation*.
- Deuxième Enquête Camerounaise Auprès des Ménages (2001) *Evolution de la pauvreté au Cameroun entre 1996 et 2001*. INS, Yaoundé.
- Deuxième Enquête Camerounaise Auprès des Ménages (2001) *Documents de méthodologies*. INS, Yaoundé.
- Deuxième Enquête Camerounaise Auprès des Ménages (2001) *Rapport d'exécution*. INS, Yaoundé.
- Deuxième Enquête Camerounaise Auprès des Ménages (2001) *Résultats*. INS, Yaoundé.
- Deville, J. C. Tillé, Y. (1998) *Unequal probability sampling without replacement through a splitting method*. Biometrika, 85.
- Dubois J.L. (1998) *Différentes approches de la pauvreté*. Contribution à la Journées des Economistes IRD dont le thème portait sur La Pauvreté, Paris. <http://kerbabel.c3ed.uvsc.edu/FIC-DDS2-C3ED-JLDB-20030702-00002.doc>
- Hurtubise D. (2003) *Estimation de la variance dans le cadre d'enquêtes complexes liées à l'utilisation de données administratives*. Assemblée annuelle de la SSC.
- Lemeshow, S. Letenneur L. Dartigues J. F. Lafont S. Orgogozo J. M. Commenges D. (1998) *Illustration of Analysis Taking into Account Complex Survey Considerations : The Association between Wine Consumption and Dementia in the PAQUID Study*. America, Journal of Epidemiology, vol 148, N°3.
- Münnich R. et Rässler S. (?) Variance Estimation under Multiple Imputation
- Verger, D. Accardo, J. Chevalier, P. Lapinte, A. (2005) *Bas revenus, consommation restreinte ou faible bien-être : les approches statistiques de la pauvreté à l'épreuve des comparaisons internationales*. Document de travail n°0503, Direction des statistiques démographiques et sociales, INSEE.
- Warszawski, J. Messiah, A. Lellouch, J. Meyer, Ll. Deville, j. C. (1997) *Estimating*

*means and percentages in a complex sampling survey : application to a french national survey on sexual behaviour (acsf)*. John wiley & sons ltd, Statistics in medicine, vol. 16, 397-423.

Winkler W. E. (2001) *Multi-way survey stratification and sampling*. Reaseach Report Series (Statistics #2001-01)Statistical Research Division, U.S. Bureau of the Census, Washington D.C. 20233.