Enquête sur les Conditions de Vie en Haiti

Conception de l’échantillon

Table des matières

Introduction 1
Exigences de l'échantillon 1
La base de sondage et l'échantillon-maître 1
Conception de l'échantillon 2
Procédures de sélection de l'échantillon 3
Sélection des UPE 3
Sélection des UPE (segmentation) 3
Cartographie et liste des ménages dans les UPE 3
Sélection des ménages 4
Sélection aléatoire d'une personne âgée de 15 ans ou plus 4
Substitution 4
Sur-échantillonnage 4
Allocation 4
Probabilités d'inclusion et poids 5
Notation 6
Sélection d'UPE, de 2 SU et de ménages 6
Probabilités de choix d'un individu 8
Poids de sondage 9
Documentation nécessaire pour l'échantillonnage 10
Vérification de l'échantillon au cours de la saisie de données 13
Non réponse et corrections pour non réponse 13
Non-réponse d'une unité : le ménage 14
Ajustement de poids et estimation de poids 15
Problèmes de sélection et RSI 16
Utilisation des poids 18
Erreurs d'échantillonnage 19
Références 22
Annexe 1 : Détermination de l'allocation de l'échantillon de ménage dans les UPE 23
Annexe 2 : Echantillonnage linéaire systématique de ménages par ordinateur 26
Mise en œuvre dans Access 26
Exemple de mise en œuvre détaillée 26

Tableaux

Tableau 1 : Correspondance terminologique 2
Tableau 2 : Allocation de l'échantillon 4
Tableau 3 : Notation utilisée 6
Tableau 4 : Structure du fichier de documentation de l'échantillon 11
Tableau 5 : Catégories de réponse 14
Tableau 6 : Notation pour l'ajustement aux non-réponse 15
Tableau 7 : Régression Logit concernant les poids relatifs aux RSI 17

Introduction

Ce document décrit l'échantillon de l'Enquête sur les Conditions de Vie en Haïti (ECVH) menée par l'Institut Haïtien de Statistique et d'Informatique (IHSI) en coopération avec Fafo. Son principal but est de documenter les procédures d'échantillonnage et celles de traitement des non-réponses dans l'enquête.

Exigences quant à l'échantillon

La conception de l'échantillon de l'ECVH était sujette comme tout autre échantillon à un certain nombre de contraintes. Les principales caractéristiques de la conception de l'échantillon étaient les suivantes :

1. La population de référence pour cette étude est l'ensemble des ménages résidant en Haïti.

2. Le budget permet un échantillon de 7000 ménages.

3. Le questionnaire fait appel à un répondant qui répond pour le ménage, à toutes les femmes composant le ménage et à une personne âgée de 15 ans ou plus sélectionnée aléatoirement.

4. L'échantillon-maître d'Haïti devrait être utilisé.

5. Un découpage en domaine d'analyse constitué d'un département ou deux au maximum devrait être possible.

6. On devrait pouvoir utiliser l'enquête en vue de créer des cartes socio-économiques d'Haïti.

La base de sondage et l'échantillon-maître

La base de sondage est la liste des Sections d'Enumération (SDE) utilisée pour le Recensement d'Haïti de 1982 telle que définie dans l'échantillon-maître d'Haïti. L'échantillon-maître (EMEM-Echantillon Maître d'Enquêtes Multiples, IHSI, 1997) définit une stratification globale et l'allocation aux strates. La structure de l'échantillon-maître est donnée à la Figure 1.

Plusieurs éléments de l'échantillon-maître méritent d'être mentionnés. Le territoire est divisé en domaines correspondant aux départements géographiques et à l'Aire Métropolitaine de Port-au-Prince. Chacun de ces domaines est ensuite divisé en strates urbaines et rurales. L'allocation de base de l'échantillon est donnée par la racine carrée de la population d'individus dans chaque domaine. Ainsi les grands domaines (départements) sont sous-représentés comparé à l'allocation proportionnelle et les petits, sur-représentés. Toutefois ceci rend possible l'analyse au niveau des départements, au moins pour quelques variables.

Chaque strate est divisée en unités territoriales qui sont soit les Sections d'Enumération (SDE) du Recensement de 1982 ou des unités nouvellement délimitées dans l'Aire

Métropolitaine. L'échantillon-maître spécifie que cette dernière devra être découpée en nouvelles unités territoriales ayant une taille d'environ 1.000 à 2000 ménages.

Dans les zones définies comme rurales dans l'échantillon-maître, la sélection des SDE s'est généralement faite en deux étapes en vue d'éviter l'énumération des ménages dans l'ensemble des 5000 SDE du Recensement de 1982. Une liste de sections communales considérées comme une liste implicite de SDE a été dressée. Ensuite les sections communales ont été sélectionnées avec probabilité proportionnelle à la taille en utilisant un échantillonnage linéaire systématique. La mesure de la taille considérée était la population des sections communales en 1997. Si la mesure de la taille de la section communale était plus grande que l'intervalle d'échantillonnage, on décompose la section communale en SDE en utilisant la définition de 1982 de la SDE et en distribuant la variation estimée de la population proportionnellement entre les SDE. On choisit alors directement la SDE. Sinon, on sélectionne une section communale, puis une SDE avec PPT dans la SDE.

Etant donné que quelques unités territoriales ainsi sélectionnées peuvent être assez larges, une étape supplémentaire d'échantillonnage est souvent nécessaire. Il s'agit de segmenter ces UP en unités plus petites, désignées sous le nom de segments. Ceci implique la réalisation d'un dénombrement rapide de chaque UP et son découpage en plusieurs segments d'environ 200 ménages chacun. On sélectionne alors un des segments avec PPT. Dans l'échantillon-maître, les segments sont dénommés « unités du premier degré » ou UPE tandis que les unités sélectionnées précédemment sont dénommées « unités supérieures » (US). Dans l'échantillon-maître, une SDE peut être soit une US soit une UPE, selon qu'elle ait été segmentée ou non. Ainsi une UPE est l'unité territoriale à partir de laquelle on construit une liste de ménages.

Dans ce document, une terminologie standardisée sera utilisée en rapport avec les différentes étapes de l'échantillonnage. La correspondance avec la terminologie de l'EMEM est établie dans le tableau qui suit :

Tableau 1: Correspondance terminologique

Terme ECVH	Terme EMEN	Notes
Domaine	Domaine	Pour les besoins de la notation, les domaines et les strates seront traités comme une seule entité.
Strate	Strate	
UPE	US/UPE	Une Section d'Enumération (SDE) de 1982, une section communale ou une SDE (grappe) nouvellement créée dans l'Aire Métropolitaine.
2SU	US/UPE	SDE ou segment de SDE (pas toujours utilisé)
3SU	UPE	Un segment d'une SDE (pas toujours utilisé)
UPE	UPE	La dernière étape de sélection d'une unité territoriale
Logement		Utilisée seulement en rapport avec la cartographie
Ménage	Ménage	
RSI		Individu sélectionné aléatoirement parmi les membres du ménage

En principe, l'EMEM définit le processus jusqu'à la sélection des segments dans le cadre de l'échantillon-maîre. Cependant une partie du travail actuel demeure parce que l'Enquête sur le Budget Consommation des Ménages (EBCM) a utilisé l'échantillon-maître et a utilisé uniquement un sous-ensemble de l'échantillon-maître.

L'EBCM a réalisé une liste de ménages et cartographié toutes les SDE rurales. Dans la plupart des cas, on n'a pas dressé de cartes ni de listes complètement nouvelles pour l'ECVH. On a plutôt procédé à une vérification sur le terrain des outils existants et apporté les modifications appropriées.

Conception de l'échantillon

Les éléments-clés de l'échantillonnage sont les suivants :

1. L'échantillon est stratifié explicitement en domaines et strates urbaines/rurales.

2. L'allocation de chaque domaine est égale à la racine carrée de la population du domaine. L'échantillon n'est pas auto-pondéré.

3. L'allocation de chaque strate d'un domaine est proportionnelle au nombre de ménages de la strate.

4. Les résultats du Recensement de 1982 tels qu'ils ont été mis à jour par l'IHSI en vue de refléter la situation de 1996 ont été utilisés comme mesure de la taille utilisée pour l'allocation d'une strate et pour la sélection d'une UPE.

5. A l'intérieur de chaque strate, on sélectionne les unités territoriales avec probabilité proportionnelle à la taille (PPT). Cette sélection peut être faite en une, deux ou trois étapes.

6. On découpe quelques UPE en plusieurs segments UPE et on en sélectionne une avec PPT.

7. On réalise un croquis pour chaque UPE sélectionnée montrant la segmentation.

8. Un croquis devrait être réalisé pour chaque UPE sélectionnée et une liste des unités de logement de l'UPE devrait être dressée.

9. A partir de la liste des UPE mises à jour, on devrait réaliser la sélection des ménages avec PPT.

10. On devrait sélectionner aléatoirement un individu âgé de 15 ans ou plus comme répondant du questionnaire « individuel ».

Procédures de sélection de l'échantillon

Sélection des UPE

La sélection des UPE a déjà été exécutée dans le cadre de l'EMEM. A l'intérieur de chaque strate, cette sélection se fait sur la base d'un échantillonnage PPT. Au total, 502 UPE ont été sélectionnées.

Sélection des UPE (segmentation)

La segmentation des SDE avait été généralement (mais pas toujours) exécutée conjointement avec l'EBCM. Dans plusieurs cas, la liste des ménages des UPE de grande taille avait été entièrement dressée et il avait été prévu d'effectuer la segmentation au bureau. Cependant, ceci n'est pas vraiment nécessaire étant donné qu'une étape supplémentaire de segmentation devrait être effectuée uniquement en vue d'éviter à établir une liste complète. A l'avenir, il serait préférable d' établir une liste rapide, d'exécuter la segmentation sur le terrain, puis établir une liste complète de ménages du segment.

Cartographie et liste des ménages des UPE

On établit la cartographie et la liste des ménages dans les UPE sélectionnées. L'objectif de la cartographie et de la liste est d'une part de permettre la sélection des ménages et d'autre part, de permettre aux enquêteurs de localiser les ménages sélectionnés.

Tous les ménages d'une UPE sélectionnée (l'UPE étant dans ce cas une SDE ou un segment) devraient être établies sur une liste, indépendamment de la taille de l'UPE.

En fait, la plupart des SDE rurales ont été entièrement cartographiées et des listes de ménage dressées. Il n'est pas nécessaire de segmenter dans ces cas, étant donné que cela ne ferait qu'introduire une étape supplémentaire dans l'échantillon. Les coordonnées géographiques exactes des UPE devraient être déterminées à l'aide d'un récepteur GPS, conjointement à la cartographie. La mesure devrait être effectuée approximativement au centre de l'UPE. La procédure est décrite dans le manuel GPS pour l'ECVH.

Sélection des ménages

Le résultat de la cartographie et du listing des UPE est une liste de ménages. La sélection de ménages est plus pratique sur cette base. La situation spécifique d'un "gran lakou" (ensemble de logements) sera traitée comme celles deplusieurs unités de logement.

On choisit les ménages à enquêter par échantillonnage linéaire systématique au bureau central à partir des listes de ménages de la base de données.

L'allocation cible pour chaque UPE est de 10 ménages dans les zones urbaines et de 20 dans les zones rurales.

Sélection aléatoire d'une personne âgée de 15 ans ou plus

L'enquêteur a la responsabilité de la sélection du RSI. Celui-ci déterminera d'abord l'appartenance au ménage à partir de la liste des membres du ménage dans le questionnaire Ménage.

Consulter le document "Sélection aléatoire d'un individu dans le ménage".

Substitution

Aucune substitution des unités sélectionnées ne devrait avoir lieu. Si pour une raison quelconque, un ménage ou une personne ne peut être interviewée, un autre ménage ou une autre personne ne devrait pas le remplacer. Néanmoins, il faut noter que la liste et la sélection seront considérées comme étant constituées d'unités de logement. Ceci signifie qu'au cours de la période entre l'établissement de la liste et l'entrevue, une famille a déménagé d'une unité de logement et une autre famille y a emménagé, la nouvelle famille devrait être interviewée.

Sur-échantillonnage

On devrait sur-échantillonner en vue de compenser la réduction de la taille effective de l'échantillon à cause de diverses sources de non-réponse. Etant donné qu'il y a peu d'informations antérieures pouvant être utilisées pour déterminer le degré de sur-échantillonnage nécessaire, le nombre de ménages supplémentaires a été fixé arbitrairement à 740 portant l'échantillon total à 7740. Le nombre 740 vient du fait que l'échantillon comprend 230 UPE urbaines et 272 rurales et que l'allocation est de 10 ménages par UPE dans les zones urbaines et de 20 dans les zones rurales. En tout, cela représente 2000 ménages en milieu urbain et 5540 en milieu rural, soit 7740.

Allocation

Sur la base des considérations précédentes, on aboutit à l'allocation de l'échantillon telle qu'elle apparaît au tableau ci-dessous :

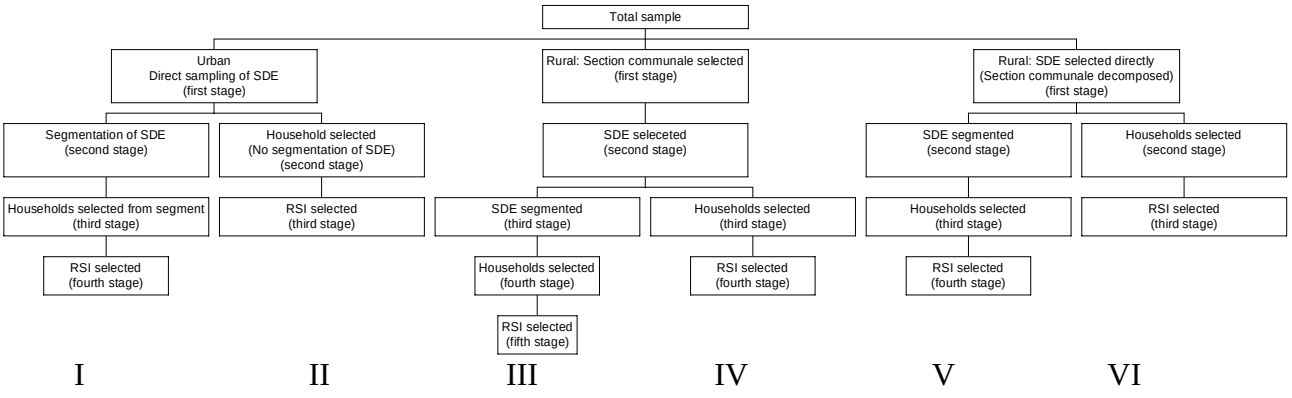
Tableau 2: Allocation de l'échantillon

			Population estimée		Somme	Rac. Carrée	Allocation			
Domaine	**Strate**	**Strate ID**	**Absolue**	**Prop-ortion**	**Domaine**	**Domaine**	**Domaine**	**Strate**	**# de UPE**	**Ménagess**
AireMetropolitaine	Faible densité	11	401,151	0.055	1,489,988	1,221	1,100	296	28	274
	Moyenne densité	12	243,556	0.033	1,489,988	1,221	1,100	180	17	168
	Forte densité	13	272,209	0.037	1,489,988	1,221	1,100	201	19	189
	Zones établies	14	243,556	0.033	1,489,988	1,221	1,100	180	17	171
	Zones d'extension	15	329,516	0.045	1,489,988	1,221	1,100	243	23	234
Reste Ouest	Urbain	21	78,545	0.011	1,082,507	1,040	938	68	10	101
	Rural	22	1,003,962	0.137	1,082,507	1,040	938	870	44	877
Sud-Est	Urbain	31	39,054	0.005	461,998	680	613	52	10	101
	Rural	32	422,944	0.058	461,998	680	613	561	28	590
Nord	Urbain	41	210,765	0.029	772,576	879	792	216	24	246
	Rural	42	561,811	0.077	772,576	879	792	576	28	561
Nord-Est	Urbain	51	63,798	0.009	252,220	502	453	114	12	128
	Rural	52	188,422	0.026	252,220	502	453	338	16	321
Artibonite	Urbain	61	239,652	0.033	1,033,370	1,017	916	212	24	259
	Rural	62	793,718	0.108	1,033,370	1,017	916	704	36	733
Centre	Urbain	71	72,739	0.010	499,538	707	637	93	12	119
	Rural	72	426,799	0.058	499,538	707	637	544	28	562
Sud	Urbain	81	94,383	0.013	662,357	814	733	105	12	109
	Rural	82	567,974	0.077	662,357	814	733	629	32	645
Grand-Anse	Urbain	91	83,362	0.011	650,997	807	727	93	12	121
	Rural	92	567,635	0.077	650,997	807	727	634	32	610
Nord-Ouest	Urbain	101	61,593	0.008	430,476	656	591	85	10	105
	Rural	102	368,883	0.050	430,476	656	591	507	28	568
Total			7,336,027	1.000	Sums of sqrt	8,322	7,500	7,500	502	7812

Probabilités d'inclusion et Poids

D'après ce qui précède, l'échantillon a un nombre de strates variables. Elles sont énumérées dans la figure ci-dessous:

Figure 1: Vue d'ensemble des strates d'échantillonnage



Comme on peut le constater, il existe 6 pistes distinctes (numérotées de I à VI) par lesquelles les ménages (et les personnes sélectionnées sur une base aléatoire) peuvent être sélectionnées. Lorsque les ménages constituent la dernière unité d'échantillonnage (ou les individus du ménage), l'échantillon est un échantillon de deux à quatre étapes. Dans le cas des Individus Sélectionnés sur une Base Aléatoire, l'échantillon est un échantillon de trois à cinq étapes. Dans les pistes I et II, c'est-à-dire les zones urbaines, les estimations de 1982 pour la taille des SDE étaient utilisées pour la sélection et de nouvelles frontières ont été définies après sélection. La SDE (UPE) était lors rapidement dénombrée et souvent segmentée (piste II). Dans les pistes III et IV, les tailles des populations estimées en 1996 dans les Sections Communales étaient utilisées comme mesure de la taille pour la sélection avec PPT. Ainsi les SDE étaient sélectionnées au cours d'une deuxième étape à partir des sections communales, encore une fois avec PPT en utilisant le nombre de ménages dans les SDE et dans la section communale en 1982 comme mesure de la taille. Les Pistes V et VI sont décrites dans la documentation de l'EMEM mais ne semblent pas avoir été utilisées dans la pratique. Ces méthodes de sélection devraient avoir été utilisées lorsqu'une section communale avait une taille qui impliquait que plus d'une SDE devrait être sélectionnée. Ainsi, un découpage de la section communale selon la taille de la SDE (en 1982) devrait ensuite être effectué avant la sélection et la SDE sélectionnée directement au cours de la même étape comme les sections communales. Cependant, les grandes sections communales ont été sélectionnées et la sous-sélection des SDE était ensuite réalisée.

Notation

Afin de décrire l'échantillon de manière précise et de calculer les probabilités d'inclusion, il nous faut introduire des éléments de notation. Ceci est fait au Tableau 3.

Tableau 3: Notation utilisée

Symbole	Signification
N	Dénombrement de ménages ou de personnes (estimation initiale)
N^l	Dénombrement sur la base de liste
N^q	Dénombrement rapide
N	Taille de l'échantillon (sur la base de l'allocation de l'échantillon)
M	Nombre d'UPE dans l'échantillon
K	Dénombrement dans le cas des 2UE
P	Probabilité d'inclusion
s	Indice relatif à la strate
c	Indice relatif aux unités territoriales (UPE,2UE,3UE (les sous-indices 2 ou 3 sont utilisés pour indiquer les étapes d'échantillonnage (on omet le sous-indice 1 pour la première étape)
k	Indice d'UPE utilisé pour la simplification lorsque l'unité se réfère à un des c, c,c_2, or c,c_2,c_3
h and I	Indice relatif au ménage (h est utilisé pour indiquer le ménage dans la phase d'échantillonnage, i variant de 1 à n dans l'échantillon pour la liste de tous les ménages)
d	Indice relatif aux personnes au sein du ménage

Sélection des PSU, 2 SU et des ménages

Les probabilités d'inclusion pour un PSU c dans une strate s sont les suivantes :

Equation 1

$$p_{s,c} = \frac{N_{s,c}}{N_s} \cdot m_s$$

Cette probabilité d'inclusion est valable pour toutes les sélections de première étape, indépendamment du type d'unité choisi. Les unités peuvent donc être des Sections Communales (zones rurales) ou des Sections d'Enumération (zones urbaines). Dans les zones rurales, la mesure utilisée pour la taille était la population en 1996. Dans les zones urbaines, la mesure utilisée pour la taille était le nombre de ménages en 1982.

La probabilité d'inclusion pour un 2SU contient deux variantes. Si une SDE est choisie à partir d'une Section communale (i.e. dans les zones rurales), dans ce cas l'équation suivante s'applique :

Equation 2

$$p_2 = \frac{N_{s,c,c_2}}{N_{s,c}} m_{s,c}$$

Dans la plupart des cas, le terme ms,c est égal à 1, i.e. une seule sélection.

(La documentation EMEM n'est pas tout à fait claire à ce point. Dans les cas où une SDE a été directement sélectionnée, l'exemple donné dans le texte (IHSI 1997, tableau 6) suggère la procédure décrite comme piste V et VI ci-dessus. La liste de l'échantillon suggère cependant que les Sections Communales sûres avaient été choisies avec « une probabilité » de plus de 1, et que en conséquence également dans ce cas on retrouve une deuxième étape de sélection où plus d'une SDE provenant de la Section Communale sélectionnée est sélectionnée).

Si la sélection de l'unité de deuxième étape (2SU) fait suite à une opération de segmentation, la probabilité est :

Equation 3

$$p_2 = \frac{N^q_{s,c,c_2}}{N^q_{s,c}}$$

où q se réfère au fait que les dénombrements rapides de ménages avec énumération devraient être utilisés. L'équation est similaire à l'Equation 2, la différence étant que le terme m dans l'Equation 2 est toujours 1 et par conséquent n'est pas inclus.

Dans les deux cas de sélection de 2SU, l'équation est une équation traditionnelle de sélection proportionnelle à la taille avec m (nombre de grappes sélectionnées) étant toujours 1. Il en est de même pour la sélection de 3SU.

La probabilité d'inclusion 3SU est actuellement similaire à celles qui viennent d'être présentées (mais le terme m n'est pas présent parce qu'il y a toujours seulement un segment sélectionné) :

Equation 4

$$p_3 = \frac{N^q_{s,c,c_2,c_3}}{N^q_{s,c,c_2}}$$

Bien sûr, l'Equation 3 et l'Equation 4 devraient être utilisées seulement lorsqu'il y a une sélection de 2SU, i.e. dans les cas où l'UPE devait être segmentée. Cependant, il devrait être utilisé indépendamment du fait que la sélection est exécutée par l'ECVH ou a été réalisée antérieurement par une autre enquête.

Dans chaque UPE, un nombre fixe de ménages doit être choisi (mais voir ci-dessus). La probabilité d'inclusion pour un ménage *h* dans un UPE *k* (indépendamment du nombre d'étapes qu'il y a eu i dans le choix de l'UPE) dans la strate *s* est alors ce qui suit :

Equation 5

$$p_h = \frac{n_{s,k}}{N^l_{s,k}}$$

Il faut noter que le nombre de ménages énumérés est utilisé au lieu de l'estimation initiale de ménages à partir du recensement. La probabilité globale d'inclusion pour un ménage i devient donc :

Equation 6

$p_i = p_1 p_2 p_3 p_h$ lorsque toutes les étapes sont présentes, ou

$p_i = p_1 p_2 p_h$ lorsque trois sont présentes, ou finalement

$p_i = p_1 p_h$ lorsque l'échantillon est un échantillon à deux étapes.

Dans la documentation de l'EMEM, p_1 et les probabilités p_2, lorsque cela est applicable sont données. (Dans la feuille de calcul qui décrit l'échantillon de l'ECVH, les probabilités qui ne sont pas applicables ont été fixés à 1 de telle sorte que la première forme de l'Equation 6 puisse être utilisée).

Probabilités pour les individus

La probabilité d'inclusion pour le RSI d parmi les N adultes du ménage h est la suivante :

Equation 7

$$p_r = p_d = \frac{1}{N_{h,d}^{\geq 15}}$$

puisque seulement un RSI est sélectionné. Cependant, comme il sera évoqué plus loin, il existe des raisons de croire que les enquêteurs n'ont pas choisi les répondants au questionnaire individuel selon la procédure prévue.

La probabilité d'inclusion totale pour un RSI est alors $p_i p_{r.}$.

Les probabilités pour l'inclusion d'autres individus (toutes les femmes, tous les enfants etc) qui ne sont pas sous-échantillonnés au sein du ménage sont les mêmes que les probabilités des ménages. Il existe cependant une exception à cette règle, principalement les hommes qui ont plus d'une femme. Dans ce cas, les femmes résident habituellement dans des ménages différents et toutes les femmes signaleront l'époux comme résident si la question leur est posée. Ceci entraîne une situation où la probabilité d'inclusion d'un homme dépend de son nombre de femmes, puisqu'il aura une chance supplémentaire d'être sélectionné pour chaque femme. Ainsi, les hommes qui n'ont pas de femme ou qui ont une femme ont la même probabilité d'inclusion que les autres personnes du ménage, mais les hommes qui ont plus d'une femme ont une probabilité différente. Cette probabilité d'inclusion peut être calculée comme l'inverse de la probabilité qu'aucun des ménages où il a une femme soit sélectionné. Une approximation est donnée plus bas (Equation 8) :

Equation 8

$$p_{h|m} = 1 - (1 - p_h)^{n_w}$$

où p_h est la probabilité d'inclusion du ménage et n_w le nombre de femmes. Ceci est seulement une approximation, parce qu'il assume que la probabilité d'inclusion pour les ménages de femmes qui n'ont pas été sélectionnés peut être estimée comme égale à la probabilité d'inclusion de ceux qui ont été sélectionnés.

En résumé, les probabilités d'inclusion sont calculées telles que définies dans le tableau suivant. Les « numéros de piste » sont ceux qui figurent dans le tableau 1.

Piste	Equation décrivant la sélection de l'unité et (la mesure de la taille)					
	Section Communale	SDE	Segment	Ménage	RSI	Hommes avec + de 2 femmes
1 (urbain)		1 (H82)		5 (L)	7 (HR)	5+8 (L,HR)
2 (urbain)		1 (H82)	3 (Q)	5 (L)	7 (HR)	5+8 (L,HR)
3 (rural)	1 (P96)	2 (H82)	4 (Q)	5 (L)	7 (HR)	5+8 (L,HR)
4 (rural)	1 (P96)	2 (H82)		5 (L)	7 (HR)	5+8 (L,HR)
5 (rural)		1 (P96/H82)	3 (Q)	5 (L)	7 (HR)	5+8 (L,HR)
6 (rural)		1(P96/H82)		5 (L)	7 (HR)	5+8 (L,HR)

Mesures de la taille: H82: Ménages tirés du Recensement de 1982, P96: Population estimée (personnes) en 1996, Q: Dénombrement rapide, L: Liste des Ménages, R: liste des membres du ménage

Poids de sondage

Il existe deux types de poids de sondage. Les poids d'expansion génèrent des estimations équivalentes aux chiffres réels de la population, tandis que les poids relatifs retiennent la taille de l'échantillon et ajustent uniquement la contribution relative de chaque unité d'analyse (ménage ou individu). Les poids d'expansion sont calculés comme l'inverse de la probabilité d'inclusion tandis que les poids relatifs sont calculés comme le poids d'expansion divisé par la moyenne de tous les poids d'expansion.

Ainsi le poids de sondage pour le ménage *i* est :

Equation 9

$$W_i^e = \frac{1}{p_i}$$

Le poids relatif de sondage est alors :

Equation 10

$$W_i^r = \frac{W_i}{\frac{\sum W_i^e}{n}}$$

Les poids de sondage ne seront pas utilisés tels quels dans l'estimation des résultats de l'enquête parce que les poids de l'échantillonnage sont ajustés pour non réponse comme on le verra plus loin.

Les poids de l'échantillonnage pour le RSI suit la même logique. Il y aura donc 2 jeux de poids d'échantillonnage, un pour le ménage et un pour le RSI.

Documentation d'échantillonnage nécessaire

L'échantillon et les différentes mesures de taille connexes doivent être documentées de manière précise. Ceci est nécessaire pour être en mesure de calculer correctement les poids. L'utilisation d'une feuille de calcul s'avère pratique pour cette tâche. Une liste de variables pour le fichier de documentation est présentée ci-après, parallèlement aux sources d'information.

Tableau 4: Structure du fichier de documentation relative à l'échantillon

Nom de la Variable	Signification	Source d'information	Nom de la variable dans le questionnaire
C01	Département	Standard IHSI	AI1
C01Name	Nom du Département	Standard IHSI	
CNO	Numéro unique PSU de l'Unité d'Echantillonnage	Standard IHSI (nouveau nom)	AI6
CODEGEO	Code géographique	Standard IHSI	AI5
STRATUM	Strate/identificateur de domaine (voir tableau d'allocation pour la numérotation	EMEM	
SNAME	Nom descriptif de la strate	EMEM	
COMMUNE	Commune (nom). Seulement pour les zones rurales	EMEM	
SECTION	Section communale (nom) Seulement pour les zones rurales	EMEM	
VILLE	Ville. Seulement pour les zones urbaines	EMEM	
PSU	Numéro de la SDE ancien style	EMEM	AI7
Secteur	Secteur d'une SDE	EMEM	
EMEMNo	Numéro séquentiel pour les grappes dans l'EMEM (de 1 à SMALLM)	EMEM	
SMALLM	Nombre de grappes dans une strate	EMEM	
UCNUM	Numéro final de grappe	Voir discussion sur l'estimation des erreurs d'échantillonnage	
MEN82	Ménages énumérés lors du recensement de 1982 dans une SDE	EMEM.SAV	
PEMEM	Probabilité d'inclusion calculée	EMEM.SAV	
MENDEPT	Ménages dénombrés lors du recensement de 1982 dans la strate	EMEM.SAV	
POP96	Population estimée dans la section communale en 1996	EMEM.SAV	
POPDEPT	Population estimée dans la strate (Département) en 1996	EMEM.SAV	
P1SPSS	Probabilité d'inclusion à la première étape comme dans EMEM.SAV	EMEM.SAV	
MENSCOM	Ménages dans la strate (section communale) en 1982	EMEM.SAV	
P2SPSS	Probabilité d'inclusion à la deuxième étape telle que donnée dans EMEM.SAV	EMEM.SAV	

P1	Probabilité d'inclusion à la première étape. Dans les zones rurales, c'est la probabilité d'inclusion de la section communale. Dans les villes, c'est celle d' une SDE	
SMALLM2	Nombre de SDE choisies dans une section communale (indiqué seulement s'il est différent de 1)	EMEM.SAV
P2	Probabilité d'inclusion à la deuxième étape. Dans les villes, elle est égale à 1 puisqu'il n'y a pas de deuxième étape de choix de SDE. Dans les zones rurales, c'est la probabilité d'inclusion de la SDE.	EMEM
NQPSU	Nombre de ménages tel qu'estimé par décompte rapide dans les SDE segmentées	Travail de segmentation sur le terrain
NQ2SU	Nombre de ménages tel qu'estimé par décompte rapide dans les segments choisis d'une SDE segmentée	Travail de segmentation sur le terrain
K2SU	Nombre de segments dans les SDE segmentées. Est égal à 1 s'il n'y a pas de segmentation. Seulement pour information (non utilisé dans les calculs sauf pour déterminer si la probabilité d'inclusion doit être fixée à 1 ou calculée)	Travail de segmentation sur le terrain
MODELIST	Mode d'énumération A : re-énumération complète, V: Vérification	
SEL2SU	Type de sélection d'une 2SU, 1: aléatoire simple , 2=PPT, 3= Certitude	
P3SEG	Probabilité d'inclusion d'un segment. Calculé selon la formule (NQ2SU*1)/NQPSU (le facteur 1 indique qu'il y a un seul segment choisi). S'il n'y a pas de segmentation, elle est fixée à 1.	Calculé dans la feuille de calcul
N2SUL	Ménages tels qu'énumérés dans l'UPE c de la strate s	Cartographie et liste
SMALLN	Taille de l'échantillon (tel que fixé par l'allocation d'échantillonnage)	Déterminé au bureau
P4	Probabilité d'inclusion d'un ménage au sein de la grappe. Calculée selon la formule	Calculé dans la feuille de calcul

	SMALLN/N2SUL	
PFINAL	Probabilité globale d'inclusion d'un ménage au sein de la grappe. Calculé selon la formule P1*P2*P3SEG*P4	Calculé dans la feuille de calcul
EW	Poids d'Expansion de l'échantillon (calculé selon 1/PFINAL)	Calculé dans la feuille de calcul
PAVER	Probabilité d'inclusion moyenne dans la strate	Voir Annexe 1 (calculé dans la feuille de calcul)
ADJUSTNA	Allocation intermédiaire ajustée SMALLN*(PAVER/PFINAL)	Voir Annexe 1 (calculé dans la feuille de calcul)
SUMADJ	Calcul intermédiaire en vue de l'ajustement final de l'allocation dans la grappe	Calculé dans la feuille de calcul
NESMALLN	Allocation ajustée finale	Voir Annexe 1(calculé dans la feuille de calcul)
PADJU	Probabilité d'inclusion étant donné l'allocation ajustée (voir PFINAL mais avec P4 remplacé par NESMALLN/N2SUL)	Calculé dans la feuille de calcul
EXPWSAMP	Poids de sondage étant donné les probabilités d'inclusion avec ajustement d'allocation	Calculé dans la feuille de calcul
LATITUDE	Latitude de l'UPE	Cartographie (GPS)
LONGITUDE	Longitude de l'UPE	Cartographie (GPS)

Vérification de l'Echantillon au cours de la saisie des données

Il convient d'entrer toute l'information concernant l'identification de l'échantillon sélectionnée dans un fichier qui peut être utilisé comme un fichier de vérification au cours de la saisie des données. Ceci garantit qu'il n'y a pas de duplication des codes d'identification et que toute incohérence entre l'allocation de l'échantillon et les entrevues soit mise en évidence le plus tôt possible.

Non-réponse et corrections aux non-réponses

Le taux de réponse obtenu au cours du travail de terrain est crucial pour la qualité des résultats de l'enquête. Lorsque les taux de réponse sont faibles, on peut raisonnablement suspecter des biais dans les résultats.

En général, on peut distinguer entre deux types de non-réponse : non-réponse d'une unité et non-réponse à une question. Le premier concerne la non-réponse d'une unité entière, telle qu'un ménage. Dans ce cas, on ne sait quasiment rien du ménage.

La non-réponse à une question fait référence au manque d'information concernant une question spécifique relative à une unité ; par exemple, une personne ne répond pas aux questions sur les revenus. On ne considère dans ce document que la non-réponse d'une unité.

Non-réponse d'une unité : Le ménage

Les résultats des entrevues ou des tentatives d'entrevues peuvent être analysées en utilisant une classification assez détaillée des non-réponses dans le questionnaire, tiré de Hidiroglou, Drew et Gray (1993). Le tableau 5 indique les catégories de réponse.

Le cadre a été élaboré sur la base de l'observation qu'une entrevue peut être manquante pour deux raisons. Tout d'abord, il est possible que le ménage sélectionné n'appartienne pas à la base de sondage. C'est le cas, par exemple, des diplomates. De plus, un ménage sélectionné qui existe en fait et est éligible, peut refuser ou peut être absent de chez lui. La classification doit aussi tenir compte de situations éventuelles où l'enquêteur ne peut pas déterminer si un ménage existe ou pas. Les enquêteurs se trouvent quelquefois dans la situation où le ménage est disponible pour l'entrevue, mais qu'il n'arrive à obtenir aucune information parce que le répondant est malade ou autrement incapable de répondre.

Tableau 5: Catégories de réponses

Catégorie	**Type de résponse**
1 Entrevue terminée	L'entrevue est possible (réponse)
2 Refus converti par le superviseur (le répondant a refusé initialement, mais a coopéré après une visite du superviseur)	L'entrevue est possible (réponse)
3 Partiellement terminée	L'entrevue est possible (réponse)
4 Statut indéterminé (L'équipe de terrain n'a pas pu déterminer si un ménage résidait à cette adresse)	Imprécise, en général réparti entre des entrevues possibles et impossibles
5 Pas d'information utilisable (par exemple, parce que le répondant était malade, mentalement malade, pas vraiment coopératif)	L'entrevue est possible, non-réponse
6 L'unité de logement n'existe pas	Aucune entrevue possible
7 L'unité de logement est vide	Aucune entrevue possible
8 L'unité de logement est en construction	Aucune entrevue possible
9 Non éligible	Aucune entrevue possible
10 Pas de contact (le ménage existe, mais n'était pas présent)	L'entrevue est possible, non-réponse
11 Refus	L'entrevue est possible, non-réponse

Correction pour non-réponse

Les non-réponses se produisent toujours. Cependant, étant donné que le degré et la gravité de la situation de non-réponse varient, le plan de la correction pour non-réponse doit être examiné à nouveau après le travail de terrain.

.

Ajustement de poids et poids estimés

Dans le cas de non-réponse d'une unité, l'utilisation directe des poids de sondage aboutira à des estimations biaisées. De façon générale, celles-ci prennent deux formes. Dans le premier cas, lorsqu'on doit estimer les totaux à l'aide des poids d'estimation de sondage, le total sera trop petit parce que la non-réponse implique que des unités qui devraient être ajoutées au total sont manquantes. Dans l'autre cas, l'estimation peut être biaisée parce que les unités n'ayant pas répondu peuvent avoir des caractéristiques particulières.

Une manière de réduire les biais produits par les non-réponses d'unité est d'ajuster les poids de sondage. La méthode de correction des poids pour tenir compte des non-réponses utilisées dans le cadre de cette enquête est celle désignée sous le nom de « méthode d'ajustement des cellules » (voir par exemple, Lehtonen et Pahkinen 1995; Little et Rubin 1987). Dans le cadre de cette approche, on identifie les ménages considérés comme étant assez similaires et le taux de non-réponse calculé pour chaque groupe de ménages, désignées sous le nom de cellules d'ajustement. Dans ce sens, lorsque les taux de non-réponse sont calculés, on ne considère que les non-réponses de ceux qui auraient pu répondre mais, pour une raison quelconque, ne l'ont pas fait.

L'inverse du taux de non-réponse dans chaque cellule d'ajustement était alors utilisée en vue d'ajuster les poids de sondage (à la fois d'expansion et relatifs) pour chaque ménage. On obtient alors ce qu'on appelle les poids estimés, à la fois d'expansion et relatifs. La taille de l'échantillon pondéré correspond à la situation où tous les ménages avaient répondu. Il en résulte une augmentation de la contribution relative aux estimations des unités similaires à celles manquantes.

Dans notre cas, les cellules d'ajustement utilisées seront constituées probablement d'UPE adjacentes géographiquement.

En utilisant la notation du Tableau 6, le facteur de correction aux poids pour tenir compte des non-réponses est donné pour l'Equation 11.

Tableau 6: Notation pour l'ajustement aux non-réponses

Symbole	Explication
C	Facteur d'ajustement (correction)
a	Indice d'ajustement de cellule
h^r	Ménages ayant répondu
h^f	Ménages n'ayant pas répondu

Le nombre d'entrevues possibles (i.e. le dénominateur dans le taux de non-réponse) est la somme des catégories 1,2,5,10 et 11 dans le Tableau 5. Le nombre d'unités n'ayant pas répondu est la somme des catégories 5, 10 et 11. La catégorie « Statut indéterminé » peut être répartie à travers les autres catégories.

Equation 11

$$C_a = \frac{1}{\frac{h_a^r}{h_a^r + h_a^f}}$$

On ajuste alors les poids selon les équations suivantes :

Equation 12

$$W_i^{estimation} = C_i W_i^e$$

Equation 13

$$W_i^{r,estimation} = \frac{W_i^{e,estimation}}{\frac{\sum W_i^{e,estimation}}{n}}$$

L'effet des corrections est d'augmenter les poids d'expansion de telle sorte que la somme de ces poids correspond à la somme des unités dans la base de sondage (moins les unités non existantes ou non éligibles). Les poids relatifs sont normalisés. Ceci signifie que la somme des poids est égale à celle des ménages dans le fichier de données.

Problèmes de Sélection et les RSI

L'examen des fichiers de données après le travail de terrain a révélé que les RSI ont été sélectionnés différemment de ce qui a été prévu. En particulier, les enquêteurs ont sélectionné les femmes plus souvent que les hommes, et il existe aussi des biais en ce qui concerne l'âge. De tels problèmes sont courants et peuvent également être corrigés dans une certaine mesure. Une procédure similaire à celle de la correction pour non-réponse aurait pu être appliquée. Cependant, ceci pourrait facilement mener à un très grand nombre de cellules d'ajustement, avec des totaux de cellules correspondants de faible valeur. Les procédures de sélection apparemment utilisées par les enquêteurs ont été plutôt modélisées en estimant un modèle de régression logit où la variable dépendante est la probabilité d'être sélectionné.

Equation 14

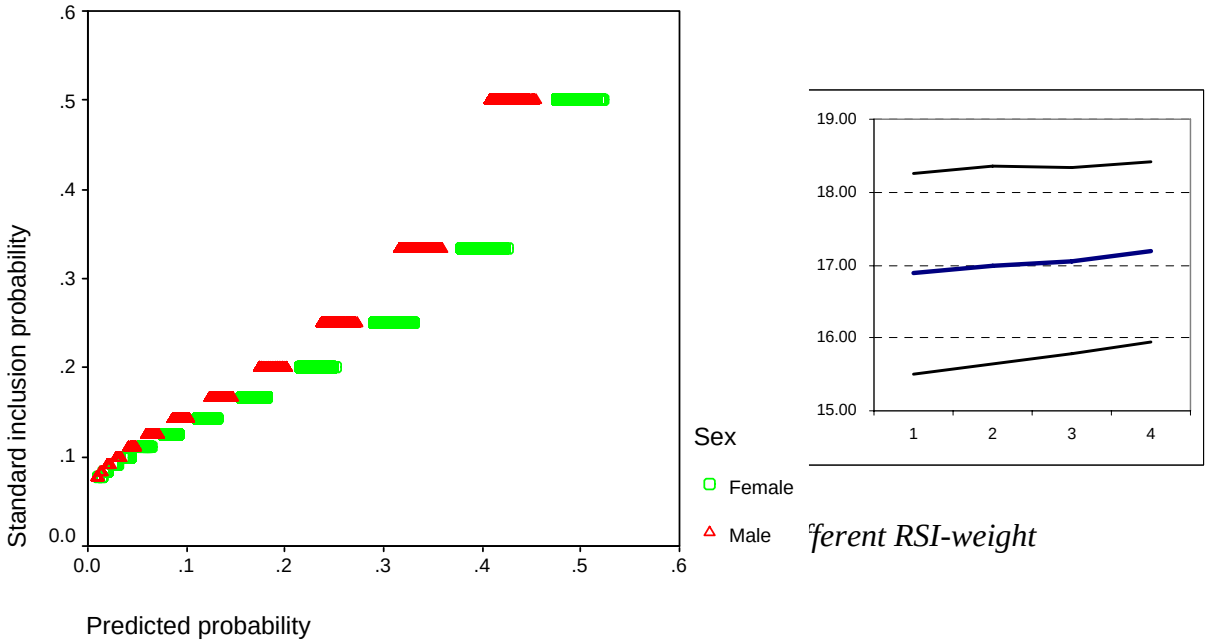
$$\log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

A partir de ce modèle, la probabilité d'inclusion des RSI dans les ménages avec deux personnes ou plus éligibles comme RSI a été estimée en résolvant l'équation, avec p comme inconnu. Les variables dépendantes (x_i) sont le nombre de personnes éligibles comme RSI dans le ménage, l'âge, l'âge au carré et le genre. Les résultats de la régression sont indiqués au Tableau 7.

Tableau 7: Régression Logit des poids reltifs aux RSIs

	B	S.E.	Wald	Sig.	Exp(B)
Sexe (Homme)	-.274	.032	71.648	.000	.760
Age	-.012	.004	8.869	.003	.988
Age au carré	.011	.004	6.052	.014	1.011
Nombre eligible/100	-.396	.012	1061.900	.000	.673
Constant	1.034	.099	108.093	.000	2.812

La figure ci-dessus montre la relation entre les poids tels que calculés directement en utilisant le nombre de members éligibles et les poids tels que modélisés. Comme on peut le voir, les femmes sont, de façon générale, modélisées avec des poids plus élevés que dans le cas hommes.



ferent RSI-weight

Figure 2:Probabilités d'inclusion prévuess et estimées

Etant donné les probabilités d'inclusion inégales des ménages et le fait que la sélection d'une personne parmi les membres éligibles produise des probabilités d'inclusion plutôt variables (de 1 à 0.08), la variation totale des poids relatifs aux RSI est considérable.

Figure 1: Estimations approximatives et intervalles de confiance pour différentes corrections de poids relatifs aux RSI

Une grande variation dans les poids contribue à la variance totale d'une estimation; en général, la variance augmente par un facteur $1+CV^2$, où CV est le coefficient de variation,

i.e. l'écart-type du poids divisé par sa moyenne (Groves et Couper 1998:318). Dans le cas des poids relatifs aux RSI calculés tel qu'indiqué précédemment, ce facteur est égal à 2.46. Autrement dit, les poids augmenteront la taille d'un intervalle de confiance d'environ 57%.

En raison de l'augmentation comparativement grande, la valeur des poids les plus extrêmes a été réduite. Ceci a pour effet de créer un biais léger, mais la diminution de l'intervalle de confiance de l'estimation a pour conséquence un gain net en précision. La figure ci-dessous montre l'effet de quelques schémas dans la réduction de la variabilité des poids, numérotés de 1 à 4 sur trois variables : le fait d'avoir participé à des élections, d'avoir écouté une émission de radio et l'indication par un enquêté de difficulté de vision nocturne.

Les schémas de poids étaient : 1 pas de correction, 2 : fixer les 5 valeurs les plus extrêmes de de part et d'autre de la distribution à la 5ème valeur la plus extrême de part et d'autre, 3 et 4 : Corriger en fixant le nouveau poids à l'aide de la formule :
by setting the new weight using the formula:

$$W_c = \begin{cases} W, W < z \\ z + \log(W - z), W > z \end{cases}$$

où z est une limite fixée à 2400 dans le cas de la méthode 3 et à 1200, pour la méthode 4. On a tilisé la méthode 3, étant donné qu'elle semble donner un compromis entre biais et variance.

Utilisation de poids

Etant donné que l'enquête n'est pas auto-pondéré, les poids estimés doivent être utilisés pour toutes les tabulations et estimations concernant l'enquête. De façon générale, les règles d'utilisation des poids sont les suivantes:

a) Les poids d'expansion devraient être utilisés lorsqu'on désire obtenir les chiffres de la population totale, par exemple le nombre de chômeurs, le nombre d'enfants de moins de 5 ans ou des paramètres similaires.
b) On peut utiliser les poids d'expansion ou les poids relatifs lorsqu'on calcule des pourcentages et des moyennes.
c) On doit utiliser les poids relatifs pour les analyses multivariées avec SPSS, mais il est préférable d'utiliser un programme tel que SUDAAN ou Stata pouvant gérer le plan d'échantillonnage complexe et aussi parce que l'échantillon est loin d'être auto-pondéré. Soulignons qu'avec SPSS, les poids relatifs donnent le nombre total de cas de façon correcte lorsque le fichier complet est utilisé, mais pas nécessairement pour les sous-groupes. En particulier, le nombre de cas utilisés pour SPSS pour le calcul des écarts-type sera nettement incorrect si on considère uniquement les zones urbaines ou uniquement les zones rurales.
   Avec SUDAAN ou STATA, on peut utiliser les poids d'expansion ou les poids relatifs, mais il est plus facile de travailler avec les premiers.
d) Le poids relatif au ménage (d'expansion ou relatif) devrait être utilisé pour les analyses liées au ménage.
e) Le poids relatif aux membres du ménage devrait être utilisé pour des membres du ménage.
f) Le poids relatif au RSI (d'expansion ou relatif) devrait être utilisé pour les analyses liées à l'Individu Sélectionné Aléatoirement.

Le tableau suivant fournit les noms des variables utilisés pour les différents poids dans les différents fichiers.

Nom	Pour le fichier	QUOI
expweigh	MAIN	Poids d'expansion lié au ménage
relweigh	MAIN	Poids lié au ménage
expweigr	Roster	Poids d'expansion lié aux membres du ménage
relweigr	Roster	Poids d'expansion lié aux membres du ménage
expwrsi	Fichier RSI	Poids d'expansion lié au RSI
relwrsi	Fichier RSI	Poids d'expansion lié au RSI

De façon générale, les fichiers constitués d'enregistrements qui découlent du ménage sans sous-échantillonnage devraient utiliser les poids liés au ménage. Le seul cas impliquant un sous-échantillonnage est le fichier RSI. Cependant, comme mentionné précédemment, il y a quelques complications concernant la liste des membres du ménage à cause des épouses multiples.

Erreurs d'échantillonnage

Les erreurs d'échantillonnage dans le cas d'une enquête avec une conception complexe, ne peuvent être calculées en utilisant les logiciels statistiques ordinaires tels que SPSS, parce que de tels logiciels supposent un échantillonnage aléatoire simple.

Cependant les erreurs d'échantillonnage peuvent être calculés en utilisant SUDAAN (Shah, Barnwell et Bieler 1997) ou STATA. Ces logiciels prennent en compte le plan de sondage dans le calcul des erreurs d'échantillonnage, en utilisant (dans ce cas) une approche de linéarisation à l'estimation.

La spécification de l'échantillon de façon à pouvoir calculer correctement l'estimation de la variance « va de soi ». Il est pratique de considérer l'échantillon comme un échantillon en grappes stratifié avec remise lors de la première étape. Ceci n'est pas strictement exact, mais simplifie considérablement les calculs. La conséquence la plus importante de cette hypothèse est que les variances calculées sont légèrement supérieures aux variances réelles. L'inflation est néanmoins négligeable. Une deuxième hypothèse simplificatrice consiste à ignorer les étapes de l'échantillonnage, autre que la première, en utilisant l'approche dénommée « grappe ultime». Une fois encore, ceci affecte légèrement la précision des résultats. On considère l'unité sélectionnée lors de la première étape de sélection aléatoire comme étant la grappe ultime.

Trois élém,ents d'information sont nécessaires en vue de spécifier la conception de l'échantillon : l'identification de la stratification de l'échantillon, la définition de la

grappe ultime et les poids ???. Lorsque le RSI est l'unité d'analyse, on doit utiliser les poids estimés relatifs aux RSI ; si l'unité d'analyse est le ménage ou l'individu, on doit utiliser les poids estimés relatifs aux ménages. La spécification des strates et des grappes ultimes n'est pas affectée par le choix de l'unité d'analyse.

Les grappes ultimes dans l'échantillon de l'ECVH sont soit des sections communales (dans les zones rurales) soit des SDE. En pratique, cela signifie que la grappe ultime est toujours une UPE, sauf dans quelques cas lorsqu'on choisit plus d'une UPE à partir d'une même section communale. L'ajout de la variable UCNUM dans le fichier d'allocation de l'échantillon reflète cela.

Pour estimer les erreurs d'échantillonnage, on doit trier le fichier de données par les variables STRATUM et UCNUM, et appliquer les poids appropriés. Si on désigne par RELWEIGH le poids estimé lié au ménage, la syntaxe correcte SUDAAN serait:

```
/* ============================================================ */
/* Fichier:ECVH01.SUD                                          */
/* But:     Example:    Erreurs d'échantillonnage pour le fichier de ménages ECVH
*/
/*         Tabulation croisée                                   */
/* Projet:    ECVH                                               */
/* Ecrit par: Jon Pedersen                                       */
/* Date:       10 MAR 2001                                         */
/* Fichiers utilisés: ECVH_HH.SAV (nom de l'exemple)                          */
/* ============================================================ */

PROC CROSSTAB DATA=ECVH_HH FILETYPE=SPSS DESIGN=WR DEFT1;
NEST STRATUM UCNUM ;
WEIGHT RELWEIGH;
SETENV LINESIZE=132 PAGESIZE=72;
TITLE "Sampling errors for ECVH household file";
SUBGROUP HA4 C01;
LEVELS   2      9 ;
TABLES  HA4 * C01 ;
PRINT    COLPER="Percent"    SECOL="Standard    Error"    DEFFCOL="Design    effect"
NSUM="Observations"
 /COLPERFMT=F5.1 STYLE=NCHS;
```

Le programme génère des écarts-types pour l'électricité (HA4) par département (C01).
(Les commandes PROC, NEST et WEIGHT incluent les spécifications de l'échantillon).
La syntaxe correspondante STATA serait alors:

```
/* ************************************************************/
/* Fichier:       ECVH01.DO                                     */
/* But:    Example:  Erreurs d'échantillonnage le fichier de ménages ECVH  */
/*             Tabulation croisée                                */
/* Projet:    ECVH                                              */
/* Ecrit par: Jon Pedersen                                      */
/* Date:       10 MAR 2001                                       */
/* Fichier utilisé: ECVH_HH.DTA (nom de l'exemple)                        */
/* ************************************************************/
clear
set memory 4m
use "D:\Haiti\ECVH\data\ECVH_HH.dta",clear
svyset strata stratum
svyset psu UCNUM
svyset pweight relweigh
generate elect=0
replace elect=100 if HA4==1
svymean elect,by(C01) ci deff
```

(La syntaxe Stata suppose qu'il n'y a pas de valeurs manquantes pour les variables)

Ecarts-type et les départements comme domaines d'analyse

Le nombre observé de ménages dans les départements varie entre 385 et 1985. Ceci rend difficile l'analyse au niveau des départements dans certains cas, à cause des écarts-types élevés. Ceci est particulièrement vrai pour quelques variables pour lesquelles on doit s'attendre à ce que la corrélation inter grappes soit très grande. Le tableau suivant relatif à l'accès à l'électricité illustre le fait que pour les variables liées aux infrastructures, il faut prendre beaucoup de précaution. Les écarts-types sont très grands. Pour les autres variables, où on s'attendrait à des corrélation inter-grappes faibles, il devrait être possible de faire des analyses au niveau du Département.

Tableau 8:Ecarts-types et effets de conception pour avoir l'électricité.

Département	Pourcentage	SE	Effet de conception	Observations
Tous les départements	27.3	1.80	11.65	1727
Aire Métropolitaine	58.9	4.32	19.80	1117
Sud-Est	8.4	2.93	3.76	54
Nord	11.6	2.75	4.84	119
Nord-Est	4.3	1.32	1.22	23
Artibonite	9.2	2.35	9.71	155
Centre	10.5	2.73	2.77	63
Sud	14.4	3.47	4.91	86
Grand-Anse	7.7	3.38	10.20	63
Nord-Ouest	8.6	2.23	2.05	47

Références

Groves, R.M. et M.P. Couper. 1998. *Nonresponse in Household Interview Surveys.* New York: John Wiley.

Hidiroglou, M., J. Drew og G. Gray. 1993, The measurement of non-response in surveys. *Survey Methodology* 19: 81-94.

IHSI. 1997, Echantillon-Maître d'Enquêtes Multiples (EMEM). Port-au-Prince: IHSI

Kish, L. 1965, *Survey sampling*. New York: Wiley.

Lehtonen, R. and E. J. Pahkinen. 1995, *Practical methods for design and analysis of complex surveys*. Chicester: Wiley.

Little, R. and D. Rubin. 1987, *Statistical analysis with missing data*. New York: Wiley.

Shah, B.V., B. Barnwell, G.S. Bieler 1997, *SUDAAN User's Manual, Release 7.5.* Research Triangle Park, NC: Research Triangle Institute.

Annexe 1
Détermination des allocations d'échantillon de ménages dans les UPE

Dans le cas d'un échantillon à deux degrés avec échantillonnage PPT lors de la première phase (ou à 3 degrés avec échantillonnage PPT lors des première et deuxième phases), il peut arriver qu'on tire un nombre constant d'unités à la dernière phase. Ceci mènera en pratique à des probabilités d'inclusion inégales si la mesure de la taille (MDT) initiale de l'UPE est différente de celle de la liste d'unités d'échantillonnage de la deuxième phase. Une façon de conserver l'auto-pondération est d'ajuster l'allocation initiale des unités secondaires par un facteur au ratio du nombre d'unités de la liste à la mesure de la taille initiale. L'équation qui suit présente la formule correspondante. En utilisant la notation définie précédemment, en y ajoutant le superscript t pour indiquer l'allocation cible, on obtient

Equation 15

$$n_{s,c} = n^{t}_{s,c} \cdot \frac{N^{l}_{s,c}}{N_{s,c}}$$

Par exemple, si l'allocation cible pour l'unité secondaire d'échantillonnage 2SU était de 20, et la MDT originale de cette unité de 130 et si la taille de la liste était de 136, le taille de l'UPE sera de 20 136/130 ou 21 en arrondissant.

Le problème est que l'application de cette formule mènera à changer globalement la taille de l'échantillon de l'enquête. Une façon d'éviter ce problème est donnée par l'équation ci-dessous, dans laquelle la correction est ajustée en référence au changement de la taille totale de la population.

Equation 16

$$n_{s,c} = n^{t}_{s,c} \cdot \frac{N^{l}_{s,c}}{N_{s,c}} \cdot \frac{\sum N_{s,c}}{\sum N^{l}_{s,c}}$$

Par exemple, si la taille totale initiale de l'UPE sélectionnée est de 120356 ménages, et la somme des tailles selon les listes est de 125301 ménages, l'équation devient (en utilisant les chiffres de l'exemple précédent):

$$n_{s,c} = 20 \cdot \frac{136}{130} \cdot \frac{120356}{125301}$$

ou 20, en arrondissant le résultat. L'allocation reste inchangée dans ce cas. Ceci s'explique du fait que la différence proportionnelle entre l'estimation initiale et la taille selon la liste de l'UPE est proche de l'augmentation relative moyenne de la taille de l'univers des ménages. (On peut utiliser alternativement la somme des allocation des nouvelles cellules non corrigées divisée par la somme des cellules corrigées comme facteur de correction).

Une façon équivalente et alternative d'obtenir le même résultat comme dans l'Equation 13 est d'utiliser les probabilités d'inclusion au lieu des tailles de population. On a dans ce cas

Equation 17

$$n_{s,c} = n^{t}_{s,c} \cdot \frac{p_{n,c,h}}{\frac{1}{m}\sum p_{n,s,h}}$$

qui représente l'allocation initiale multipliée par le résultat de la division de la probabilité d'inclusion initiale des ménages de l'UPE par la moyenne de ces probabilités dans la strate. Ceci est préférable pour ce qui concerne l'EMEM, parce que les mesures de taille initiales n'ont pas été conservées dans la documentation de l'échantillon.

Cependant, l'Equation 17 partage avec l'Equation 15 la caractéristique qu'elle change l'allocation globale de l'échantillon. Ceci peut être rectifié comme dans le schéma précédent, à l'aide de l'équation suivante:

$$n_{s,c} = n^{a}_{s,c} \cdot \frac{\sum n^{t}_{s,c}}{\sum n^{a}_{s,c}}$$

où le supscript *a* indique la valeur initiale ajustée.

Annexe 2: Abbréviations

SDE	Section d'Enumeration
US	Unité Supérieure
UPE	Unité du premier degré
PSU	Unité Primaire d'Echantillonnage (US, SDE)
2SU	Unité d'Echantillonnage du 2ème degré
3SU	Unité d'Echantillonnage du 3ème degré
MOS	Mesure de la taille
PPT	Probabilité Proportionnelle à la Taille
RSI	Personnes Sélectionnées sur une base Aléatoire

Annexe 3:

Echantillonnage linéaire systématique de ménages à l'aide de l'ordinateur

Il existe plusieurs façons de tirer des échantillons linéaires systématiques. De façon générale, elles consistent à :

a) Disposer d'une liste d'unités à être sélectionnées avec la taille N

b) Connaître le nombre (n) d'unités à être sélectionnées

c) Calculer un intervalle d'échantillonnage k = N/n

d) Tirer un nombre aléatoire (r) compris entre 0 et k

e) Sélectionner la première unité comme étant celle sur la liste d'ordinal o vérifiant la condition $o-1 < r \leq o$

f) Ajouter séquentiellement l'intervalle k à r de telle sorte que la prochaine unité sélectionnée vérifie la condition $o-1 < (r+k) \leq o$, la suivante $o-1 < (r+k+k) \leq o$ et ainsi de suite. (En général, une unité est sélectionnée si $o-1 < (r+(i-1k) \leq o$, i variant de 1 au nombre d'unités à être sélectionnées.

La procédure marche avec des nombres réels ou des entiers, mais l'avantage d'utiliser un ordinateur est que on peut utiliser facilement les nombres réels. Ceci élimine le problème rencontré couramment dans l'échantillonnage linéaire systématique utilisant des entiers, à savoir que le processus de sélection dépasse la longueur de la liste ou est plus court.

Implémentation sur Access

Ce qui suit décrit une méthode d'implémentation de la sélection de l'échantillon sur Microsoft Access 2000. Le résultat de la sélection peut être présenté sous forme d'une liste de tous les ménages choisis parmi chaque grappe. Cela présuppose que les résultats de l'énumération des ménages sont stockés dans une base de données relationnelle. En ce qui concerne notre exemple, on suppose qu'il y a 2 tableaux à l'intérieur de la base de données:

a) Un tableau au niveau de la grappe (UPE) fournissant l'information relative à la grappe dans son ensemble. Chaque enregistrement inclut l'identification, la localisation, le nombre de ménages et la taille de l'échantillon (le nombre de ménages à être sélectionnés)

b) Un tableau au niveau du ménage donnant l'information relative à chaque ménage. Chaque enregistrement inclut l'identification de la grappe, celle du ménage et des variables relatives au ménage telles que le nom du chef de ménage et le nombre de personnes dans le ménage.

L'implémentation générale est de créer un rapport de la base de données et puis d'utiliser les événements mFormat qui sont exécutés lorsque chaque enregistrement du tableau au niveau du ménage est lu de façon à décider si un ménage est choisi ou pas. A la fin du processus, seulement les informations relatives aux ménages qui sont sélectionnés sont imprimées dans le rapport.

Une fois le rapport généré, il peut être imprimé, sauvegardé dans un tableau Access ou converti en feuille de calcul Excel.

Exemple d'implémentation détaillée

Les deux bases de données sont dénommées SDE Info (niveau grappe) et SDE Level (niveau ménage), dont voici les structures:

SDEInfo (Tableau niveau grappe)

Champ	Type	Description
ID	Autogénéré	Séquence numérique
Nom de la Localisation	Texte	Nom de la localisation
SDE (Key)	Nombre	Clé unique – Numéro de la grappe, lien au niveau SDE
Ménages	Nombre	Nombre de ménages
Echantillon	Nombre	Taille de l'échantillon

La base de données SDELevel des ménages contient

Champ	Type	Description
ID	Autogénéré	Séquence numérique
HHNO	Nombre	Numéro du ménage au sein de la grappe, doit être séquentiel de 1 aux Ménages SDEInfo.
SDE	Nombre	Numéro de la grappe, lien aux SDELevel
Personnes	Nombre	Nombre de personnes dans le ménage
Nom	Texte	Nom du chef de ménage

Le rapport dénommé CreateSample possède les propriétés importantes suivantes:

a) Il doit être un rapport relationnel qui contient les SDE, les Ménages et l'Echantillon issus de la base de données SDEInfo (les variables de localisation et autres variables d'information, s'il y a lieu, sont optionnelles)
   La base de données SDELevel doit inclure les variables HHNO, SDE et de façon optionnelle, des variables d'information telles que des noms de personnes.
b) Le SDELevel Detail band doit être trié selon la variable HHNO au moment de la constitution du rapport (il convient aussi de trier selon la SDE en général).
c) Dans le Detail band il y a une autre textbox dénommé SelSeq, utilisée pour emmagasiner la séquence numérique de sélection.
d) Il existe un gestionnaire d'événement onFormat associé au Group header band et le detail band et optionnellement au report header. Le code est le suivant:

```
Option Compare Database
Rem Ci-aprèes les variables utilisées pour la sélection
```

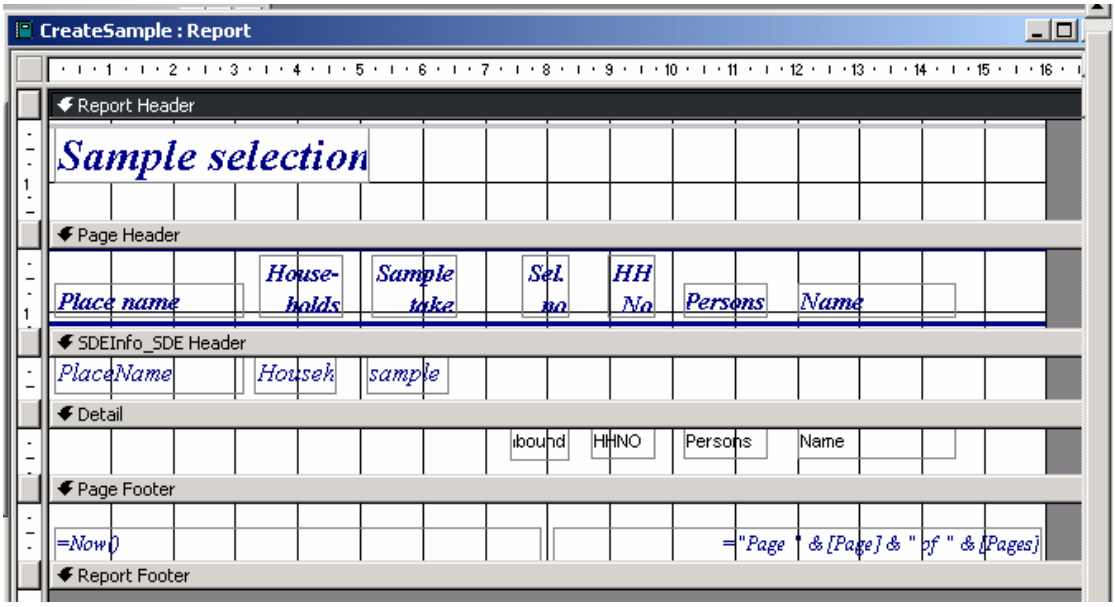
```
Rem Elles doivent être publique ((de telle sorte que detail Format puisse les utiliser
de manière répétée)
Public interval     ' selection interval
Public start        ' random start
Public NumSelected As Integer  ' Counter for selected

Private Sub Detail_Format(Cancel As Integer, FormatCount As Integer)
 Rem la selection a lieu ici
 Rem Logic: verifier si le ménage doit être sélectionné
 Rem        si oui, imprimer et augmenter le décompte des ménages sélectionnés
 Rem         si non, aller au prochain enregistrement sans imprimer
 seqnum = Report_CreateSample.HHNO
 cnum = start + interval * NumSelected
 If (seqnum - 1) < cnum And cnum <= seqnum Then
   NumSelected = NumSelected + 1
   Report_CreateSample.SelSeq = NumSelected
 Else
 With Report_CreateSample
 .MoveLayout = False      ‘ Does not advance one line on the page
 .NextRecord = True       ‘ moves to next record
 .PrintSection = False    ‘ Does not print current
 End With
End If
End Sub


Private Sub GroupHeader0_Format(Cancel As Integer, FormatCount As Integer)
    Rem Ce code est execute lorsqu’on initialize une nouvelle grappe
    Rem Les variables necessaries pour la selection initialisées ici
    Rem Initialisés ici
    interval = Report_CreateSample.Households / Report_CreateSample.sample
    start = Rnd * interval
    NumSelected = 0
End Sub

Private Sub ReportHeader_Format(Cancel As Integer, FormatCount As Integer)
    Rem Ce code n’est pas essential, il initialize le générateur de nombre aléatoire
    Rem seed generator
    Rem Eliminer le commentaire de l’énoncé pour fixer la sélection
    Rem (la valeur du nombre n’est pas importante)
    Rem Randomize (0.234)
End Sub
```

Screen dump of report design



Le tableau est le rapport produit par Access :

Sélection de l'échantillon

Nom du lieu	Ménages	Echantillon	No. Sel	No Ménage	Personnes	Noms
Pétion Ville	81	19				

The following is the report produced by Access:

Sample selection

Place name	House-holds	Sample take	Sel. no	HH No	Persons	Name
Petion Ville	81	19				
			1	5	6	Moulel
			2	9	4	Roland
			3	13	8	Borkman
			4	17	12	Gregory
			5	22	7	Laurie
			6	26	2	Edwin
			7	30	6	Maria
			8	34	4	Dieter
			9	39	9	Maldon
			10	43	2	Emma
			11	47	6	Raymond
			12	51	10	Theresa
			13	56	15	Wang
			14	60	5	Gurl
			15	64	9	Darwin
			16	68	6	Johannes
			17	73	11	Nina
			18	77	2	Tommy
			19	81	6	Lambert

Sample selection

Place name	*House-holds*	*Sample take*	*Sel. no*	*HH No*	*Persons*	*Name*
Petion Ville	*81*	*19*				
			1	5	6	Moulel
			2	9	4	Roland
			3	13	8	Borkman
			4	17	12	Gregory
			5	22	7	Laurie
			6	26	2	Edwin
			7	30	6	Maria
			8	34	4	Dieter
			9	39	9	Maldon
			10	43	2	Emma
			11	47	6	Raymond
			12	51	10	Theresa
			13	56	15	Wang
			14	60	5	Gurl
			15	64	9	Darwin
			16	68	6	Johannes
			17	73	11	Nina
			18	77	2	Tommy
			19	81	6	Lambert