DN 2277

PROJET MIMAP
CEDRES
BURKINA FASO

ANALYSE DE LA QUALITE DES DONNES DE L'ENQUETE PRIORITAIRE

Juin 1999

SOMDA PROSPER
KONE MICHEL
SAWADOGO SITA MALICK

1-Introduction

Les enquêtes de ménage représentent des outils de décisions importantes en termes de politiques économiques. Les données qui en sont issues sont donc d'une importance stratégique et même souvent politiques. Les statisticiens ont toujours eu le souci de mesurer la qualité des données d'enquête de ménages, et ont développé pour cela des instruments adéquats. Cela n'est pas toujours le cas par contre dans d'autres milieux où les données sont manipulées pour en tirer des conclusions fort hâtives.

Mesurer la qualité des données d'une enquête revient à mesurer l'ampleur des erreurs réalisées dans la collecte des données de l'enquête. On peut distinguer trois types d'erreurs dans les enquêtes de ménage: les erreurs d'observation, les erreurs d'échantillonnage et les erreurs d'enregistrement des données. La somme de ces trois types d'erreurs donne l'erreur totale.

Les erreurs d'observation sont les plus difficiles à mesurer. Elles peuvent être liées à plusieurs facteurs: les compétences des enquêteurs dans l'administration du questionnaire d'enquête, la fatigabilité de l'enqueté qui à certains moments donne des informations erronées, l'importance des non réponses, les erreurs d'interprétation des réponses par l'enquêteur...En ce qui concerne les erreurs d'échantillonnage, elles sont bien connues par les statisticiens dans la mesure ou elles sont liées aux difficultés liées au choix d'un échantillon représentatif de la population. Quant aux erreurs d'enregistrement des données, elles apparaissent depuis le travail des enquêteurs sur le terrain jusqu'à la phase d'analyse en passant par la codification et la saisie des données. Elles représentent avant tout des erreurs humaines qu'on peut réduire à travers des méthodes de contrôles rigoureux à tous les stades de l'enquête.

Pour apprécier la qualité des données de l'EP I, quatre (4) approches ont été utilisées :

- La comparabilité des données de l'EP avec d'autres sources de données
- L'analyse des taux de non réponse concernant certaines variables clé
- L'analyse de la cohérence de certaines données de l'enquête
- Le calcul des erreurs d'échantillonnage

2-Comparabilité des données d'enquêtes auprès de ménages

2.1- Les données de l'EDS

Les données issues de l'EDS, méritent une mention particulière. En effet en dehors des enquêtes démographiques, c'est l'une des premières enquêtes spécialisées en démographie et en santé qui adopte une méthodologie claire et bien structurée. On peut recenser les points d'analyses suivants sur cette enquête:

L'une des originalités de cette enquête a été la traduction des questionnaires dans les trois (3) langues nationales : mooré, dioula, fulfudé. De même les enquêteuses ont bénéficié d'une formation en langues nationales.

L'enquête a bénéficié du concours d'organismes spécialisés dans les domaines de la santé de la nutrition, du planning familial comme la Direction de la santé de la famille, la Direction du

contrôle des maladies transmissibles, et du Comité National de lutte contre le Sida. Ceci a permis de donner aux enquêteuses tout le background nécessaire pour comprendre les questions posées et les aspects techniques liés à l'enquête.

La note méthodologique de l'EDS est très fournie et fait apparaître toutes les informations utiles pour comprendre la procédure d'enquête et les erreurs possibles d'estimation.

Le grief que l'on peut porter à l'encontre de cette enquête est qu'elle ne fournit pas d'indications sur les revenus des ménages, ni sur les consommations des ménages. Pourtant on aurait fait des économies d'échelle en étendant l'étude à ces aspect du niveau de vie des ménages. Des informations sur le niveau de vie sont données uniquement à travers les données sur la nutrition des enfants et des mères qui sont de bonne qualité.

2.2-Analyse des données de l'Enquête Prioritaire (EP)

L'EP a connu beaucoup de problèmes dans son organisation et le déroulement du travail de terrain. Cela va du choix de la période d'enquête, à l'insuffisance ou l'imprévision des moyens matériels et financiers. Ces problèmes sont décrits dans la note méthodologique.

Le questionnaire n'a pas fait l'objet d'une large traduction dans les langues nationales. De même les résultats de l'enquête pilote n'ont pas été exploités entièrement pour permettre de reformuler autrement les questions afin de les adapter aux différentes situations. Tout cela fait que l'on peut se poser des questions sur la fiabilité des données recueillies. En effet selon Scott et Overs (1988), les erreurs commises par les enquêteurs étaient 2 à 4 fois plus élevées lorsqu'on procédait à des interprétations orales sur le terrain que lorsqu'on effectuait des traductions écrites du questionnaire.

La formation des agents enquêteurs a été très courte pour leur permettre de connaître à fond les différents aspects du questionnaire. Certaines informations concernant par exemple les mesures locales de la production agricole leur ont été fournies à la dernière minute. On a donc laissé beaucoup de liberté d'interprétation du questionnaire par les enquêteurs.

Les données sur les revenus des ménages sont les plus sujets à controverse dans la mesure ou c'est à ce niveau que l'estimation est la plus difficile. Très souvent en effet les enquêtés ont du mal à se rappeler leurs revenus annuels. Et comme le constate Lachaud J.Pierre (1997), la méthode d'estimation des revenus non salariaux et en particulier les revenus du secteur informel et du secteur rural est empreinte de beaucoup d'incertitudes.

Pour les données sur la consommation, la période de référence retenue dans l'EP est d'un mois. Pourtant il a été montré qu'il est préférable d'avoir deux (2) périodes de référence, l'une courte (2 semaines à 1 mois), l'autre longue (une année). Ceci permet de réaliser des ajustements appropriés afin d'éviter les erreurs liées à la défaillance de la mémoire de l'enquêté, et à l'instabilité de la consommation de certains ménages dans le temps.

La méthode de construction des indices de prix est approximative. L'Enquête Communautaire n'a pas permis une collecte des prix. Les données sur les prix du mil et du sorgho ont été obtenues auprès du système d'Information sur le marché céréalier du Ministère du Commerce de l'Industrie et de l'Artisanat portant sur le prix du mil et du sorgho. Les données sur les produits alimentaires périssables ont été recueillies par le service des prix de l'INSD, en

février et juillet. Un indice spatial a ensuite été construit. La dimension temporelle a été incluse en prenant en compte l'indice des prix de Ouaga en octobre 1994. Mais le fait d'avoir uniquement considéré l'inflation à Ouaga rend cette dimension temporelle trop restrictive. Et la variabilité des prix (coût de la vie) d'un lieu à un autre peut biaiser les mesures de la pauvreté.

Les résultats du questionnaire communautaire n'ont pas été présentés. Il aurait permis de disposer de données économiques de premier plan sur l'infrastructure, l'équipement...

L'EP ne fait nullement cas des biens d'équipements agricoles que ce soit dans la rubrique "production agricole" du questionnaire ou "biens des ménages". Il aurait été intéressant de pouvoir disposer de ce genre de données pour apprécier le niveau de l'équipement des ménages en milieu rural (charrues, tracteurs...). Les "achats des ménages" ne recouvrent pas non plus ce genre d'informations.

2.3- Comparaison des données de l'EP et de celles de l'EDS.

Il peut être intéressant de comparer les données recueillies sur les ménages à l'aide des deux (2) grandes enquêtes à ce sujet : l'EP et EDS. Même si ces deux (2) enquêtes ne répondaient pas aux mêmes objectifs, la méthodologie assez semblable qu'elles ont utilisé permet d'effectuer un petit rapprochement des données ayant trait à certaines caractéristiques des ménages qu'elles ont toutes les eux (2) calculé. Ceci permet d'avoir une idée sur la fiabilité de certaines données. Comme nous l'avons souligné plus haut la note méthodologique de l'EDS fournit elle même les éléments d'appréciation sur la fiabilité des estimations faites et la collecte des données.

-On a par exemple une présentation détaillée des erreurs de sondage par régions. La lecture de ces mesures montre que l'échantillon est très représentatif dans la mesure où la plupart des paramètres se retrouvent dans les intervalles de confiance à 95%. En outre par rapport au remplissage des questionnaires, on a un tableau qui présente le taux des réponses au questionnaire qui s'élève à 94%. Dans l'EP, on a aucune information de ce type. On se contente de signaler qu'il y a eu des difficultés pour obtenir des entrevues de la part des agriculteurs absents pour cause de travaux champêtres, et un refus de réponse de la part de certains chefs de ménages urbains. Cependant l'importance de ces refus n'est signalée nulle part.

En comparant les données de ces deux enquêtes, on obtient cependant des données assez compatibles sur les taux de scolarisation, les logements. A titre d'exemple au niveau du primaire l'EDS trouve un taux de scolarisation de 13.6% pour la population féminine âgée de 6 ans et plus, contre 13.8% pour l'EP.

Par contre, les données d'ensemble sur la structure par âge de la population offrent des différences importantes sur certains points. En effet si le haut de la pyramide des âges dans les deux enquêtes présente les mêmes contours, il n'en est pas de même pour la base. Dans l'EDS, les enfants âgés de 0 à 4 ans qui forment la base de la pyramide représentent la composante la plus importante de la population burkinabé (près de 20%). Par contre la pyramide des âges présentée par l'EP, laisse voir plutôt une composante majoritaire des enfants âgés de 5 à 9 ans. Les enfants âgés de 0 à 4 ans venant en deuxième position. L'EP impute ce résultat à une réduction de la fécondité ces dernières années ou à des mauvaises déclarations d'âge

conduisant à un vieillissement des enfants de 0 à 4 ans. Si la première hypothèse est plausible la deuxième tient difficilement du fait qu'on sait intuitivement que l'enquêté se souvient plus facilement de l'âge d'un jeune enfant que de celui d'un enfant d'âge avancé. De ce fait les risques de se tromper sur l'âge des enfants de 0 à 4 ans sont beaucoup plus faibles par rapport à ceux de 5 à 9 ans. Cette non correspondance des données soulève des questions importantes sur les mutations démographiques au Burkina.

Lorsqu'on observe également les données sur la nutrition des enfants, on se rend compte que les données collectées par l'EDS sont plus précises et plus complètes car elles intègrent non seulement les mesures poids-âge mais aussi les mesures poids-taille.

En conclusion, on peut dire qu'en ce qui concerne certaines données (scolarisation des ménages, type de logements...), les résultats de ces deux enquêtes sont assez comparables. Mais lorsqu'il s'agit de données beaucoup plus incertaines (âges des enfants) la correspondance est moindre. De ce point de vue la méthodologie de l'EDS basée sur un questionnaire accessible aux ménages du fait de la traduction en langues nationales, le choix d'une période d'enquête plus favorable (décembre-mars) et la qualité de la formation reçue par les enquêteuses apparaissent plus efficace. En l'absence d'éléments sur les erreurs de sondage de l'EP, l'analyse menée plus haut suggère que les erreurs de mesure ("non sampling errors") semblent beaucoup plus élevées dans l'EP que dans l'EDS. Ces erreurs étant dues justement à la gestion administrative de l'enquête, aux erreurs possibles des enquêteurs, à la fatigue du répondant (cas probable des paysans de retour des champs), au refus de collaborer, au planning inadapté.

Par contre, en termes d'opérationnalisation, l'EP l'emporte sur l'EDS dans la mesure où la désagrégation des paramètres par provinces et catégories socioprofessionnelles rend le ciblage facile et trace les lignes d'action de la lutte contre la pauvreté.

2.4- Comparaison des données sur la pauvreté

Les trois (3) principales sources complètes de données sur la pauvreté sont celles de l'INSD issues de l'Enquête Prioritaire (1994-1995), celles du projet CEDRES/Laval issues de l'enquête Permanente Agricole, et enfin celles collectées pour les travaux de Savadogo et al (1995). Une revue des données sur la pauvreté a été déjà effectuée par Taladidia et al (1997).

Cette étude montre que les données issues de ces trois (3) sources sont assez comparables. La différence entre les données de l'INSD et celles de CEDRES Laval ou de Savadogo et al (1995) étant attribuées au fait que l'enquête couvre les zones urbaines contrairement aux deux (2) autres qui se sont focalisées sur le milieu rural.

Les données quantitatives et qualitatives de Savadogo et al (1995)

Les travaux de Savadogo et al ont permis de disposer de données de premier plan sur les conditions de vie des ménages ruraux. Elles ont été recueillies à travers des enquêtes auprès des ménages. Cette étude a porté sur un échantillon assez réduit constitué de six (6) villages caractérisant les différences socio-économiques et agro-écologiques présentes au Burkina. Les trois (3) zones retenues sont respectivement le plateau central, le sud-Ouest, le Nord. Trois (3) provinces ont été retenues dans ces régions : il s'agit du Namentenga, du Kossi, et du Soum.

L'enquête a adopté deux (2) approches : une approche qualitative ou enquête participative basée sur les récits de vie et des entretiens ou l'on cherche à déterminer les perceptions des populations elles-mêmes. La seconde approche de type quantitatif est l'approche classique basée sur un questionnaires. Le questionnaire ménage a porté sur un échantillon de 300 ménages au total à raison de 50 ménages par village. Quant aux entrevues, elles ont été réalisées auprès de dix (10) personnes par villages qui sont choisis sur la base de leurs positions sociales (vieux, jeunes, personnes influentes, personnes indigentes...).

L'enquête Permanente Cedres/Laval

C'est l'une des rares enquêtes à avoir collecter des données sur une période de deux (2) ans. En ce sens les données recueillies permettent d'apprécier l'évolution de la pauvreté entre 1993 et 1994. Mais tout comme l'enquête de Savadogo et al, les données recueillies concernent uniquement le milieu rural. Les zones couvertes sont identiques à celles retenues par l'enquête de Savadogo et all. Les résultats de cette étude montrent <<une prévalence de l'insécurité alimentaire chez plus de 80°/ des ménages jugés pauvres. Sur les deux (2) périodes l'étude trouve une réduction de la pauvreté rurale entre 1993 et 1994.

L'enquête Prioritaire de l'INSD (1994-1995)

Initiée par la Banque Mondiale dans le cadre de ses travaux sur la Dimension Sociale de l'Ajustement Structurel (DSA), cette enquête est l'une des plus importantes en termes de collecte de données sur la pauvreté. Elle a porté sur 8642 ménages répartis sur toute l'étendue du territoire.

Sur la base du seuil de pauvreté absolu, l'INSD estime le nombre de pauvre au Burkina à 44,5% de la population totale. Ce nombre s'élève à 54 % lorsqu'on utilise le seuil de pauvreté relatif de 48 522 F CFA. Savadogo et al ont retenu un seuil de pauvreté relatif correspondant aux 2/3 de la moyenne de la consommation de kg/EC/EA/mois de la population. Ils ont sur cette base estimé le nombre de pauvres à 42%. Ces résultats sont presque identiques à ceux obtenus sur la base des données du CEDRES/Laval (41.5% de pauvres).

Les données issues de ces différentes enquêtes sont donc très comparables.

Tableau 1 : Comparaisons des mesures de la pauvreté individuelle modérée

SOURCE	Localité	Pourcentage de la population (°/)	Mesures		
			P0	P1	P2
CEDRES/Laval (1994) Seuil de 254 kg Equivalent-Céréales/Equivalent Adulte/an	Namentenga (Plateau Central)		38.50	11.72	6.44
	Soum (Nord)		50.70	15.24	6.97
	Kossi (Sud-Ouest)		9.20	3.48	2.15
	Nahouri		66.20	25.86	13.01
	National		**41.50**	**14.12**	**7.14**
Savadogo et al (1995) Seuil de Pauvreté = <<kg Equivalent Céréales/Equivalent Adulte>>	Namentenga (Plateau Centra)	32.89	62.87	17	6.73
	Soum (Nord)	34.17	40.37	12.4	5.46
	Kossi (Sud-Ouest)	32.93	22.85	5.4	1.766
	National	**100**	**42.00**	**11.6**	**4.65**
INSD (1994-1995) Seuil absolu de pauvreté = 41099 F	Centre Nord	23	61.2	20.9	9.5
	Sud-est	4.3	54.4	18.7	9.5
	Centre sud	24	51.4	18.7	8.2
	Nord	5.4	50.1	14.6	5.80
	Sud	8.9	45.1	14.0	5.6
	Ouest	18.5	40.1	11.9	4.8
	Autres villes	4.8	18.1	4.9	1.9
	Ouaga-Bobo	11.4	7.8	1.50	0.5
	National	**100**	**44.5**	**13.9**	**6.0**

Source : THIOMBIANO et al, 1997.

3 - Analyse des taux de non réponse

On abordera pas ici les questions portant sur le remplissage des questionnaires. En effet le mode d'administration de l'enquête a permis d'éviter certains problèmes liés au remplissage des questionnaires. A ce niveau, c'est la formule du remplacement de tout ménage absent ou non identifié qui a été adoptée selon une méthodologie bien définie. Cela a permis de disposer de questionnaires entièrement remplis.

L'analyse présente des non réponse a porté sur les données non extrapolées de l'enquête Prioritaire. Toute analyse sur les données extrapolées devrait se faire en utilisant les coefficients de pondération établis à cet effet afin d'avoir une représentativité nationale.

On distingue deux (2) types d'erreurs généralement : les erreurs systèmes qui sont représentés dans le fichier par les cases vides et les non réponses codifiées par << je ne sais pas >> sous les numéros en 9. Les cases vides dans un questionnaire désignent principalement les non concernés par la question.

L'analyse des taux de non réponse se fera en considérant uniquement les non réponses des personnes concernées par les questions.

Dans ces conditions il semble approprié d'analyser la qualité des données du point de vue des non réponse.

Certaines variables clé ont été retenues dans le cadre de l'analyse

Présence des Chefs de ménage

Les données montrent que 15,5% des ménages avaient le chef ménage absent au moment de l'enquête. L'analyse des substituts au chef de ménage montre que la plupart des substituts étaient des personnes extérieures ayant un lien de parenté quelconque avec la famille. Ceci tient aux caractéristiques culturelles des sociétés africaines ou un membre de la famille élargie est toujours associé aux décisions importantes de la famille (oncle,...).

Les enquêteurs en charge de l'EP avaient reçu une formation dans les langues. Cependant quelquefois, on a eu recours à des interprètes pour certaines questions. 10,7 % des interviews ont fait l'objet d'interprétation.

Age

Les données sur l'âge proviennent de deux (2) sources :

Il y a des cas où le ménage connaît le nombre d'années de naissance de ses membres sans avoir la date exacte de naissance. Il est clair que cette donnée ne pourra être précise et des erreurs qui porteront sur quelques mois seront fréquentes. Pour certaines variables comme les mesures anthropométriques ou il est indispensable d'avoir les données précises sur les date de naissance, cela peut sans doute créer un biais.

Le deuxième cas c'est lorsqu'on dispose de la date de naissance, alors il suffit de créer simplement une variable âge calculée à partir des informations recueillies.

L'analyse des données fournit un taux faible de non réponse en ce qui concerne les informations portant à la fois sur l'année de naissance, l'âge en mois de naissance et en année de naissance.

Ce taux est de 0,01 % soit 13 individus sur les 65014 concernés.

Santé

A ce niveau l'analyse des données montre que 100% des personnes avant déclarées être tombées malades vont en consultation. Parmi les agents consultés, la médecine moderne prédomine même si on note un recours appréciable à la médecine traditionnelle.

Au niveau des dépenses de santé, on a 5,9% de ménages soit 3823 sur 4335 qui ont déclaré n'avoir rien dépensé malgré le fait que l'un des membres soit tombé malade.

Par contre lorsqu'on considère les dépenses pour les visites médicales, on trouve que 73,3% des individus ont déclaré ne rien dépensé pour les visites médicales OF, ce qui est trop élevé. Ce taux pourrait cependant s'expliquer par la forte gratuité des consultations médicales surtout dans les zones rurales.

Education

Les données sur l'éducation montrent que 73,6 % des individus ne présentent aucun niveau d'instruction. Les informations manquantes concernent 14,3 % des individus concernés.

Revenus annuels
Tous les ménages ont déclaré avoir au moins une source de revenus

Revenus des trente derniers jours en CFA

Les déclarations de revenus ont été difficiles à obtenir surtout en milieu urbain comme l'a bien noté la note méthodologique de l'EP INSD (1995). Les réticences à ce niveau provenant le plus souvent des personnes <<intellectuelles>>.

Les personnes concernées par cette question sont les personnes occupées ou ayant occupées un emploi. Les informations manquantes concernent 19816 individus. L'information disponible porte sur 1157 individus parmi lesquels on a 25 non réponse. Ce qui donne un taux de réponse de 5,39 °/.

Les données obtenues sont sujettes à caution dans la mesure ou 63,4% des individus ont déclaré les revenus des trente dernier jours à 0 F et près de 73 % des revenus déclarés sont inférieurs à 2 400 F CFA.

Revenus agricoles

On note à ce niveau un déficit d'informations. Cependant on relève que 1919 ménages tirent leurs revenus de la culture de l'arachide. Ce taux est le plus élevé pour ce qui est des sources

de revenus agricoles. Ce résultat semble surprenant dans la mesure où la principale culture en milieu rural reste le mil, le sorgho ou le mais et la principale source de richesse reste le coton. En effet respectivement 902 ménages et 240 ménages s'adonnent à la culture du sorgho et du mais sur un total de 8642 ménages.

Données anthropométriques

Les ménages sélectionnées pour les mesures anthropométriques sont au nombre de 3010. Les non mesures ont concerné 655 ménages soit un aux de non réponse de 21,76 °/. Les raisons de non mesure sont reprises sur le tableau suivant :

Tableau 2 : raisons de non mesure

Raisons	Fréquence	Pourcentage	Pourcentage Cumulé
Absent	459	70.1	70.1
Malade	43	6.6	76.6
Refus de pesée	25	3.8	80.5
Refus de mesure	82	12.5	93.0
Autre	46	7.0	100
Total	655	100	

Source : données de base, EP.1.

Ce taux de non mesure élevé appelle à des réserves sur les données anthropométriques, d'autant plus que les mesures retenues sont parfois invraisemblables.

Conclusion

Cette analyse sommaire montre que les données de l'EP 1 sont satisfaisantes pour la plupart des variables retenues. Les réserves qui existaient sur les données portant sur les revenus sont confirmées par la présente analyse. D'autre part quelques limites ont été trouvées dans les données anthropométriques.

4- Analyse des erreurs d'échantillonnage

Les erreurs d'échantillonnage sont donc liées au fait que le choix d'un échantillon revient à exclure une partie de la population du champ d'observation. L'erreur d'échantillonnage permet donc de savoir la validité de l'interpolation des conclusions d'une enquête auprès d'un échantillon à la population totale.

Le calcul des erreurs d'échantillonnage se résume traditionnellement à un calcul des écarts-types pour certaines variables construites à partir des données portant sur un échantillon constitué de façon aléatoire. Il est estimé à partir de la racine carrée de la variance des réponses dans l'échantillon. La variance donne une idée de la précision des estimateurs. La variance est faible avec un échantillon large et adéquat L'une des propriétés de l'écart-type (ET) est qu'il permet de connaître l'intervalle de confiance dans lequel doit se trouver la vraie valeur du paramètre estimé (moyenne ou proportion) dans 95% des échantillons de taille similaire et de caractéristiques identiques. Cet intervalle est égal à] -2ET; +2ET [. Ainsi si on obtient un intervalle de confiance de]0.5; 0.60[, cela signifie qu'on a 95 % de chances que la vraie valeur du paramètre calculé (proportion de pauvres par exemple ou niveau de revenus moyen) se situe dans cet intervalle. Ainsi donc, la qualité des données d'enquête se mesurera

par rapport à la valeur de l'écart type. Un écart-type proche de zéro traduit une très bonne qualité des données. Des erreurs d'échantillonnage allant jusqu'à 10 % sont acceptables. Au delà on a une moindre qualité des données. De façon générale, plus la taille de l'échantillon est élevée, plus faible serait l'erreur d'échantillonnage.

e calcul de l'erreur type pour les sondages aléatoires simples pose peu de problèmes. Des formules simples permettent de le calculer. Ainsi si on note ET l'écart-type et Var (Y), la variance, Y la valeur du paramètre et Y la moyenne des différentes valeurs prises par le paramètre Y.

$$ET = \sqrt{Var(Y)} = \sqrt{\sum_{i}^{n} (Yi - \bar{Y})^2}$$

La difficulté apparaît des lors qu'on se retrouve avec des échantillons sélectionnés suivant d'autres méthodes de sondages à l'exemple des sondages stratifiés à double degré comme c'est le cas dans la plupart des enquêtes sur les conditions de vie des ménages (l'EP-I et l'EDS 1993). Dans ces cas précis des formules complexes permettent d'obtenir la variance et l'écart-type[1].

En désignant par r le paramètre à estimer avec r=y/x où y représente la valeur du paramètre Y pour l'échantillon total et x le nombre total de cas dans l'ensemble (ou sous ensemble) de l'échantillon. La variance r est estimée par:

$$ET(r)^2 = var(r) = \frac{1-f}{x^2} \sum_{h=1}^{H} \left[\frac{m_h}{m_{h-1}} \left(\sum_{i=1}^{m_h} Z_{hi}^2 - \frac{Z_h^2}{m_h} \right) \right]$$

Avec

$Z_{hi} = Y_{hi} - r\, x_{hi}$

$Z_h = y_h - r.\, x_h$

h = la strate allant de 1 à H

m_h= le nombre total d'unités primaires de sondages (UPS), tirées dans la hième strate

y_{hi}= la somme des valeurs du paramètre y dans l'UPS i dans la h ième strate

x_{hi}=la somme des nombres de cas dans l'UPS i dans la hième strate

f = taux global de sondage

La mise en œuvre de telles formules nécessite naturellement des logiciels sophistiqués. La recherche à ce niveau a beaucoup avancé. Si les premiers logiciels développés dans ce cadre étaient d'une manipulation ardue, les développements récents dans ce domaine ont permis d'alléger la tâche des chercheurs. On arrive ainsi à calculer en quelques heures et de façon précise l'erreur d'échantillonnage.

Parmi les logiciels utilisés à cet effet, on a le logiciel Clusters développé par l'International Statistical Institute pour l'Enquête Mondiale sur la Fécondité. Ce logiciel a été notamment appliqué par Macro-International dans le cadre de l'Enquête EDS de 1993 au Burkina. Plus récemment nous avons le logiciel IMPS développé par l'International System Team du Bureau de Recensement du Gouvernement américain pour ses recensements. C'est ce dernier qui sera

[1] EDS 1993 (op. cit)

utilisé dans le cadre de cette étude. Ce dernier logiciel grâce à sa nouvelle version IMPS 4.1 couplée à l'ancienne version IMPS 3.1 offre l'avantage d'être plus convivial[2] même s'il demeure perfectible.

Pour le calcul des erreurs d'échantillonnage de l'EP-I, nous avons utilisé le logiciel IMPS. Le guide d'utilisation pratique de ce logiciel figure en annexe de ce document.

Nous avons retenu essentiellement deux variables d'intérêt pour calculer les erreurs d'échantillonnage. Il s'agit des variables dépenses par /tête et de la variable Incidence de pauvreté P0. La variable Incidence de pauvreté est une variable binomiale prenant la valeur 1 pour les pauvres et la valeur 0 pour les non pauvres. Le seuil de pauvreté considéré est le seuil de pauvreté de l'EP-I placé à 41099 F. Les résultats des calculs sont repris dans les deux tableaux ci-dessus. Ces variables ont été calculées par rapport à différentes sous populations, les strates d'analyse, les zones (rural, urbain), le Genre (hommes, femmes) et les Groupes socio-économiques.

Les tableaux de résultats donnes différents indicateurs d'appréciation de la qualité des données. Parmi ces indicateurs, nous avons:
- La valeur de l'estimateur (estimator) qui correspond ici à l'estimation des dépenses par tête et de l'Incidence de la pauvreté . La variable dépenses par tête correspond au rapport (ratio) entre deux variables : dépenses et taille du ménage. La variable Incidence de pauvreté est obtenue également comme un rapport entre le nombre de pauvres et la population totale. Pour cela il a fallu créer une variable prenant la valeur 1 pour tous les ménages pour servir de déflateur pour la variable binomiale.
- L'écart type (standard error) et l'intervalle de confiance (confidence Interval) pour les différentes variables d'intérêt
- Le coefficient de variation (CV) ou écart type relatif qui permet d'apprécier la précision des estimateurs en termes relatifs et surtout de comparer la précision des estimateurs dans des sous populations différentes. On l'obtient en faisant le rapport de l'écart type et de la variance. Un coefficient de variation faible traduit une bonne fiabilité des données. On admet en générale qu'un CV supérieur à 20 % dénote une forte variabilité des données qui entachent leur fiabilité.
- Le Design Effect encore appelé effet de grappe permet d'apprécier l'efficience du choix d'un Plan de Sondage donné, en l'occurrence ici le sondage à double degré stratifié en le comparant à un Plan de sondage simple (aléatoire). Il est donné par le rapport entre l'écart-type observée pour le Plan de sondage considéré sur l'erreur type observée si le sondage avait été de type aléatoire simple. En d'autres termes il permet donc de savoir dans quelle mesure le Plan de sondage choisi se rapproche d'un échantillon aléatoire de même taille. Ainsi s'il tend vers un (1), alors le Plan de sondage choisi donne une erreur type similaire à celle observée dans les conditions d'un Plan de sondage aléatoire idéal.

Les deux tableaux ci-dessous reprennent les résultats des calculs réalisés à partir du logiciel IMPS.

[2] Cf en annexe; le guide d'utilisation d'IMPS.

Tableau 3: Calcul des erreurs autour de la variable d'incidence de la pauvreté

ANALYSIS TYPE: RATIOS

Analysis Ratio: PZ / DENOMINATEUR::

	Estimate (%)	Standard Error(%)	C.V. (%)	95% Confidence Interval Lower(%)	Upper(%)	Design Effect	Number of observations
Incidence globale P0	**44.5**	**1.5**	**3.38**	**41.5**	**47.4**	**7.90**	**8642**
GSE							
Salariés du sect public	2.2	1.2	52.63	-0.1	4.6	2.36	675
Salariés du sect privé	6.7	2.4	36.04	2.0	11.5	2.15	482
Artisants,commercants	9.8	1.4	14.80	6.9	12.6	1.28	1,026
Autres actifs	19.4	6.1	31.34	7.5	31.4	1.36	104
Agriculteurs de rente	50.1	3.3	6.56	43.6	56.5	3.89	486
Agriculteurs vivriers	51.5	2.1	4.00	47.5	55.5	9.99	5,154
Inactifs	41.5	3.3	8.04	35.0	48.0	3.01	715
Genre							
hommes	45.2	1.6	3.44	42.2	48.3	8.07	7,863
femme	28.2	2.8	9.93	22.7	33.7	1.49	779
ZONE							
rural	51.0	1.9	3.70	47.3	54.7	10.33	5,924
urbain	10.4	1.3	12.85	7.8	13.0	2.68	2,718
STRATAN							
Ouest	40.1	5.9	14.72	28.5	51.7	22.76	840
Sud	45.1	4.0	8.86	37.3	53.0	4.96	457
Centre Nord	61.2	1.9	3.03	57.5	64.8	2.86	1,959
Centre Sud	51.4	2.4	4.74	46.6	56.2	4.93	1,098
Nord	50.1	2.7	5.44	44.8	55.5	1.39	1,290
Autres villes	18.1	3.4	19.03	11.3	24.8	3.33	780
Ouaga/Bobo	7.1	1.1	14.96	5.0	9.2	1.69	1,938
Sud-Est	54.4	5.8	10.57	43.1	65.7	5.01	280

Tableau4: Calcul des erreurs autour de la variable dépenses par tête

ANALYSIS TYPE: SUBPOPULATION RATIOS

Analysis Ratio: **NVIE / TAILLE**

Category	Estimate	Standard Error	C.V. (%)	95% Confidence Interval Lower	Upper	Design Effect	Number of Observations
GSE							
Salariés du sect public	64,403	4,968	7.71	54,665	74,141	1.42	675
Salariés du sect privé	48,071	5,459	11.36	37,372	58,771	1.43	482
Artisants,commercants	32,479	2,420	7.45	27,736	37,222	1.54	1,026
Autres actifs	36,357	6,485	17.84	23,647	49,067	1.04	104
Agriculteurs de rente	7,632	997	13.06	5,679	9,585	6.05	486
Agriculteurs vivriers	7,349	194	2.64	6,968	7,730	1.86	5,154
Inactifs	13,428	1,238	9.22	11,001	15,854	1.03	715
Genre							
hommes	12,352	532	4.30	11,310	13,394	3.34	7,863
femme	29,994	1,992	6.64	26,090	33,899	1.22	779
ZONE							
rural	9,121	440	4.83	8,258	9,984	4.91	5,924
urbain	33,967	2,227	6.56	29,601	38,333	2.33	2,718
STRATAN							
Ouest	11,440	1,369	11.97	8,756	14,124	7.18	840
Sud	11,466	1,655	14.43	8,223	14,709	6.40	457
Centre Nord	6,130	268	4.37	5,606	6,655	1.54	1,959
Centre Sud	9,072	1,001	11.03	7,111	11,034	5.64	1,098
Nord	10,355	743	7.17	8,899	11,811	1.13	1,290
Autres villes	23,839	3,234	13.56	17,502	30,177	2.41	780
Ouaga/Bobo	38,247	2,739	7.16	32,879	43,615	2.16	1,938
Sud-Est	9,183	1,154	12.56	6,922	11,444	1.63	280

Commentaires sur les résultats:

Comparaisons intra sous populations

D'une manière générale, les données sont d'une bonne qualité. Aussi bien pour les estimations de dépenses par tête que pour les estimations sur l'incidence de la pauvreté, on trouve que les erreurs d'échantillonnage sont faibles. Elles sont dans l'ensemble inférieures à 10%. Ainsi pour l'incidence globale de la pauvreté on trouve que l'erreur faite sur l'estimation à 44.5% du nombre de pauvres au Burkina donne une erreur relative de 1.5% assez proche de 0. Pour les résultats sur l'incidence de la pauvreté dans les sous populations, les erreurs relatives sont inférieures dans l'ensemble à 5 % sauf en ce qui concerne les sous populations Autres actifs (6.1%), Ouest (5.9%), Sud-est (5.8%). Pour des zones comme Ouaga-Bobo, et les salariés du secteur public l'erreur d'estimation tend pratiquement vers 0.

Comparaison inter sous populations

Le coefficient de variation permet de comparer les différences d'erreurs entre sous populations. Les résultats généraux obtenus montrent que les estimations sont dans l'ensemble très bonnes. En effet la plupart des coefficients de variation pour l'incidence de la pauvreté sont inférieurs à 20%. Le coefficient de variation pour la population totale est assez faible de l'ordre de 3.38%. Il est supérieur à 20% pour les sous populations salariés du secteur public, salariés du secteur privé , agriculteurs de rente et agriculteurs vivriers. Cela pourrait signifier une plus forte variabilité des estimateurs dans ces sous populations comparativement aux autres sous populations. On observe d'autre part que ce coefficient est moyennement élevé pour les sous populations suivantes: les artisans commerçants, le milieu urbain, l'Ouest, les autres villes, Ouaga/Bobo. Il est faible pour tout le reste.

Par contre l'analyse des coefficients de variation pour la variable dépenses par tête montre des coefficients très faibles qui se situent tous en dessous de 20%.

On remarque par ailleurs que la variabilité des estimateurs est plus élevée pour les sous-échantillons de taille faible et plus faible pour les sous échantillons de taille élevée. Cela pourrait corroborer l'hypothèse selon laquelle plus la taille de l'échantillon est élevée plus fiable seront les estimateurs.
Il existe cependant des exceptions à cette règle. En effet pour les sous populations artisans commerçants, Urbain, Ouaga/Bobo, nous avons des tailles d'échantillon assez importants avec des CV quand même élevés par rapport à la moyenne. D'autre part pour la catégorie Urbain, le CV est faible malgré la taille réduite de l'échantillon.

L'efficience du Plan de sondage

A ce niveau on trouve que le Design Effect est particulièrement élevé pour la variable Incidence de pauvreté et ce pour la population globale (7.80%) et pour les différentes sous populations. Ceci pourrait démontrer une moindre efficacité du choix du Plan de sondage utilisé dans l'enquête EP-I comparativement à un sondage aléatoire simple. Cela signifierait que les erreurs sur les estimateurs quoique faibles sont pour la plupart dues au Plan de sondage stratifié à double degré.

Par contre pour la variable Dépenses /tête, on trouve des Design Effect plus faibles que ceux observés pour l'incidence de la pauvreté. Il restent cependant très supérieurs à 1 en dehors des sous populations sur les groupes socio-économiques ou la tendance est vers l'unité.

5-Conclusion

Cette analyse a permis de confirmer la bonne qualité des données de l'EP-I. Les erreurs d'échantillonnage sont assez faibles. Le choix du Plan de sondage pourrai être davantage amélioré pour obtenir une plus grande précision des estimateurs dans les différentes sous populations étudiées.

Annexe 1 Guide d'utilisation d'IMPS

ETAPES POUR L'UTILISATION D'IMPS

1-Travail préliminaire sur SPSS

On recherche dans le fichier maître les variables requises pour calculer les erreurs d'échantillonnage

Faire les transformations de variables sur SPSS avant de passer à IMPS car après il est difficile de faire les Transformations nécessaires
-tout ramener aux ménages (8642 ménages)
>niveau de vie /capita*taille<nvieindexée

Les variables à retenir

-numéro d'identification du ménage
-Dépenses alimentaires du ménage
-Dépenses non alimentaires totales hors loyers
-Dépenses totales –dépenses non alimentaires loyer compris
-Taille du ménage
-Equivalent adulte
-Créer la variable Strate d'Analyse consistant à aller chercher certaines provinces (STRATANA)
-Il faut avoir les codes des ZD Pour pouvoir informer l'ordinateur sur la structure échantillonnage
-On peut prendre aussi le numéro de ménage mais cela n'est pas nécessaire
-Variable type de ménage à recréer
*recréer la catégorie de ménage
*sexe du chef de ménage
variable polygame 2 femmes /trois femmes
méthode = compute (créer nouvelle variable)+recode (remplacer les missing par 0)+aggregate+break variable+sum+ add value
-GSE : groupe socio économique
-POND et TPOND = pondération = taille *Pond
-Loyer
-Dépenses
-Niveau de vie Nvie
-DEPNALT= dépenses non alimentaires
NB : il faut recréer les strates en cas de besoin

Ranger les variables (Sort) selon les G5 et G0

Veiller à formater les variables sans les virgules et e, leur donnant le format correspondant au maximum, de colonnes qu'elles peuvent contenir.

Vérifier que vos variables sont bien choisies en faisant quelques tableaux sur les fréquences ou les moyennes

Imprimer votre dictionnaire de données

Il faut sortir sur SPSS en sauvegardant le fichier transformé avec les variables retenues comme fichier ASCII.dat.

2- Travaux sur Word Pad

Ouvrir le fichier QSCII.dat sous Word pad.
Il faut vérifier qu'il n y a pas de décimales. Le fichier ne prend pas les décimales et non plus pas de . Remplacer les points (.) par des zéro (0).
-Eliminer les chiffres décimaux non significatifs
-Si les décimaux sont significatifs la solution consiste à multiplier par un multiple de 10 et de ne pas oublier après en imprimant le dictionnaire des données
-Remplacer aussi les blancs par zéro sous Word Pad.
- Sauvegarder le fichier ASCII.dat transformé dans un répertoire bien identifié.

3-Travaux sur IMPS

3.1-Passer de SPSS à fichier ASCII

3.2-Travail sur IMPS 4.1

Etape 1-Construire dictionnaire de données sous IMPS 4.1

Ouvrir d'abord la feuille du dictionnaire de données

Entrer d'abord les informations générales sur les données et le fichier

Aller ensuite aux common values

Il apparaît deux champs
-le premier indique les strates G5
-le deuxième identifiant l'unité primaire G0.

NB : il faut donc toujours communiquer la structure échantillonnale à l'ordinateur
-Il faut que le fichier soit trié par ordre strate et ZD
Sort Champ 1 (strate) champ 2 (ZD)
tri1 = par strate
tri 2 = par Unité primaire (ZD)
NB : il est possible de le faire sous SPSS et même sous IMPS

NB :IMPS calcule les variances à l'intérieur de chaque strate puis les additionne
NB : si le programme ne tourne pas rapidement alors il faut supprimer un certain nombre de variables
Type value= pas besoin d'informations à entrer
Hierachy 1
Max 1
Entrer les record item (# command items)
Entre les item label ensuite
stats= 1 (automatique)

length= Longuer du champ
Occurrence= forme de modalités, nombre de fois = toujours 1
Decimal = Nombre de décimaux
Variable strate= introduire les nombres de strates dans Value set
Variables groupes socio-économiques
Pour certaines variables, il faut introduire les value label
Etc....
-Enregistrez votre dictionnaire de données (par save)
-sauvegarder le dictionnaire sous IMPS 3.1 (extension dd) (par save as) dans le répertoire de travail de base.
-Dans Menu, s'assurer qu'il n y a pas de common items mais des record items
-Enregistrement

Etape 2

Imprimer le dictionnaire sous forme de text-viewer en allant à
File Report
File Print

Etape 3

Sauvegarder de nouveau
Sauvegarder de nouveau le même dictionnaire sous IMPS 3.1 car c'est sous celui-ci que fonctionne cenvar

Etape 4

Vérifier la cohérence des données en utilisant cross-tabulation
-en lignes : groupes socio-économiques
-en colonnes dépenses par tête
run tabulate permet de vérifier que cela marche bien,

3.3-Travail sous IMPS 3.1

Etape 5

-Aller à IMPS 3.1
On rentre dans le DOS pour arriver dans le CENVAR
CENVAR : calculer les variances
F2 Data dictionnary (Fichier dictdr)
F2 Data filemaître M.S.R.T (FICHIER TRI2)

Etape 6

Menu IMPS
-Introduire le design
a-stratum field = G5
b-Cluster field=G0 (recodification des ZD)

c-Weight field =POND ou TPOND (= poids correspondant à la structure échantillonale)
d-Two stage options =no
e-Design = yes

Etape 7

Il y a à ce niveau les analyses sur la population totale et les analyses sur les sous populations. Pour chaque catégorie, il y a trois types d'analyse / analyse sur les ratios, les loyennes, les proportions, les totaux.

Exemple:

Analysis Type= subpopulations ratios
-Analyse par strates : analyse sur les sous populations

Ex : classification var	Analysis (numerator)	Denominator
1-GSE	Nvie	taille

Il existe aussi des options comme les means

Vous pouvez choisir le format de vos résultats en utilisant les touches F

Etape 8

-Run
-Définir fichier des résultats

3.4-Lecture des résultats sur IMPS 4.1

Etape 9

-sortir de IMPS 3.1
-Aller dans IMPS4.1
-Aller dans le fichier de résultats
-Lire les résultats sous l'extension VAR dans text-viewer

Bibliographie

Macro-International et INSD: Enquête Démographique et de santé (EDS), 1993.

ASSELIN Louis Marie: Techniques de sondage avec applications à l'Afrique., CECI, 1994.

International Systems Team, Bureau of the Census, Washington, DC: Centry, IMPS Version 3.1, Users Guide; January 11, 1995.