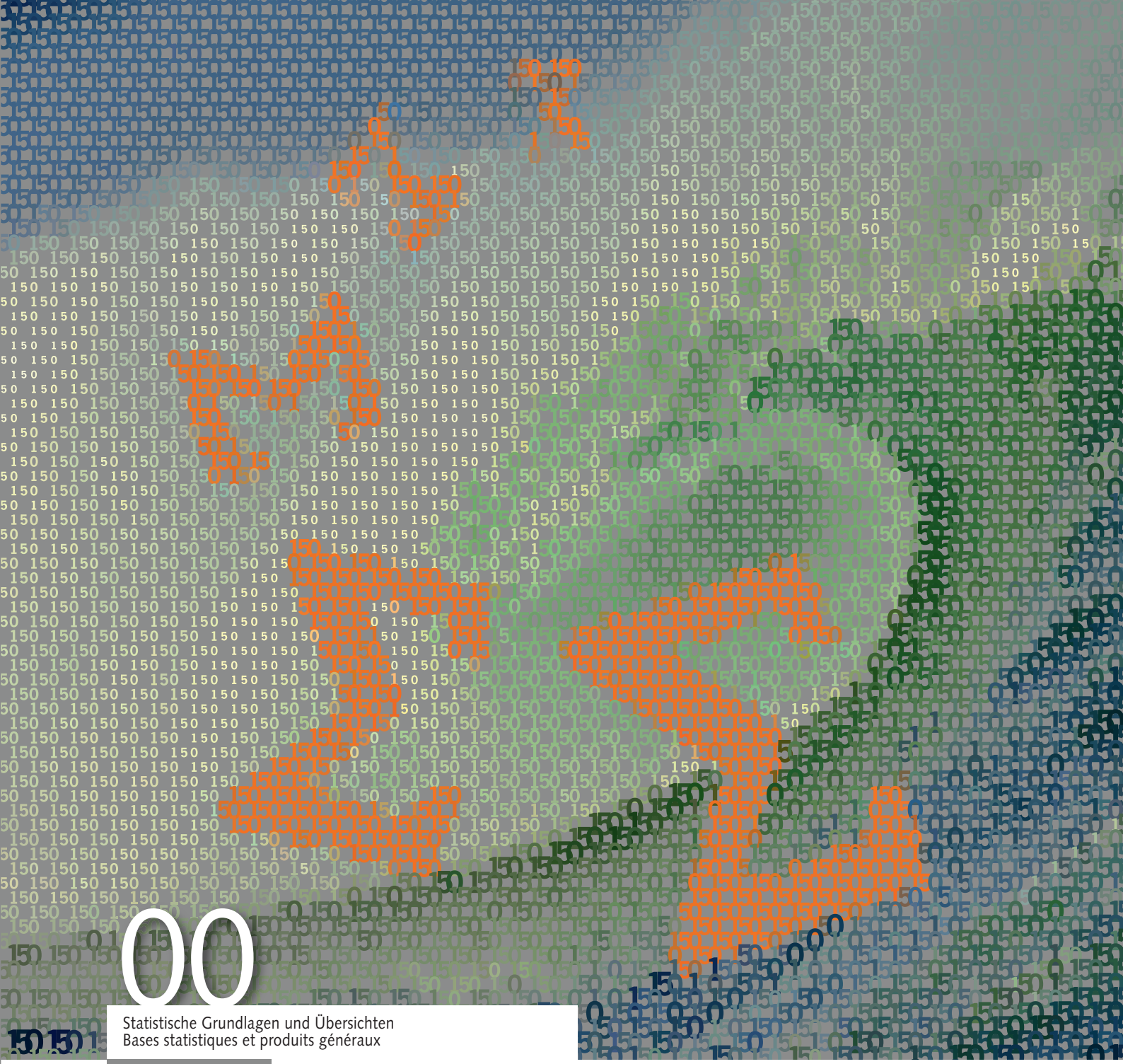


00

Statistische Grundlagen und Übersichten
Bases statistiques et produits généraux

338-0061

Rapport de méthodes

Enquête suisse sur la santé 2007

Plan d'échantillonnage, pondérations et analyses pondérées des données

150 jahre BFS
ans OFS
anni UST
years FSO

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Bundesamt für Statistik BFS
Office fédéral de la statistique OFS

Neuchâtel, 2010

Statistik der Schweiz

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

0	Statistische Grundlagen und Übersichten
1	Bevölkerung
2	Raum und Umwelt
3	Arbeit und Erwerb
4	Volkswirtschaft
5	Preise
6	Industrie und Dienstleistungen
7	Land- und Forstwirtschaft
8	Energie
9	Bau- und Wohnungswesen
10	Tourismus
11	Mobilität und Verkehr
12	Geld, Banken, Versicherungen
13	Soziale Sicherheit
14	Gesundheit
15	Bildung und Wissenschaft
16	Kultur, Medien, Informationsgesellschaft, Sport
17	Politik
18	Öffentliche Verwaltung und Finanzen
19	Kriminalität und Strafrecht
20	Wirtschaftliche und soziale Situation der Bevölkerung
21	Nachhaltige Entwicklung und Disparitäten auf regionaler und internationaler Ebene

Statistique de la Suisse

La série «Statistique de la Suisse» publiée par l'Office fédéral de la statistique (OFS) couvre les domaines suivants:

0	Bases statistiques et produits généraux
1	Population
2	Espace et environnement
3	Vie active et rémunération du travail
4	Economie nationale
5	Prix
6	Industrie et services
7	Agriculture et sylviculture
8	Energie
9	Construction et logement
10	Tourisme
11	Mobilité et transports
12	Monnaie, banques, assurances
13	Protection sociale
14	Santé
15	Education et science
16	Culture, Médias, société de l'information, sport
17	Politique
18	Administration et finances publiques
19	Criminalité et droit pénal
20	Situation économique et sociale de la population
21	Développement durable et disparités régionales et internationales

Statistik der Schweiz

Methodenbericht

Statistique de la Suisse

Rapport de méthodes

Enquête suisse sur la santé 2007

Plan d'échantillonnage, pondérations et analyses pondérées des données

Auteur **Eric Graf**
Office fédéral de la Statistique

Editeur **Office fédéral de la statistique**

Neuchâtel, 2010

Préambule

L'enquête suisse sur la santé 2007 (ESS07) est une enquête quinquennale réalisée pour la quatrième fois consécutive par l'Office fédéral de la statistique (OFS) sur mandat du conseil fédéral. Le service des méthodes statistiques (METH) de l'OFS a été mandaté pour réaliser les travaux méthodologiques, notamment le tirage et le suivi des échantillons durant l'année d'enquête, puis le développement de modèles de pondération et le calcul de la précision, une fois les données plausibilisées disponibles. Ces travaux ont été menés, pour la première partie, par M. Christian Panchard, et par M. Eric Graf pour ce qui est des travaux relatifs à la pondération et l'extrapolation.

Que soient ici remerciées les personnes ayant pris une part active dans les discussions et la collaboration avec METH: M. Renaud Lieberherr, responsable des analyses de l'ESS au sein de la section GESB, ainsi que M. Marco Storni, chef du projet ESS, et Mme Marilina Galati.

Toute notre gratitude va également à Lionel Qualité et Jean-Pierre Renfer de METH pour leurs remarques avisées après une lecture approfondie de ce rapport.

Résumé

L'ESS07 a pour objectif principal de fournir, à partir d'un échantillon de près de 19'000 personnes résidant en Suisse et âgées de 15 ans et plus, des résultats sur l'état de santé de la population, les comportements en matière de santé, l'assurance-maladie et le recours aux prestations du système de santé.

Ce rapport décrit les méthodes utilisées pour établir le plan d'échantillonnage en abordant les questions relatives à la délimitation de la population, au cadre de sondage et aux unités d'échantillonnage. Le tirage, la gestion et l'érosion des échantillons sont également abordés dans la première partie.

L'ESS se compose en fait de deux enquêtes: la première est une enquête téléphonique alors que la seconde, sous forme écrite, ne s'adresse qu'aux personnes ayant répondu à la première. La pondération relative à chacune de ces enquêtes y est décrite dans la deuxième partie.

Enfin, la troisième partie tente d'apporter des réponses aux questions les plus courantes reçues par le service des méthodes statistiques à propos de l'utilisation et de l'exploitation des données de l'ESS à l'aide des pondérations.

Mots-clé

Rapport de méthodes; cadre de sondage; non-réponse; plan d'échantillonnage; pondération; stratification; calage; variance; segmentation.

Complément d'information: Eric Graf, tél. 032 713 64 01
Eric.Graf@bfs.admin.ch

Réalisation: Service de méthodes statistiques, OFS

Diffusion: Office fédéral de la statistique
CH-2010 Neuchâtel
Tél. 032 713 60 60 / Fax 032 713 60 61
Order@bfs.admin.ch

Internet: http://www.statistik.admin.ch

Numéro de commande: 338-0061

Prix: gratuit

Série: Statistique de la Suisse

Domaine: 0 Bases statistiques et produits généraux

Langue du texte original: Français

Page de couverture: OFS; concept: Netthoevel & Gaberthüel, Bienne;
photo: © NorthShoreSurfPhotos – Fotolia.com

Graphisme/Layout: OFS

Copyright: OFS, Neuchâtel 2010
La reproduction est autorisée, sauf à des fins commerciales,
si la source est mentionnée.

ISBN: 978-3-303-00436-4

Table des matières

I Introduction à l'enquête suisse sur la santé **5**

1 Généralités **5**
1.1 Déroulement de l'enquête suisse sur la santé 2007 dans les ménages privés 5
1.2 Délimitation de la population 5

2 Échantillonnage **6**
2.1 Cadre de sondage et unités d'échantillonnage 6
2.2 Plan d'échantillonnage 6
2.3 Érosion des échantillons et taux de réponse 7

II Pondérations **11**

3 Introduction à la méthodologie utilisée **11**
3.1 Étapes de la construction des poids 11

4 Pondération initiale **13**
4.1 Poids d'échantillonnage ménage : étape 1 13
4.2 Poids initial ménage : étapes 2 et 3 14
4.3 Poids initial individu : étapes 4 et 5 14

5 Modélisation de la non-réponse par segmentation **16**
5.1 La méthode de segmentation 16
5.2 Modélisation de la non-réponse au questionnaire téléphonique : étapes 6 et 6b 17
5.3 Modélisation de la non-réponse au questionnaire écrit : étapes 7 et 7b 19

6 Introduction théorique sur le calage utilisé **21**
6.1 Notations 21
6.2 Approche de la pondération en deux phases pour corriger la non-réponse 21
6.3 Vecteurs et information auxiliaire 22
6.4 Le calage 23
6.4.1 La méthode "classique" 23
6.4.2 La méthode "lune et étoile" 24

7 Calcul des poids finaux **25**
7.1 Pondération de l'enquête téléphonique 25
7.2 Pondération de l'enquête écrite 31
7.3 Commentaires sur les poids finaux obtenus 33

III Analyses des données avec pondération **38**

8 Documentation existante sur les échantillons de 1992 à 2007 **38**
8.1 Définition des strates d'échantillonnage 38
8.2 Tailles des cadres de sondages 38

9 Comparaisons géographiques et temporelles **42**
9.1 Calculs de précisions . . . 42
9.2 Comparaisons régionales . . . 42
9.3 Comparaisons temporelles . . . 43

10 Questions fréquentes, conseils pour les utilisateurs **44**
10.1 Faut-il utiliser les poids ? . . . 44
10.2 Quelle pondération utiliser : téléphonique/écrite ? . . . 44
10.3 Quelle est la différence entre la strate et le canton ? . . . 44
10.4 Faut-il tenir compte de la poststratification dans l'analyse ? . . . 44
10.5 Les poids peuvent-ils être utilisés tels quels sur un sous-groupe de l'échantillon ? . . . 44
10.6 Dans quelles conditions doit-on standardiser les poids ? . . . 45
10.6.1 Régression logistique . . . 45
10.6.2 Test binomial . . . 48
10.6.3 Test du Chi-carré . . . 48
10.6.4 Conclusions . . . 49
10.7 La somme des poids ne correspond pas aux nombres publiés dans l'ESPOP . . . 50

Conclusions **51**

Références **52**

Annexes **53**

A Modélisation de la non-réponse à l'enquête CATI/CAPI **53**

B Modélisation de la non-réponse à l'enquête écrite **56**

Première partie

Introduction à l'enquête suisse sur la santé

1 Généralités

1.1 Déroulement de l'enquête suisse sur la santé 2007 dans les ménages privés

La santé de la population suisse a été étudiée pour la quatrième fois en 2007 dans le cadre de l'Enquête Suisse sur la Santé (ESS07). Cette étude a lieu tous les cinq ans (les trois premières enquêtes ont été réalisées en 1992/93, 1997 et 2002). L'ordonnance concernant l'exécution des relevés statistiques fédéraux du 30 juin 1993 (RS 431.012.1) en est la base légale. Il s'agit d'une vaste enquête basée sur un échantillon qui permet d'extrapoler à la population résidante âgée d'au moins 15 ans (voir paragraphe 2.1). La section "Santé de la Population" de l'Office Fédéral de la Statistique est responsable de la réalisation et de la coordination du projet.
La réalisation sur le terrain a été confiée, en 2007, à l'institut M.I.S Trend (Lausanne et Gümligen), en 2002, à l'Institut IHA-GfK (Lucerne et Lausanne), en 1997, à l'Institut LINK (Lucerne et Lausanne) et, en 1992, à l'Institut IPSO (Zürich et Genève).

L'objectif principal de l'ESS est de recueillir des données représentatives sur l'état de santé, les attitudes face à la santé, les comportements, les conditions de vie et la prévention, et les recours aux prestations de système de santé.

Les instruments de mesure de l'ESS07 sont comparables à ceux des trois enquêtes précédentes : un questionnaire téléphonique suivi d'un questionnaire écrit. L'enquête a été menée en deux étapes. Dans une première phase, un échantillon de personnes disposant d'un raccordement téléphonique et résidant en Suisse a été contacté pour une interview (en allemand, français ou italien) soit par téléphone (technique CATI : "*Computer Assisted Telephone Interview*"), soit en face-à-face (CAPI "*Computer Assisted Personal Interview*") pour les personnes de 75 ans et plus qui le préféraient ou pour qui ce mode d'interrogation était plus adapté. Dans cette première phase, si la personne cible n'était pas en mesure de répondre aux questions par téléphone ou en face-à-face pour cause de maladie, acccident, de handicap de naissance, de vieillesse, de problème linguistiques ou d'absence pendant plus de 4 semaines, une interview téléphonique avec une personne de substitution proche de la personne cible (interview PROXY) a été réalisée. La deuxième phase consiste en un questionnaire écrit transmis par la poste ou via Internet (eSurvey) aux personnes ayant pris part aux interviews téléphoniques ou face-à-face, proxys mis à part[1]. Le questionnaire écrit était accompagné d'un cadeau de remerciement pour les renseignements fournis.

1.2 Délimitation de la population

Comme pour les enquêtes précédentes, la population cible que l'on désire étudier est constituée de l'ensemble des personnes âgées de 15 ans et plus résidant de manière permanente en Suisse. Les personnes vivant dans les homes pour personnes âgées, dans les prisons, ou dans d'autres institutions et ménages collectifs ne font pas partie de la population étudiée.

[1] Le contenu du questionnaire écrit est, pour certaines questions, trop détaillé pour qu'une autre personne que la personne cible le remplisse à sa place. On devrait développer un questionnaire écrit PROXY plus court pour que cela puisse se faire, ce qui n'a pas été fait.
En 1992 le questionnaire écrit a été envoyé uniquement aux personnes âgées de 15 à 74 ans ; pour 1997, 2002 et 2007 il n'a pas été fixé de limite supérieure d'âge.

2 Échantillonnage

2.1 Cadre de sondage et unités d'échantillonnage

L'échantillon de l'enquête suisse sur la santé 2007 est issu du SRH, le registre pour les enquêtes auprès des ménages de l'OFS. Ce dernier est actualisé trimestriellement. Il contient les numéros de téléphone des ménages qui y sont répertoriés et fournit ainsi le cadre de sondage pour le tirage d'échantillons de ménages privés. Une fois le ménage contacté, ses caractéristiques peuvent être déterminées (comme la taille du ménage, l'âge, le sexe ou la nationalité de ses membres, etc.) ; lorsque ces informations ont été collectées, on dit que le ménage a été documenté. On est alors en mesure de former un échantillon aléatoire de personnes de la population cible (âgée de 15 ans ou plus) et de l'interroger. La plupart des personnes de 15 ans ou plus résidant de manière permanente en Suisse vivent dans des ménages privés et disposent d'un raccordement téléphonique enregistré dans le registre. Cependant, le recouvrement entre la population et le registre n'est pas parfait. Il convient de traiter les effets de ce défaut de couverture (tout comme ceux liés à la non-réponse aux questionnaires) par des méthodes appropriées permettant de réduire les risques de biais. Dans cette enquête, le recours à des calages sur marges basés sur des modèles de non-réponse et sur la statistiques annuelle de la population ESPOP se sont avérés judicieux.
Le fait de disposer d'un cadre de sondage des ménages privés ne permet pas de procéder directement à un échantillonnage de personnes. Il faut donc considérer un tirage aléatoire à deux niveaux. Au premier niveau, l'unité d'échantillonnage est le ménage privé (l'unité de tirage correspond au raccordement téléphonique du ménage). On parle dans ce cas d'**unités primaires** (ménages). Pour chaque ménage de l'échantillon tiré au premier niveau qu'on a pu documenté, on procède à un tirage aléatoire d'une seule personne âgée de 15 ans ou plus. Il s'agit du deuxième niveau dont les éléments sont appelés **unités secondaires** (personnes).

2.2 Plan d'échantillonnage

La récolte de l'information de l'ESS dure toute une année : l'échantillon est construit en 4 vagues. A chaque saison, l'une des 4 vagues est enquêtée.
La taille de chacune des 4 vagues est approximativement identique. Chaque sous échantillon provient d'un sondage aléatoire stratifié. Les strates sont définies par les cantons, excepté les cantons de Zurich et Berne qui ont été partitionnés en respectivement deux et trois strates. On a donc 29[2] strates au total (voir tableau 15). Cette stratification géographique est justifiée par le besoin de traiter spécialement les cantons (ou régions) ayant demandé d'augmenter la taille de leur échantillon à des fins d'analyses à leurs niveaux respectifs. A l'intérieur de chacune des 29 strates, un tirage aléatoire à deux niveaux est effectué.
Selon l'ordonnance concernant l'exécution des relevés statistiques fédéraux du 30 juin 1993 (RS 431.012.1), l'enquête suisse sur la santé doit comporter environ 10'000 interviews réalisées. Ce nombre minimum d'interviews est réparti par grande région proportionnellement à sa taille en terme de population. Et à l'intérieur de chaque grande région dans les cantons, aussi proportionnellement à leur taille. En accord avec la conception détaillée du projet, l'OFS a finalement financé une base d'au moins 11'700 interviews. Comme c'était le cas pour les enquêtes passées, les cantons avaient la possibilité d'agrandir l'échantillon qui leur est attribué par défaut en finançant des interviews supplémentaires. Ainsi, sur les 26 cantons, 13 cantons ainsi que la ville de Zurich ont sollicité une augmentation de leur échantillon pour pouvoir réaliser des exploitations plus précises (au niveau du canton et de Zurich). Il s'agit des cantons suivants : Argovie, Appenzell Rhodes-Extérieures, Berne, Fribourg, Genève, Jura,

[2] Notons qu'en 2007 les cantons de AI et AR ont été chacun considérés comme une strate contrairement à l'ESS02 où ils avaient été réunis pour former une seule strate.

Lucerne, Neuchâtel, Schwyz, Tessin, Uri, Vaud et Valais. Le canton de Berne a en outre élargi séparément des échantillons régionaux. Il a ainsi été partitionné en trois strates : le Jura bernois, l'Oberland bernois et les autres parties du canton. Au total les cantons ont financés 6'637 interviews supplémentaires, l'échantillon net visé était donc de 18'337 interviews (voir tableau 1).
La gestion des échantillons n'est pas triviale et doit être conduite avec le plus grand soin de manière à satisfaire au mieux le but recherché. Dans notre cas, il s'agit de garantir un nombre minimum d'interviews réalisées réparties aussi uniformément que possible sur les quatre vagues de l'enquête, sans pour autant sacrifier les taux de réponse en activant un nombre trop élevé d'adresses pour cela. En effet, si l'on active trop d'adresses, les ménages ou personnes faciles à atteindre et coopératifs sont, par la force de choses, interrogés d'abord. Si le nombre d'interviews minimal à effectuer est trop vite atteint, on risque de s'en contenter au lieu de faire tous les efforts d'usage pour épuiser l'échantillon (conversions de refus, rappels à d'autres jours et heures, etc.). Le résultat est un moins bon taux de réponse et un risque de biais plus important. Un avantage indéniable de procéder en quatre vagues en tirant un nouvel échantillon chaque trimestre durant une année est de pouvoir gérer les tailles brutes de ces échantillons en les adaptant en fonction de l'information obtenue sur les vagues précédentes. La gestion des échantillons par strates pour les 4 vagues a été menée de manière tout à fait similaire à ce qui a été fait pour l'ESS02, et nous renvoyons le lecteur au rapport de méthodes correspondant [2] pour plus d'informations à ce sujet.

2.3 Érosion des échantillons et taux de réponse

Pour les quatre vagues de l'enquête, l'OFS a livré un échantillon brut de 30'179 adresses (numéros de téléphone) à l'institut de sondage. La répartition de ces adresses par strate géographique est donnée dans le tableau 3. La description de l'érosion est particulièrement importante. Elle se situe à différents niveaux dans l'enquête :

1. au niveau des adresses elles-mêmes qui ne sont pas parfaites, 1'847 adresses ou numéros se sont révélés être invalides ou ne pas correspondre à un ménage privé (maisons de vacances, ménages collectifs, entreprises, etc.).
2. Puis au niveau des ménages privés qui ne peuvent ou ne veulent pas répondre (6'185 pertes supplémentaires),
3. des personnes cible de l'enquête téléphoniques (3'414 n'ont pas participé alors que le ménage avait été documenté)
4. et enfin de l'enquête écrite (4'367 questionnaires en moins). Cette érosion de l'échantillon est résumée dans le tableau 2 avec les pourcentages correspondants à la base y relative.

A partir des informations contenues dans le tableau 2, il est possible de définir différents taux dont l'interprétation permet de quantifier la qualité des résultats obtenus. On peut ainsi définir les *taux de validité*, *taux d'éligibilité*, *taux de participation* des ménages, taux de participation des personnes (tél.) et taux de participation des personnes (écrit) suivants :

t_{ve} = taux de validité/éligibilité[3] = $\frac{\textit{adresses valides et éligibles}}{\textit{total des adresses}} = \frac{28332}{30179} = 94\%$;

t_p^M = taux de participation (ménages) = $\frac{\textit{ménages participants}}{\textit{adresses valides et éligibles}} = \frac{22174}{28332} = 78\%$;

$t_{p,tel}^P$ = taux de participation (tél.) = $\frac{\textit{personnes cible participantes CATI, CAPI, PROXY}}{\textit{ménages participants}} = \frac{18760}{22174} = 85\%$;

$t_{e,ect}^P$ = taux d'éligibilité (écrit) = $\frac{\textit{personnes participantes sans CATI, CAPI}}{\textit{personnes participantes CATI, CAPI, PROXY}} = \frac{17931}{18760} = 96\%$;

$t_{p,ect}^P$ = taux de participation (écrit) = $\frac{\textit{personnes cible participantes}}{\textit{personnes participantes CATI, CAPI}} = \frac{14393}{17931} = 80\%$.

[3] Les statistiques de suivi de l'ESS07 ne permettent malheureusement pas de distinguer les adresses valides des adresses éligibles comme il se devrait. Ces taux ne sont donc pas à comparer directement avec ceux obtenu en 2002.

Strate géographique et grandes régions	N° des cantons et grandes régions	Tailles nettes ESS07 prévues par l'OFS	Augmentations régionales tailles nettes prévues	Tailles nettes visées (prévues)	N° de la strate (Stratnum)
VD	22	955	400	1355	220
VS	23	425	575	1000	230
GE	25	640	400	1040	250
Rég. lémanique	**1**	**2020**	**1375**	**3395**	
BE-Oberland		293	307	600	21
BE-jura	2	78	522	600	22
BE-reste		1105		1105	23
FR	10	357	643	1000	100
SO	11	379		379	110
NE	24	256	400	656	240
JU	26	104	400	504	260
Espace Mittelland	**2**	**2571**	**2272**	**4843**	
BS	12	303		303	120
BL	13	409		409	130
AG	19	843	400	1243	190
Suisse du N-O	**3**	**1555**	**400**	**1955**	
ZH-Ville	1	549	400	949	11
ZH-reste		1379		1379	12
Zurich	**4**	**1928**	**400**	**2328**	
GL	8	58		58	80
SH	14	114		114	140
AR	15	80	400	480	150
AI	16	22		22	160
SG	17	682		682	170
GR	18	285		285	180
TG	20	342		342	200
Suisse orientale	**5**	**1584**	**400**	**1984**	
LU	3	531	490	1021	30
UR	4	53	400	453	40
SZ	5	198	400	598	50
OW	6	49		49	60
NW	7	58		58	70
ZG	9	153		153	90
Suisse centrale	**6**	**1042**	**1290**	**2332**	
TI	21	1000	500	1500	210
Tessin	**7**	**1000**	**500**	**1500**	
SUISSE		**11700**	**6637**	**18337**	

T 1 Plan d'échantillonnage de l'ESS07, nombres d'interviews téléphonique nettes visées avant l'enquête. La colonne "Taille nette ESS07" indique le nombre d'interviews prévues par l'OFS et la colonne "Augmentations régionales" détaille les densifications par canton ou partie de canton financées par ces derniers. "Stratnum" définit les strates d'échantillonnage.

Échantillon brut (lettres envoyées)	30179	100%				
Adresses valides et éligibles	28332	94%	100%			
Ménages participants	22174	73%	78%	100%		
Personnes participantes (CATI+CAPI)	18760	62%	66%	85%	100%	
Personnes participantes CATI	17931	59%	63%	81%	96%	100%
Personnes cible participantes (écrit)	14393	48%	51%	65%	77%	80%

T 2 Érosion de l'échantillon au niveau des ménages et des personnes. Une seule personne étant sélectionnée et interviewée par ménage, on peut se permettre de confondre les deux niveaux de collecte de l'information dans un même tableau comme cela est fait ici.

Strate géographique (Stratnum)	Echantillon brut, adresses activées	Adresses valides et éligibles	Ménages documentés	Individus interviewés	Proxys	Quest. écrits valides
VD 220	2190	2059	1616	1365	57	1066
VS 230	1953	1776	1319	1079	51	825
GE 250	1851	1705	1252	1036	68	730
Rég. Lémanique	**5994**	**5540**	**4187**	**3480**	**176**	**2621**
BE-Oberland 21	991	916	743	641	27	493
BE-jura 22	1046	982	764	632	25	496
BE-reste 23	1601	1527	1264	1105	44	843
FR 100	1612	1527	1210	1009	57	777
SO 110	562	530	434	377	10	299
NE 240	1018	974	778	666	38	525
JU 260	796	754	595	514	25	399
Espace Mittelland	**7626**	**7210**	**5788**	**4944**	**226**	**3832**
BS 120	473	446	343	300	6	237
BL 130	642	617	485	408	12	329
AG 190	1964	1883	1523	1267	59	968
Suisse du N.-O.	**3079**	**2946**	**2351**	**1975**	**77**	**1534**
ZH-Ville 11	1606	1482	1128	969	38	718
ZH-reste 12	2133	2037	1647	1401	45	1102
Zurich	**3739**	**3519**	**2775**	**2370**	**83**	**1820**
GL 80	120	111	82	69	1	55
SH 140	158	154	120	108	4	89
AR 150	751	714	579	497	14	389
AI 160	37	37	28	22	0	18
SG 170	1102	1049	823	689	34	522
GR 180	580	501	345	286	14	227
TG 200	527	506	407	339	16	263
Suisse orientale	**3275**	**3072**	**2384**	**2010**	**83**	**1563**
LU 30	1653	1590	1248	1074	35	833
UR 40	736	703	548	468	18	360
SZ 50	987	940	735	615	34	455
OW 60	88	78	59	49	1	43
NW 70	97	95	74	63	1	55
ZG 90	246	235	186	154	9	117
Suisse centrale	**3807**	**3641**	**2850**	**2423**	**98**	**1863**
TI 210	2659	2404	1839	1558	86	1160
Tessin	**2659**	**2404**	**1839**	**1558**	**86**	**1160**
SUISSE	**30179**	**28332**	**22174**	**18760**	**829**	**14393**

T 3 Résultats du terrain de l'ESS07. Le nombre d'adresses valides et éligibles (28332) diffère légèrement de celui du rapport final de M.I.S-Trend (28319) en raison de certaines adaptations effectuées ultérieurement.

Deuxième partie

Pondérations

3 Introduction à la méthodologie utilisée

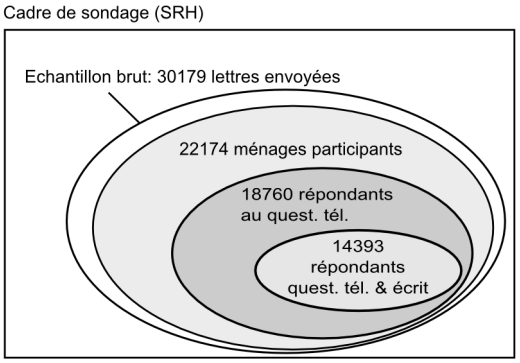
Les données des fichiers livrés de l'OFS aux utilisateurs ne sont pas pondérées. Il est du ressort des utilisateurs des données d'appliquer les poids appropriés en fonction des analyses qu'ils veulent produire. On risque de forts biais et une sous-estimation de la variance si l'on n'applique pas les poids. Pour des recommandations et des exemples sur comment et quels poids utiliser, nous renvoyons le lecteur à la troisième partie de ce rapport qui est consacrée à ces questions.
Les pondérations développées ici suivent la logique du processus de collecte de l'information dans l'enquête. Les grandes étapes de la production des poids sont les suivantes :

1. Une *pondération initiale* est effectuée de manière à tenir compte à la fois du plan d'échantillonnage et de la non-réponse. Ces poids initiaux découlent des probabilités d'inclusion et d'une première correction pour la non-réponse.
2. Des ajustements de *poids pour la non-réponse* à différents niveaux permettent de réduire les risques de biais en modélisant finement le mécanisme de réponse des individus et des ménages.
3. Le recours à des données auxiliaires de référence permet de procéder à *un calage* fournissant les pondérations finales à utiliser dans le calcul des différents estimateurs.

Les deux dernières étapes de la production des poids listées ci-dessus ont été, grâce à la méthodologie développée récemment par C.-E. Särndal et S. Lundström (voir [10]), synthétisées en un seul calage dit "lune et étoile" qui permet à la fois de corriger pour la non-réponse au niveau de l'échantillon et de respecter les bons totaux au niveau de la population pour certaines variables-clé.
Cette partie du rapport est consacrée au développement des modèles de pondération de l'ESS07 de la pondération initiale aux pondérations finales, en passant par les calages sur marge et le choix des variables de calage.

3.1 Étapes de la construction des poids

Les étapes principales de la construction des poids suivent le processus de collecte des données de l'enquête sur le terrain. On se calque également sur le déroulement de la prise d'information lorsqu'on désire calculer les taux de réponse atteints (voir section 2.3).

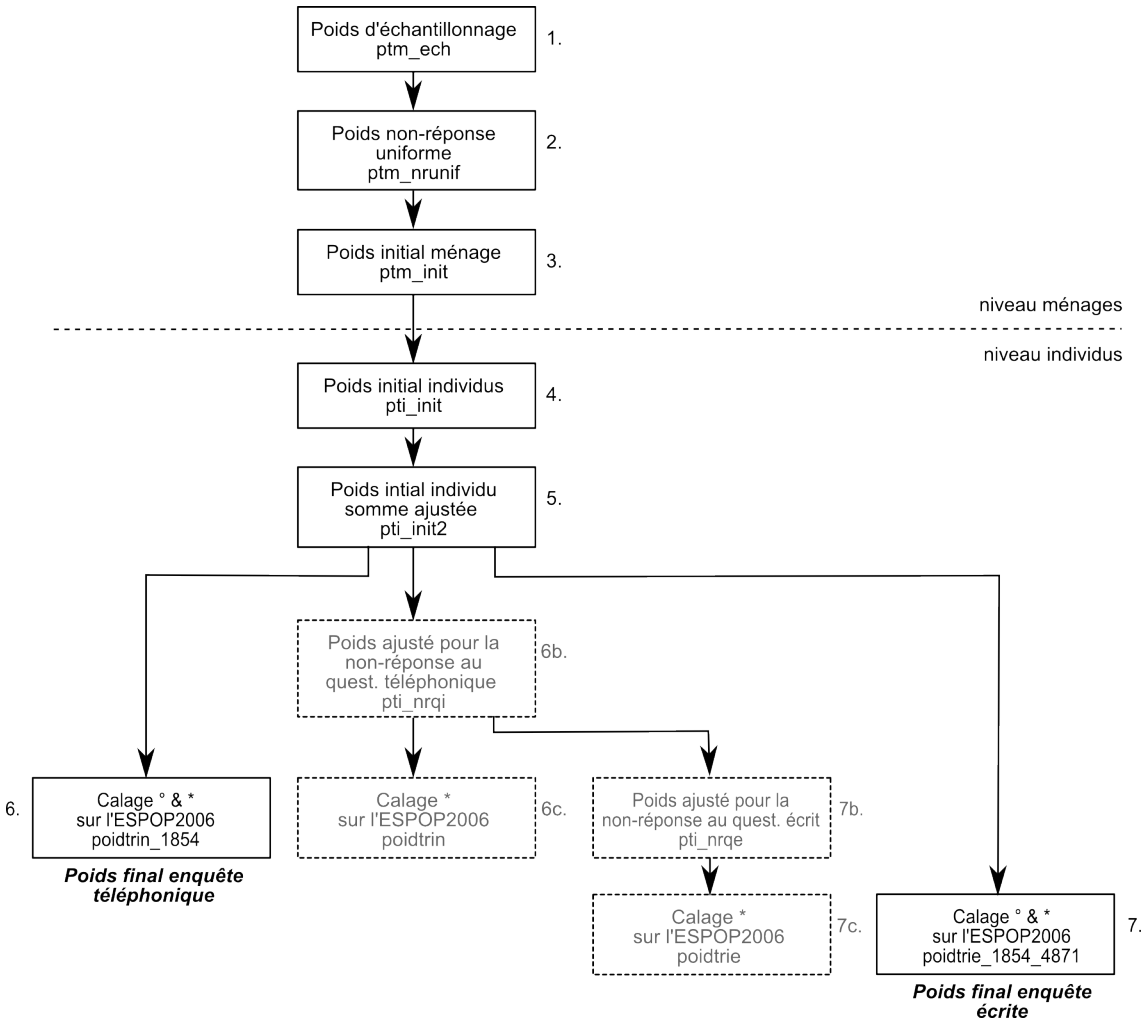


G 1 Schéma de la non-réponse dans l'échantillon.

La figure 1 schématise encore une fois le processus dont les étapes-clé sont les suivantes :

– l'échantillon brut : défini par les adresses et numéros de téléphone sélectionnés dans le cadre de sondage pour participer à l'enquête,
– les ménages participants : ceux qui ont pu être documentés,
– les personnes cibles ayant participé à l'enquête téléphonique,
– les personnes cibles ayant participé à l'enquête écrite.

La figure 2 résume schématiquement les différents processus de construction des poids envisagés et détaillés dans les paragraphes qui suivent.



G 2 Schéma du processus de construction des pondérations des enquêtes téléphonique et écrite. Les case en trait-tillé sont des variantes qui ont aussi été calculées mais non retenues dans le choix final du modèle de pondérations. Dans le texte de ce rapport on se réfère à ces variantes comme étant la manière conventionnelle de calculer les poids par contraste avec la "nouvelle" méthodologie (calage en une seule étape "lune et étoile") qui a finalement été appliquée et retenue ici. La numérotation des cases correspond aux étapes de la pondération détaillées dans les paragraphes du texte qui s'y réfèrent.

4 Pondération initiale

Chaque individu ayant participé à l'enquête se voit attribuer un poids initial qui n'est rien d'autre que le produit de l'inverse de sa probabilité d'inclusion et du taux de réponse par strate. Nous distinguerons ainsi les poids d'échantillonnage des facteurs correctifs de poids dus à la non-réponse dont le produit fournit les poids initiaux. Pour cela, fixons les notations suivantes. Soit $U = 1, ..., k, ..., N$ la population des personnes de 15 ans et plus résidant dans des ménages privés en Suisse. Cette population est partitionnée en M unités primaires (ménages) notées $U_1, ..., U_i, ..., U_M$. L'ensemble des unités primaires (la population des ménages privés) est notée par $U^M = 1, ..., i, ..., M$. La taille d'un ménage U_i (le nombre de personnes de 15 ans ou plus qu'il contient) est désignée par N_i. Le tableau 7 résume ces notations et en introduit des supplémentaires.

k	indice utilisé pour désigner des personnes
i	indice utilisé pour désigner des ménages
U_i	ensemble des personnes d'un même ménage
N_i	taille du ménage U_i
h	strate du ménage U_i ou de l'individu k
$s_h^{b,M}$	échantillon brut de ménages dans la strate h, de taille m_h^b
$s_h^{n,M}$	échantillon net de ménages dans la strate h, de taille m_h^n (interviews réalisées)
$s_h^{b,P}$	échantillon brut de personnes dans la strate h
M_h	nombre de ménages dans la population de la strate h.

T 4 Notations.

L'enquête ESS se base sur quatre échantillons (quatre vagues) provenant chacun d'un tirage aléatoire stratifié à deux niveaux. Les échantillons sont tirés de manière disjointe. Cependant, pour simplifier et parce que certaines strates ont un échantillon par vague de très petite taille qui mène à des volatilités trop grandes et indésirables des poids d'échantillonnage, on a décidé de traiter les échantillons par strate des 4 vagues comme un seul. L'effet sur les poids finaux entre cette solution et celle de considérer séparément les échantillons des-dites 4 vagues a été testé et la première solution évoquée a été jugée préférable. On est donc dans le cas d'un tirage aléatoire simple aux deux niveaux suivants :

– *Premier niveau :* un échantillon s^M de ménages est tiré dans $U^M (s^M \subset U^M)$ selon le plan $p_M(\cdot)$.
– *Deuxième niveau :* dans chaque ménage $i \in s^M$, on tire un échantillon s_i^P de taille 1 (une personne), $s_i^P \subset U_i$, selon le plan $p_i(\cdot \mid s^M)$.

4.1 Poids d'échantillonnage ménage : étape 1

Notons par $\pi_i^M = Pr(i \in s^M)$ la probabilité d'inclusion d'un ménage i dans l'échantillon. Cette probabilité d'inclusion est la même pour tous les ménages au sein d'une même strate (canton ou région) :

$$\pi_i^M = Pr(i \in s_h^{b,M}) \qquad \text{si} \quad i \in h.$$

Ladite probabilité d'inclusion π_i^M des ménages de la strate h est donnée par :

$$\pi_{h,i}^M = \frac{m_h^b}{M_h} \qquad \text{avec} \quad i \in \text{strate} h.$$

On peut ainsi en déduire le poids (transversal) d'échantillonnage du ménage i dans la strate h (les M_h sont donnés à la fin du document dans la dernière colonne du tableau 23) :

$$ptm_ech_{i,h} = \frac{M_h}{m_h^b} = \frac{1}{\pi_{h,i}^M} \qquad \text{avec} \quad \in U^M \quad \text{et} \quad i \in \text{strate}h \tag{1}$$

4.2 Poids initial ménage : étapes 2 et 3

Pas tous les ménages (ou toutes les personnes dans les ménages) de l'échantillon participent à l'enquête, et ce pour différentes raisons. De ce fait, certains ménages n'ont pas pu être documentés.
Si τ_i est le taux de réponse du ménage ou plus exactement la probabilité de participer à l'enquête (conditionnellement au fait que le ménage i a été tiré dans l'échantillon), $\tau_i = Pr(i \text{ participe } | i \in s^M)$, alors la correction du poids due à la non-réponse, w_i^{NR}, est donnée par

$$w_i^{NR} = \frac{1}{\tau_i}.$$

En faisant une hypothèse d'homogénéité de la non-réponse par strate au **niveau des ménages**, on a :

$$\tau_i = \tau_h, \forall i \in s_h^{b,M}.$$

Ce τ_h est estimé par $\frac{m_h^b}{m_h^n}$. Finalement, on obtient le poids initial au niveau du ménage noté ptm_init_i :

$$\begin{aligned} ptm_init_{i,h} &= ptm_ech_i \cdot w_i^{NR} \\ &= \frac{M_h}{m_h^b} \cdot \frac{m_h^b}{m_h^n} = \frac{M_h}{m_h^n}, \forall i \in s_h^{n,M} \end{aligned} \tag{2}$$

4.3 Poids initial individu : étapes 4 et 5

Une fois le ménage U_i documenté, la probabilité d'inclusion π_k d'un individu k dans l'échantillon est égale à $1/N_i$. On choisit 1 individu parmi N_i. Le poids initial de la personne cible, noté pti_init_k (au niveau individuel) s'écrit donc :

$$\begin{aligned} pti_init_{k,h,i} &= ptm_init_{i,h} \cdot \frac{1}{\pi_k}, \\ &= \frac{M_h}{m_h^n} \cdot N_i \qquad \text{avec } k \in U_i, U_i \in h. \end{aligned} \tag{3}$$

Afin de ne pas alourdir inutilement le texte et comme une seule personne par ménage peut participer à l'enquête, nous dénoterons désormais par N_k la taille du ménage de la personne k ($N_k = N_i$ si $k \in U_i$). Le poids initial au niveau individu se notera alors :

$$pti_init_{k,h} = \frac{M_h}{m_h^n} N_k.$$

Avant de procéder aux étapes ultérieures de la construction des pondérations, on a encore multiplié les poids initiaux par un facteur, noté f_{ind}^{corr}, pour que leur somme corresponde à la taille de la population[4]. On obtient alors les poids initiaux individuel $pti_init2_{k,h}$:

$$pti_init2_{k,h} = pti_init_{k,h} \cdot f_{ind}^{corr}$$

Cet ajustement n'est pas une nécessité à ce niveau, mais il permet de mieux interpréter les modèles de correction de la non-réponse développés dans la suite. Il sera impératif pour les calages finaux (aux niveaux "lune" et "étoile", voir section 6) qui modélisent la non-réponse tout en respectant certains totaux au niveau de la population pour des variables-clé. Le poids initial $pti_init2_{k,h}$ d'un individu k dépend :

- de la strate h (canton) dans laquelle il/elle se trouve,
- du nombre d'unités primaires (ménages) existantes dans cette strate (M_h),
- du nombre d'unités primaires sélectionnées et répondantes dans la strate h (m_h^n)
- du nombre d'unités secondaires N_k dans l'unité primaire, c'est-à-dire du nombre de personnes éligibles dans le ménage de k.

La somme des poids initiaux dans l'échantillon fournit une estimation du nombre de personnes de 15ans et plus vivant en Suisse.
Le tableau 5 fournit, pour l'échantillon des répondants à l'enquête téléphonique, la répartition des ménages selon le nombre de personnes de 15 ans et plus qui y vivent.

N_k	Fréquence	%	Fréq. cumulée	%
1	6379	28.77	6379	28.77
2	11771	53.08	18150	81.85
3	2293	10.34	20443	92.19
4	1375	6.2	21818	98.39
5	296	1.33	22114	99.73
6	44	0.2	22158	99.93
7	13	0.06	22171	99.99
8	3	0.01	22174	100

T 5 N_k : nbr. de personnes de 15ans+ vivant dans les ménages ayant participé à l'enquête téléphonique.

Rappelons encore que le registre dans lequel on a "tiré"[5] les adresses est mis à jour trimestriellement. Il est donc inévitable d'avoir quelques cas de déménagements entre les strates dans notre échantillon stratifié. C'est-à-dire qu'un ménage dont on a tiré le numéro peut avoir déménagé entre sa sélection et le passage dans l'enquête. Ces mouvements de population ont été pris en compte de la manière suivante : le calcul des poids initiaux a été fait en gardant la strate initiale. L'exploitation des données a cependant été faite dans les cantons où habitaient les personnes au moment de l'interview. Ainsi les données de l'ESS07 contiennent deux variables pour le canton :

- La variable `Kanton` (livrée) correspond au canton dans lequel la personne interviewée habite au moment de l'interview (variable de post-stratification),
- La variable `Kanton_1` (non-livrée) correspond au canton dans lequel la personne habite (habitait) selon le plan d'échantillonnage (variable de stratification).

Voir aussi les section 8.1 et 10.3 pour plus de détails à ce sujet.

[4] selon la statistique ESPOP, état décembre 2006, corrigée pour refléter la population résidant dans des ménages privés (donc sans les personnes institutionnalisées, en prison, etc.).

[5] En fait on tire des numéros de téléphone dans le SRH (StichprobenRegister für Haushalterhebungen).

5 Modélisation de la non-réponse par segmentation

Dans l'ESS07, la méthode de segmentation utilisée permet d'analyser et modéliser d'une part le mécanisme de réponse des individus au questionnaire téléphonique parmi les ménages documentés, d'autre part celui de réponse des individus au questionnaire écrit parmi les répondants au questionnaire téléphonique (CATI).

5.1 La méthode de segmentation

Une fois les poids initiaux calculés, il s'agit de calculer un facteur d'ajustement pour la non-réponse (totale) de chaque unité (ménage ou individu selon les cas) considérée comme répondante. Le facteur d'ajustement pour la non-réponse appliqué à chaque unité répondante est défini comme étant l'inverse du taux de réponse du groupe homogène de réponse (GHR) dans lequel l'unité se situe. Les GHR sont formés en regroupant des unités dont les caractéristiques explicatives de la non-réponse sont similaires. Si les GHR sont définis de sorte que la non-réponse soit complètement aléatoire à l'intérieur d'un groupe, alors le biais dû à la non-réponse est négligeable (voir [12], [4]). Pour trouver un ajustement adéquat des unités répondantes, afin de pouvoir extrapoler les résultats à l'ensemble de la population, on procède à une modélisation de la non-réponse.

La première étape de modélisation de la non-réponse consiste à déterminer quelles sont les variables explicatives qui doivent être considérées dans le modèle afin de bien prédire la variable dichotomique indicatrice de réponse. Le choix des variables est souvent assez restreint car on doit détenir l'information tant pour les unités répondantes que pour les non-répondantes. Il s'agit de dresser une liste des variables disponibles pouvant décrire le processus de non-réponse. La nature de cette liste varie selon les thèmes abordés par l'enquête, le mode d'interview, le plan d'échantillonnage, etc. Tout facteur susceptible d'expliquer la non-réponse doit être pris en compte dans la mesure du possible. Les variables peuvent êtres des grandeurs mesurées (à une phase précédente) dans l'enquête elle-même ou des informations auxiliaires que l'on peut obtenir par d'autres moyens (par exemple par des registres) sur les unités répondantes et non-répondantes.
Notons encore que les variables catégorielles dont le nombre de modalités (m) est supérieur à 2 ainsi que les variables de comptage[6] doivent être recodées en m variables dichotomiques.

La méthode utilisée pour la modélisation de la non-réponse est l'analyse par segmentation (voir [6]). Cette méthode est utilisée pour les deux ajustements nécessaires dus à la non-réponse et présentés dans ce document. La modélisation par segmentation utilise un processus itératif pour partitionner le fichier de données et ainsi former l'arbre de décision. Le premier noeud de l'arbre de décision est tout simplement l'ensemble complet des données. La première itération définit le premier embranchement de l'arbre. Ce dernier est créé en déterminant la variable qui influence le plus significativement sur la variable réponse parmi les variables retenues pour la modélisation. Une mesure statistique reflétant l'importance de la variable pour discriminer la non-réponse est calculée pour chaque variable ; d'après cette mesure, la variable qui est la plus importante est choisie. Le premier embranchement est ainsi formé, créant un noeud à chaque extrémité. Pour chaque noeud formé, une nouvelle itération est effectuée en répétant le même mécanisme. De nouveaux embranchements et de nouveaux noeuds sont ainsi créés. Si pour un noeud donné, aucune variable n'est significative pour expliquer la non-réponse, ou qu'un des critères d'arrêt de l'algorithme est rencontré, aucun autre embranchement n'est créé.

Plusieurs méthodes peuvent être employées pour déterminer la variable qui influence le plus sur la propension à répondre (voir [6]). Nous utilisons ici l'algorithme de CHAID[7], qui sélectionne la variable

[6] Par exemple le nombre de personnes du ménage.
[7] CHAID : Chi-Square Automatic Interaction Detection, algorithme inventé par Kass (voir [5]).

qui conduit à une valeur de la statistique du chi-deux de Pearson la plus élevée.
On utilise une macro SAS qui fut développée par Jean-François Naud de Statistique Canada lorsque ce dernier reçut le mandat d'effectuer les pondérations du Panel Suisse de Ménages[8].

5.2 Modélisation de la non-réponse au questionnaire téléphonique : étapes 6 et 6b

On crée un fichier contenant tous les ménages dont un individu a répondu au questionnaire téléphonique (CATI ou CAPI) ainsi que tous les ménages qui ont seulement pu être documentés (et pour lesquels la personne cible a refusé l'interview). On définit la variable[9] de réponse à l'interview téléphonique :

$$CODEREP = \begin{cases} 4 & \textit{pour les répondants,} \\ 3 & \textit{pour les non-répondants.} \end{cases}$$

Il s'agit ensuite de créer des variables dichotomiques que l'on pense être pertinentes pour expliquer le mécanisme de réponse. On est bien sûr limité à l'information disponible pour les répondant et pour les non-répondants. On contrôle également qu'il n'y pas de trop grandes corrélations (> 0.5) entre ces variables dichotomiques et qu'elles identifient toutes des groupes d'observations d'au moins 30 unités dans le cas qui nous occupe ici. On a soumis les variables retenues au modèle de segmentation avec les contraintes d'un taux de réponse pondéré d'au moins 50% et un effectif minimal de 30 unités pour chaque GHR. Ce sont là des critères d'arrêt à paramétrer. Si l'on permet un taux de réponse trop bas et/ou un nombre minimal d'observations par GHR trop petit, l'arbre de segmentation peut devenir très grand et détaillé, mais aussi instable. A l'inverse des critères d'arrêt de la segmentation trop sévères ne permettent qu'une modélisation très grossière du mécanisme de réponse. Il s'agit donc de trouver un bon consensus en testant différents critères d'arrêt.
Le tableau 6 liste les variables dichotomiques qui ont finalement été retenues dans modèle obtenu. L'arbre de segmentation obtenu est présenté dans l'annexe A. Le modèle de non-réponse créé avec l'analyse par segmentation a conduit à la constitution de 54 GHR. Le tableau 7 donne quelques informations sur le facteur d'ajustement ainsi calculé et appliqué aux poids intiaux pti_init2_k pour obtenir les poids corrigés pour la non-réponse à l'interview téléphonique pti_nrqi_k.

[8] www.swisspanel.ch
[9] Ce nom et ces codes étant imposés par les besoins de la macro.

Nom	signification des codes
AC_1	=1 pour les personnes de 15 à 19 ans, 0 sinon
AC_2	=1 pour les personnes de 20 à 24 ans, 0 sinon
AC_3	=1 pour les personnes de 25 à 34 ans, 0 sinon
AC_4	=1 pour les personnes de 35 à 44 ans, 0 sinon
AC_8	=1 pour les personnes de 75+ ans, 0 sinon
HH_1	=1 pour les ménages de 1 (NBPERS=1) personne, 0 sinon
HH_2	=1 pour les ménages de 2 (NBPERS=2) personnes, 0 sinon
HH_3	=1 pour les ménages de 3 (NBPERS=3) personnes, 0 sinon
NOKIDS	=1 si pas d'enfants dans le ménage, 0 sinon
PLUS_VIEUX	=1 si la personne est la plus âgée du ménage, 0 sinon
STRATNUM_1	=1 si le ménage appartient à la strate d'échantillonnage Zurich ville, 0 sinon
STRATNUM_2	=1 si le ménage appartient à la strate d'échantillonnage Zurich reste, 0 sinon
STRATNUM_20	=1 si le ménage appartient à la strate d'échantillonnage St Gall, 0 sinon
STRATNUM_21	=1 si le ménage appartient à la strate d'échantillonnage Grisons, 0 sinon
STRATNUM_26	=1 si le ménage appartient à la strate d'échantillonnage Valais, 0 sinon
STRATNUM_28	=1 si le ménage appartient à la strate d'échantillonnage Genève, 0 sinon
STRATNUM_3	=1 si le ménage appartient à la strate d'échantillonnage Berne Oberland, 0 sinon
STRATNUM_5	=1 si le ménage appartient à la strate d'échantillonnage Berne reste, 0 sinon
STRATNUM_6	=1 si le ménage appartient à la strate d'échantillonnage Lucerne, 0 sinon
WELLE_1	=1 si le ménage fait partie de la première vague d'enquête, 0 sinon
WELLE_2	=1 si le ménage fait partie de la deuxième vague d'enquête, 0 sinon
WELLE_3	=1 si le ménage fait partie de la troisième vague d'enquête, 0 sinon
WELLE_4	=1 si le ménage fait partie de la quatrième vague d'enquête, 0 sinon

T 6 Définition des variables dichotomiques intervenant dans la modélisation de la non-réponse au niveau individuel dans le cadre de la modélisation de la non-réponse au questionnaire téléphonique.

Moyenne	1.18	Min	1.05
Médiane	1.17	Max	1.74
Écart-type	0.09	Étendue	0.69

T 7 Statistiques descriptives du facteur d'ajustement *f_aj_nrTel* pour les 18760 ménages répondants à l'interview CATI ou CAPI (ceux pour qui *CODEREP* = 4). $pti_nrqi_k = pti_init2_k \cdot f_aj_nrTel_k$.

5.3 Modélisation de la non-réponse au questionnaire écrit : étapes 7 et 7b

On crée un fichier contenant tous les ménages dont un individu a répondu au questionnaire téléphonique (CATI sans les interviews PROXY[10]) avec l'indication de si la personne cible a ensuite aussi complété le questionnaire écrit. On définit la variable[11] de réponse à l'interview écrite :

$$CODEREPECT = \begin{cases} 4 & \textit{pour les répondants,} \\ 3 & \textit{pour les non-répondants.} \end{cases}$$

Il s'agit ensuite de créer des variables dichotomiques qu'on pense être pertinentes pour expliquer le mécanisme de réponse. On est bien sûr à nouveau limité à l'information disponible pour les répondant et pour les non-répondants. Cependant, à ce niveau, ladite information est bien plus riche que pour la modélisation de la non-réponse exposée au paragraphe précédant. En effet, dans le cas présent, les non-répondants au questionnaire écrit on tous tout de même répondu à l'enquête téléphonique. Autrement dit, on en sait beaucoup plus à leur sujet. Par exemple, dans les variables dichotomiques supplémentaire dans la modélisation de la non-réponse à l'enquête écrite on a ajouté des variables au sujet de la formation, de l'état de santé, du nombre d'appels ayant été nécessaires pour pouvoir faire l'interview CATI, l'aptitude à lire, la durée de l'interview CATI, etc.
Pour toutes les variables binaires crées, on contrôle également qu'il n'y pas de trop grandes corrélations[12] (> 0.5) entre elles et qu'elles identifient toutes des groupes d'observations d'au moins 30 unités dans le cas qui nous occupe ici. On a soumis les variables retenues au modèle de segmentation avec les contraintes d'un taux de réponse pondéré d'au moins 60% et un effectif minimal de 50 unités pour chaque GHR.
Le tableau 8 liste les variables dichotomiques qui ont finalement été retenues dans modèle obtenu. L'arbre de segmentation obtenu est présenté dans l'annexe B. Le modèle de non-réponse créé avec l'analyse par segmentation a conduit à la création de 71 GHR. Le tableau 9 donne quelques informations sur le facteur d'ajustement ainsi calculé et appliqué aux poids corrigés pour la non-réponse à l'enquête téléphonique pti_nrqi_k pour obtenir les poids corrigés pour la non-réponse au questionnaire écrit pti_nrqe_k.

[10] En effet, ces dernières sont non-éligibles pour l'enquête écrite.
[11] Ce nom et ces codes étant imposés par les besoins de la macro.
[12] En pratique une corrélation de 0.5 est encore loin d'être vue comme une colinéarité. Cependant des variables corrélant à 0.5 ou plus déstabilisent facilement l'arbre de segmentation...

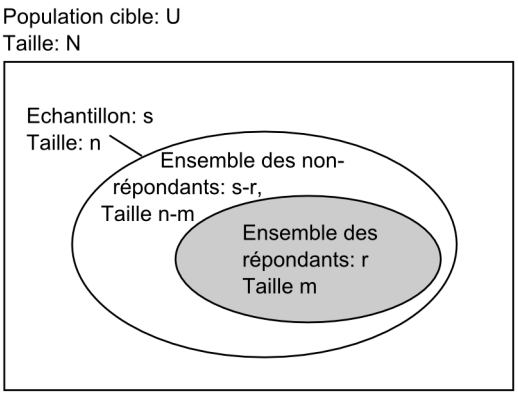
Nom	signification des codes
AC_1	=1 pour les personnes de 15 à 19 ans, 0 sinon
AC_2	=1 pour les personnes de 20 à 24 ans, 0 sinon
AC_3	=1 pour les personnes de 25 à 34 ans, 0 sinon
AC_4	=1 pour les personnes de 35 à 44 ans, 0 sinon
AC_5	=1 pour les personnes de 45 à 54 ans, 0 sinon
AC_6	=1 pour les personnes de 55 à 64 ans, 0 sinon
AC_7	=1 pour les personnes de 65 à 74 ans, 0 sinon
ALLEM	=1 si langue d'interview était l ?allemand (tkont10=1), 0 sinon
CoupleAvEnf	=1 si tsode61a=3, 0 sinon
CoupleSaEnf	=1 si tsode61a=2, 0 sinon
Form1	=1 si bildun3=1, 0 sinon
Form2	=1 si bildun3=2, 0 sinon
Form3	=1 si bildun3=3, 0 sinon
HH_3	=1 pour les ménages de 3 (NBPERS=3) personnes, 0 sinon
IntLenght_5	=1 si interview téléphonique réalisée en plusieurs fois, 0 sinon
intLength_2	=1 si interview téléphonique a duré 21< CDAUERN <=42 minutes, 0 sinon
IntLength_4	=1 si interview téléphonique a duré plus de 58 minutes (58< CDAUERN), 0 sinon
LIRE_1	=1 si TSBHD03=1 or TSBHD03<0, 0 sinon
Men1Pers	=1 si tsode61a=1, 0 sinon
MenAutre	=1 si tsode61a>=4, 0 sinon
NBAPP_1	=1 si nombre d'appels NBAPP <=4, 0 sinon
NBAPP_2	=1 si nombre d'appels 4 < NBAPP <=10, 0 sinon
NBAPP_3	=1 si nombre d'appels 10< NBAPP, 0 sinon
PLUS_VIEUX	=1 si la personne est la plus âgée du ménage, 0 sinon
SANTE_1	=1 si TSUBG03 in (1 2) or TSUBG03<0, 0 sinon
SANTE_2	=1 si TSUBG03=3, 0 sinon
SANTE_3	=1 si TSUBG03 in (4 5), 0 sinon
STRATNUM_1	=1 si le ménage appartient à la strate d'échantillonnage Zurich ville, 0 sinon
STRATNUM_2	=1 si le ménage appartient à la strate d'échantillonnage Zurich reste, 0 sinon
STRATNUM_22	=1 si le ménage appartient à la strate d'échantillonnage Argovie, 0 sinon
STRATNUM_26	=1 si le ménage appartient à la strate d'échantillonnage Valais, 0 sinon
STRATNUM_28	=1 si le ménage appartient à la strate d'échantillonnage Genève, 0 sinon
STRATNUM_5	=1 si le ménage appartient à la strate d'échantillonnage Berne reste, 0 sinon
WELLE_1	=1 si le ménage fait partie de la première vague d'enquête, 0 sinon
WELLE_2	=1 si le ménage fait partie de la deuxième vague d'enquête, 0 sinon
WELLE_4	=1 si le ménage fait partie de la quatrième vague d'enquête, 0 sinon

T 8 Définition des variables dichotomiques intervenant dans la modélisation de la non-réponse au niveau individuel dans le cadre de la modélisation de la non-réponse au questionnaire écrit.

Moyenne	1.25	Min	1.04
Médiane	1.20	Max	1.65
Écart-type	0.16	Étendue	0.61

T 9 Statistiques descriptives du facteur d'ajustement *f_aj_nrEct* pour les 14393 individus répondants questionnaire écrit (ceux pour qui *CODEREPECT* = 4).
$pti_nrqe_k = pti_nrqi_k \cdot f_aj_nrEct_k$.

6 Introduction théorique sur le calage utilisé



G 3 Schéma classique de la non-réponse dans un échantillon.

6.1 Notations

Soit Y une variable d'intérêt mesurée sur un échantillon s. Si des valeurs y_k de Y ne sont relevées que sur une partie de s (comme représenté schématiquement dans la figure 3), du fait de la non-réponse, on risque d'estimer le total $Y = \sum_U y_k$ avec un biais.

- On désigne la probabilité d'inclusion de k dans s par π_k (π_k correspondra pour nous à pti_init2_k).
- On note $q(\cdot|s)$ le mécanisme de réponse conditionné à s. Il s'agit d'un plan de sondage sur s.
- Soit R_k la variable aléatoire binaire, indicatrice de réponse (0=pas de réponse, 1=réponse), pour $k \in s$. Il s'agit d'une variable de Bernoulli.
- On note $\theta_k = E_q(R_k|s)$ la probabilité de réponse.
- On fait l'hypothèse que θ_k n'est pas fonction de s. Ainsi, il existe un vecteur θ de probabilités de réponse.

6.2 Approche de la pondération en deux phases pour corriger la non-réponse

L'idée générale de cette approche est la suivante : la sélection d'un échantillon s est suivie de la réalisation d'un ensemble de répondants r vu comme un sous-échantillon de s (voir figure 3).
Si l'on admet que la distribution de réponse $q(r|s)$ est connue, alors $P(k\&l \in r|s) = \theta_{kl}$ et les θ_k sont connus. Dans ce sas, si les θ_k sont > 0, le total $Y = \sum_U y_k$ est estimable par $\hat{Y} = \sum_r \frac{1}{\pi_k}\frac{1}{\theta_k} y_k = \sum_r \frac{d_k}{\theta_k} y_k$, $k \in r$. Ce qui est l'estimateur de Horvitz-Thompson dans le cas de la sélection en deux phases.
Mais θ_k est toujours inconnu pratique. On estime donc θ_k par $\hat{\theta}_k$, avec $k \in r$ et $d_k = \frac{1}{\pi_k}$ le poids de sondage, grâce à de l'information auxiliaire et une modélisation de $q(\cdot|s)$. On obtient alors l'estimation du total Y suivante : $\hat{Y} = \sum_r \frac{d_k}{\hat{\theta}_k} y_k$.
Dans ce cas le facteur $\frac{d_k}{\hat{\theta}_k}$ joue le rôle de poids pour l'élément répondant k. Cette approche structurée suppose :

(i) l'existence d'un mécanisme de réponse,

(ii) qu'on peut formuler un modèle réaliste pour ce mécanisme dans lequel θ_k apparaît comme paramètre inconnu,

(iii) qu'on estime θ_k en utilisant des variables (de l'information) auxiliaires et le fait qu'un sous-ensemble de s a été observé comme répondant.

Un modèle fréquemment utilisé consiste à former des groupes homogènes de réponse (GHR) : on partitionne la population en des groupes distincts de telle façon que les unités d'un même groupe répondent avec la même probabilité et de manière indépendante (voir section 5.1). L'information auxiliaire est utilisée ici pour classifier les répondants et les non-répondants dans ces GHR. On remplace alors θ_k par le paramètre estimé $\hat{\theta}_k$ dans chaque GHR.

6.3 Vecteurs et information auxiliaire

L'utilisation efficace d'information auxiliaire est la clé pour une estimation correcte de la non-réponse. L'information auxiliaire est transmise par un vecteur auxiliaire. Le *vecteur auxiliaire* est un vecteur dont les valeurs sont connues pour tous les éléments répondants, $k \in r$, et l'on dispose de l'information sur ce vecteur pour un ensemble plus grand que r. Cette information sera utilisée tant pour la protection contre la non-réponse que pour la réduction de la variance. Nous dénotons le vecteur auxiliaire par X et sa valeur pour un élément k, x_k.
L'information pour le plus grand de ces ensembles fournit l'input d'information nécessaire au calcul des poids calés.
Pour une présentation des techniques de calage, on peut se référer à [1]. L'approche par calage de C.-E. Särndal distingue trois cas différents appelés InfoU, infoS et infoUS, dépendants de l'information disponible :

– *InfoU* : Information disponible au niveau de la population U. On dénote par $\mathbf{x}_k^*$ un vecteur de dimension $J^* \geqslant 1$, dit le vecteur *étoile*, tel que :
  (i) le total sur la population $\mathbf{X}_k^* = \sum_U \mathbf{x}_k^*$ est connu,
  (ii) $\forall k \in r$, le vecteur $\mathbf{x}_k^*$ est connu.
– *InfoS* : Information disponible au niveau de l'échantillon s, sans que ladite information ne soit disponible au niveau de la population. On dénote par $\mathbf{x}_k^\circ$ un vecteur de dimension $J^\circ \geqslant 1$, dit le vecteur *lune*, tel que :
  (i) $\forall k \in s$, le vecteur $\mathbf{x}_k^\circ$ est connu, alors que la somme sur la population $\sum_U \mathbf{x}_k^\circ$ est inconnue,
  (ii) $\forall k \in r$, le vecteur $\mathbf{x}_k^\circ$ est connu.

  Le vecteur lune contient de l'information à un niveau plus bas que le vecteur étoile.
– *InfoUS* : Les deux types d'informations InfoU et InfoS sont disponibles, et utilisés de manière combinée pour le calcul des poids. On peut alors formuler le vecteur auxiliaire comme un agrégat $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^\circ \\ \mathbf{x}_k^* \end{pmatrix}$ de dimension $J^\circ + J^*$.

Nous renvoyons le lecteur à la section 7 pour des éclaircissements et applications pratiques sur l'ESS07 des concepts présentés ici. Le tableau 10 résume les définitions développées ci-dessus.

Type d'information	Vecteur auxiliaire $\mathbf{x}_k$	Input d'information $\mathbf{X}$
InfoU	$\mathbf{x}_k^*$	$\mathbf{X}_k^* = \sum_U \mathbf{x}_k^*$
InfoS	$\mathbf{x}_k^\circ$	$\mathbf{X}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$
InfoUS	$\begin{pmatrix} \mathbf{x}_k^\circ \\ \mathbf{x}_k^* \end{pmatrix}$	$\begin{pmatrix} \mathbf{X}_k^\circ \\ \mathbf{X}_k^* \end{pmatrix}$

T 10 Définitions et notations pour l'information auxiliaire.

Malgré la non-réponse, on aimerait que l'estimateur de $Y = \sum_U y_k$ ait les propriétés suivantes :

(i) un biais petit,

(ii) une petite variance,

(iii) un système de poids qui reproduise l'input d'information auxiliaire lorsqu'il est appliqué aux variables auxiliaires.

(iv) un système de poids utilisable pour l'estimation du total de chacune des variables y de l'enquête.

6.4 Le calage

Le but d'une correction à posteriori comme le *calage sur marges* est d'améliorer l'estimation en diminuant les biais dus à la non-réponse. Toutes les informations disponibles simultanément sur l'échantillon et la population peuvent alors être utilisées.
Comme annoncé plus haut de le texte, dans le cadre des pondérations de l'ESS07, nous avons utilisé un calage "lune et étoile" qui est détaillé ci-après. Nous commençons cependant par rappeler brièvement la technique du calage sur marges dans le cas "classique" où l'on utilise que de l'information auxiliaire au niveau de la population, i.e. au niveau "étoile" uniquement.

6.4.1 La méthode "classique"

On ne dispose ici que de l'information *InfoU*. Les totaux sur la population $\mathbf{X}_k^*$ sont donc connus pour $J^* \geqslant 1$ variables auxiliaires. On aimerait estimer le total de la variable d'intérêt Y. Les totaux sur la population sont une information dont on aimerait tenir compte. L'approche par calage produit un estimateur par calage de Y, noté $\hat{Y}_c$, et un estimateur de la variance correspondant, noté $\hat{V}(\hat{Y}_c)$. L'index c utilisé suggère *calibré*.
Les poids obtenus après calage sont notés w_k, $k \in r$. L'estimateur par calage est donné par :

$$\hat{Y}_c = \sum_{k \in r} w_k y_k,$$

où les poids w_k sont proches (dans un sens à préciser) des poids initiaux d_k, et vérifient les *contraintes de calage* :

$$\sum_{k \in r} w_k x_{jk} = \mathbf{X}_{jk}^* \ , \ j = 1, ..., J^*.$$

La capacité des w_k à protéger efficacement contre le biais dû à la non-réponse dépend de la force de l'information auxiliaire.
On cherche donc un estimateur "peu différent" de l'estimateur de Horvitz-Thompson qui *cale* l'échantillon sur les totaux des variables auxiliaires.
Avant de réaliser le calage, il faut choisir une *fonction de distance* G, d'argument w_k/d_k, pour mesurer les distances entre w_k et d_k. La fonction G doit vérifier les conditions suivantes : elle doit être positive et convexe avec $G(1) = G'(1) = 0$.
Une fois la fonction G choisie, le problème consiste à déterminer les poids $w_k (k \in r)$ solutions du problème d'optimisation sous contraintes suivant :

$$\min_{w_k} \sum_{k \in s} d_k G(\frac{w_k}{d_k}) \text{ sous la contrainte } \sum_{k \in r} w_k x_k = \mathbf{X}^*. \tag{4}$$

On cherche ainsi à minimiser une somme pondérée (par les d_k) des "distances" entre les poids initiaux d_k et les pondérations recherchées w_k, sous les contraintes de calage.
La fonction de distance utilisée pour l'ESS est celle correspondant à la méthode *raking ratio* et les résultats sont obtenus par l'algorithme *iterative proportional fitting* (voir [8]).

6.4.2 La méthode "lune et étoile"

Dans le cas général, InfoUS, où le vecteur auxiliaire est $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^{\circ} \\ \mathbf{x}_k^{*} \end{pmatrix}$ de dimension $J^{\circ} + J^{*}$ et l'apport d'information est $\mathbf{X} = \begin{pmatrix} \mathbf{X}_k^{\circ} \\ \mathbf{X}_k^{*} \end{pmatrix}$, on cherche un système de poids w_k pour $k \in r$ qui satisfassent les équations de calage :

$$\sum_r w_k \mathbf{x}_k = \mathbf{X}. \tag{5}$$

L'équation (5) implique :

$$\sum_r w_k \mathbf{x}_k^{*} = \sum_U \mathbf{x}_k^{*}, \quad et$$
$$\sum_r w_k \mathbf{x}_k^{\circ} = \sum_s d_k \mathbf{x}_k^{\circ}.$$

Les poids qui satisfont (5) sont dits calibrés à l'apport d'information $\mathbf{X}$ ou calibrés sur $\mathbf{X}$. De part l'échantillonnage, chaque élément k possède un poids $d_k = \frac{1}{\pi_k}$. Avec la non-réponse, les poids d_k pour $k \in r$ sont trop petits en moyenne pour produire des estimations acceptables. La somme pondérée $\sum_r d_k y_k$ sous-estime clairement le total $Y = \sum_U y_k$. Les d_k doivent être gonflés. On cherche de nouveau poids qui sont plus grands que les d_k, pour toutes, ou au moins pour la plupart, des unités répondantes. Notons v_k le facteur, par lequel on multiplie les d_k pour obtenir le nouveau poids de l'élément k : $w_k = d_k v_k$.
Les v_k doivent refléter les caractéristiques individuelles connues de l'élément $k \in r$ contenues dans la valeur du vecteur $\mathbf{x}_k$. Plusieurs formes sont possibles pour les v_k. Une forme simple est donnée par $v_k = 1 + \boldsymbol{\lambda}^{'} \mathbf{x}_k$ où λ est un vecteur à déterminer. Dans cette formulation, on laisse v_k dépendre linéairement des valeurs connues $\mathbf{x}_k$ pour l'élément k. C'est la méthode linéaire, une des options possible pour effectuer le calage via les macros SAS CALMAR développées en France par l'INSEE.
Pour trouver les poids calibrés $w_k = d_k + d_k \boldsymbol{\lambda}_r^{'} \mathbf{x}_k$, il faut déterminer le $\boldsymbol{\lambda}$, noté $\boldsymbol{\lambda}^{'}$ qui satisfait les contraintes de calage en résolvant (5). On obtient $\boldsymbol{\lambda}^{'} = \boldsymbol{\lambda}_r^{'}$ où :

$$\boldsymbol{\lambda}_r' = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \tag{6}$$

pour autant que la matrice $\sum_r d_k \mathbf{x}_k \mathbf{x}_k'$ soit inversible. L'estimateur calibré (ou calé) qui en résulte est :

$$\hat{Y}_c = \sum_r w_k y_k \tag{7}$$

A part l'apport d'information $\mathbf{X}$, le calcul des w_k nécessite d'effectuer deux sommes sur l'ensemble des répondants r. Toutes les deux peuvent être calculées puisque $\mathbf{x}_k$ est connu pour tout $k \in r$. La matrice $(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')$ qui doit être inversée est de dimension $(J^{*} + J^{\circ}) x (J^{*} + J^{\circ})$.
La méthode de calage (en fait la fonction de distance) utilisée dans le cadre des pondération de l'ESS07 n'est pas la méthode linéaire car celle-ci est susceptible de produire des poids négatifs, ce qui est difficilement concevable pour une enquête auprès des ménages ou personnes. On a donc préféré utiliser la méthode du raking ratio ou iterative proportional fitting (comme pour l'ESS02) qui ne fournit que des poids strictement positifs. Dans ce cas, on pose la forme suivante pour les v_k : $v_k = exp(\boldsymbol{\lambda}^{'} \mathbf{x}_k)$. Les nouveaux poids sont alors donnés par $w_k = d_k exp(\boldsymbol{\lambda}^{'} \mathbf{x}_k)$ après qu'on ait résolu pour $\boldsymbol{\lambda}^{'}$ l'équation (5) :

$$\sum_r d_k \mathbf{x}_k exp(\boldsymbol{\lambda}^{'} \mathbf{x}_k) = \mathbf{X}. \tag{8}$$

Un cas particulier de la méthode du raking ratio est le calage sur marges. Dans ce cas les composantes x_{ki} du vecteur $\mathbf{x}_k$ valent 0 ou 1 selon que l'unité i appartienne à la sous-population $U_i \subset U$. On a alors pour $k \in U$:

$$w_k = d_k \prod_{i|U_i \ni k} exp(\lambda_i). \tag{9}$$

7 Calcul des poids finaux

7.1 Pondération de l'enquête téléphonique

Afin de corriger l'effet de la non-réponse, un calage (raking ratio) a été effectué sur les différents types de marges suivantes :

- *Niveau lune* (InfoS) : marges informant sur le mécanisme de réponse au niveau individuel : 48 marges définies par les sommes des poids pti_init2_k sur les 22174 ménages documenté pour les 48 groupes homogènes de réponse obtenus par l'arbre de segmentation décrit dans la section 5.2 ainsi que dans l'annexe A. Les valeurs de ces marges sont données dans tableau 27 et correspondent aux sommes des poids initiaux pti_init2_k (répondants + non-répondants à l'enquête CATI/CAPI). Contrairement à la manière classique de faire (étapes 6b et 6c dans la figure 2) on n'utilise pas les poids corrigés pour la non-réponse pti_nrqi_k produits. Toute l'information auxiliaire au niveau lune, $\mathbf{X}_k^{\circ}$ ou *InfoS*, est contenue dans les marges (sommes des pti_init2_k) ainsi calculées.
- *Niveau étoile* (InfoU) :
  - 54 marges contenant les informations socio-démographiques (définies dans le tableau 12).
  - 18 marges géographiques définies dans le tableau 15.

  Pour définir les marges de calage sociodémographiques (voir tableau 12) on a pris tous les croisements possibles entre les tranches d'âge (15-19, 20-24, ..., 65-75 et 75+), le genre, la nationalité (suisse ou non-suisse) et le fait d'être marié ou non. La distinction entre les tranches d'âges 15-19 et 20-24 est due au fait, qu'en pratique, on observe souvent un comportement de réponse différent pour ces deux groupes. Il en est de même pour les trois dernières tranches d'âges utilisées. Dans certains cas on a dû regrouper certaines cellules par rapport à ces croisements, notamment lorsqu'on tombait en-dessous de 30 observations dans l'échantillon des répondants (18760 unités, voir tableau 14).

  Les marges géographiques correspondent aux strates d'échantillonnage pour les régions ou cantons ayant augmenté la taille de leur échantillon. Les cantons n'ayant pas financé d'interviews supplémentaires ont été regroupés dans ce qui restait de la grande région (selon la nomenclature NUTSII) qui les contient.

  Pour les marges sociodémographiques et les marges géographiques on a utilisé comme source la statistique ESPOP[13] (état déc. 2006) corrigée pour refléter la population des ménages privés (donc sans les personnes en institutions, prisons, etc.).

Notons que la somme de chacune de ces trois séries de 54, 18 et 48 marges vaut 6'186'711 et correspond à la taille de la population.

On a défini ainsi beaucoup plus de marges de calage que dans les pondérations de l'ESS02 et on a ainsi utilisé plus d'information auxiliaire ce qui tend à minimiser les biais lors de la production de résultats avec les données de l'ESS07. Notons qu'il faut tout de même se limiter à des marges qu'on pense être pertinentes car un nombre inutilement trop élevé de marges risque de faire augmenter l'erreur quadratique moyenne.

Pour des remarques et commentaires sur les poids finaux obtenus se reporter à la section 7.3.

[13] ESPOP est une statistique de synthèse, qui s'appuie sur les résultats du recensement fédéral de la population (RFP), de la statistique de la population résidante de nationalité étrangère (PETRA), de la statistique du mouvement naturel de la population (BEVNAT) et de la statistique des migrations des personnes de nationalité suisse.

Type d'information	Vecteur auxiliaire $\mathbf{x}_k$	Input d'information **X**
InfoU	54 marges sociodémographiques + 18 marges géographiques, $J^* = 72$	voir les tableaux 12 et 15, totaux de l'ESPOP
InfoS	48 GHR selon le modèle de non-réponse à l'enquête tél., voir la section 5.2, $J^\circ = 48$	voir tableau 27 et annexe A, sommes au niveau des ménages documentés

T 11 Définitions et notations pour l'information auxiliaire dans le cas de la pondération de l'enquête téléphonique pour l'ESS07.

âge	Hommes suisses mariés	Hommes suisses non-mariés	Hommes non-suisses mariés	Hommes non-suisses non-mariés	Femmes suisses mariés	Femmes suisses non-mariés	Femmes non-suisses mariés	Femmes non-suisses non-mariés
15-19		214282		51767		203253		45973
20-24	91358	160043	91452	41019	140620	149053	127998	27659
25-34		233235		71600		196294		48205
35-44	275695	162738	124192	45146	309485	147082	119065	33851
45-54	299908	114698	97402	22110	307349	132883	70946	19152
55-64	284902	83564	61300	13559	273956	124366	38900	14566
65-74	194733	46567	34337	12275	177562	117501	24931	14309
75+	136079	46925	11063		96409	191620		11775

T 12 Définition des marges sociodémographiques (ESPOP décembre 2006). Les cellules vides ont été exclues puisqu'aucune observation de l'échantillon des répondants n'y correspondait. Certaines cellules ont été fusionnées dû au peu d'observations correspondantes présentes dans l'échantillon des répondants.

âge	Hommes suisses mariés	Hommes suisses non-mariés	Hommes non-suisses mariés	Hommes non-suisses non-mariés	Femmes suisses mariés	Femmes suisses non-mariés	Femmes non-suisses mariés	Femmes non-suisses non-mariés
15-19	0	436	2	68	0	407	2	53
20-24	6	354	7	50	23	353	16	36
25-34	290	615	117	108	470	609	207	93
35-44	883	534	218	141	998	660	258	117
45-54	768	426	157	71	844	569	112	60
55-64	896	369	102	34	954	654	45	64
65-74	649	266	68	27	634	769	31	52
75+	435	265	39	23	294	892	14	46

T 13 Définition des marges sociodémographiques au niveau de l'échantillon pour la pondération de l'enquête téléphonique. Les cellules vides ont été exclues puisqu'aucune observation de l'échantillon des répondants n'y correspondait. Les cellules de moins de 30 observations ont été regroupées avec une (ou plusieurs) cellule(s) voisine(s), comme l'indique le tableau 14.

âge	Hommes suisses mariés	Hommes suisses non-mariés	Hommes non-suisses mariés	Hommes non-suisses non-mariés	Femmes suisses mariés	Femmes suisses non-mariés	Femmes non-suisses mariés	Femmes non-suisses non-mariés
15-19		MAR_6	MAR_50	MAR_19		MAR_30	MAR_53	MAR_41
20-24	MAR_49	MAR_7		MAR_20	MAR_52	MAR_31		MAR_42
25-34		MAR_8		MAR_21		MAR_32		MAR_43
35-44	MAR_1	MAR_9	MAR_14	MAR_22	MAR_25	MAR_33	MAR_38	MAR_44
45-54	MAR_2	MAR_10	MAR_15	MAR_23	MAR_26	MAR_34	MAR_39	MAR_45
55-64	MAR_3	MAR_11	MAR_16	MAR_24	MAR_27	MAR_35	MAR_40	MAR_46
65-74	MAR_4	MAR_12	MAR_17	MAR_51	MAR_28	MAR_36	MAR_54	MAR_47
75+	MAR_5	MAR_13	MAR_18		MAR_29	MAR_37		MAR_48

T 14 Définition de la variable `MAR` déterminant les marges sociodémographiques utilisées dans les calages pour les pondérations de l'ESS07. Ce tableau permet de lire le graphique 4.

	Canton	Strates d'échantillonnage	Strates géographiques de calage STRATE_CALAGE	Marges niveau population	Effectis niveau échantillon
1	1	Zurich ville	1	1066454	2370
2	1	Zurich reste			
3	2	Berne Oberland	2	799070	2378
4	2	Berne Jura			
5	2	Berne reste			
6	3	Lucerne	3	293158	1074
7	4	Uri	4	28563	468
8	5	Schwyz	5	112586	615
9	10	Fribourg	6	207103	1009
10	11	Soleure	7	206009	377
11	15	Appenzell r. ext.	8	42904	497
12	19	Argovie	9	471952	1267
13	21	Tessin	10	271973	1558
14	22	Vaud	11	537345	1365
15	23	Valais	12	242662	1079
16	24	Neuchâtel	13	137966	666
17	25	Genève	14	353889	1036
18	26	Jura	15	56131	514
19	12	Bâle ville	16	380616	708
20	13	Bâle campagne			
21	8	Glarus	17	830526	1513
22	14	Schaffhouse			
23	16	Appenzell r. int.			
24	17	St Gall			
25	18	Grisons			
26	20	Turgovie			
27	6	Obwald	18	147805	266
28	7	Nidwald			
29	9	Zoug			

T 15 Définition des marges géographiques utilisées dans les calages pour les pondérations de l'ESS07 en fonction des 29 strates d'échantillonnage. Les marges au niveau de l'échantillon indiquent la répartition des 18760 répondants à l'enquête téléphonique par rapport à chacune de ces marges ; elles sont calculées en utilisant la variable KANTON_1 (canton lors du tirage de l'échantillon).

Statistique	wght (= poidtrin_1854)	poidtrin	wecrit (= poidtrie_1854_48	poidtrie
Nombre d'obs.	18760.0	18760.0	14393.0	14393.0
Somme poids	6186711.0	6186711.0	6186711.0	6186711.0
Moyenne	329.8	329.8	429.8	429.8
Médiane	253.6	262.3	318.8	332.1
Écart-type	267.2	253.9	402.8	381.7
Étendue	4354	2808	8696	6063
Minimum	15.0	23.6	15.8	27.0
p1	28.8	33.3	33.4	41.1
p99	1287.3	1255.2	1950.2	1895.2
Maximum	4368.6	2831.5	8711.4	6090.0

T 16 Statistiques descriptives des poids finaux produits pour l'ESS07. Les `poidtrin_1854` sont les poids finaux (`wght` dans les données livrées) de l'enquête téléphonique. Les `poidtrin` sont les mêmes poids calculés selon un manière conventionnelle (i.e. selon les étapes 6a et 6b de la figure 2). De même les poids `poidtrie_1854_4871` sont les poids finaux (`wecrit` dans les données livrées) de l'enquête écrite. Les `poidtrie` sont les même poids calculés selon un manière conventionnelle (i.e. selon les étapes 6a, 7b et 7c de la figure 2).

Strate de calage	Moy.	Ecart-type	Max	P99	P95	Méd.	P5	P1	Min	Max/Min
1	450.0	286.5	3142.0	1394.3	980.6	420.4	132.7	118.1	88.2	35.6
2	336.0	291.5	2189.8	1349.6	895.1	260.2	34.6	21.1	15.6	140.4
3	273.0	162.5	1685.4	853.0	582.7	234.1	95.4	73.9	63.4	26.6
4	61.0	32.3	194.7	164.1	134.6	53.9	22.0	17.1	15.0	13.0
5	183.1	103.7	641.5	561.4	391.1	155.3	63.5	53.1	44.5	14.4
6	205.3	117.7	1034.8	605.1	421.8	169.0	68.5	61.0	46.2	22.4
7	546.4	368.4	4368.6	1815.8	1084.1	487.0	203.0	163.0	135.0	32.4
8	86.3	47.7	304.3	266.1	181.0	75.4	30.9	24.5	21.6	14.1
9	372.5	232.4	3999.2	1128.5	769.8	323.0	132.2	104.8	87.9	45.5
10	174.6	104.5	987.2	550.5	365.9	151.4	63.1	48.9	41.9	23.5
11	393.7	225.7	1826.4	1127.5	807.1	347.9	145.0	112.4	96.5	18.9
12	224.9	137.0	1394.9	773.4	476.6	189.0	77.3	59.9	51.4	27.1
13	207.2	121.2	950.4	637.7	419.3	186.9	78.0	60.4	51.8	18.3
14	341.6	214.5	1444.7	1053.5	782.7	293.5	112.6	87.3	74.9	19.3
15	109.2	58.0	343.7	293.7	214.0	97.0	38.4	33.0	26.3	13.1
16	537.6	305.6	2461.7	1546.9	1096.9	491.8	204.8	171.1	150.6	16.4
17	548.9	325.4	2619.0	1664.1	1187.2	460.0	178.2	140.0	106.2	24.7
18	555.7	300.0	1924.4	1672.0	1151.2	487.3	192.4	163.6	159.8	12.0

T 17 Caractéristiques des poids finaux `Poidtrin_1854` (=`wght`) pour les données du questionnaire téléphonique par strate de calage. Se reporter au tableau 15 pour la définition des strates.

Strate d'échan-tillon-nage	Moy.	Ecart-type	Max	P99	P95	Méd.	P5	P1	Min	Max/Min
ZH_Ville	329.4	206.9	1928.7	1065.5	688.4	291.0	120.8	102.8	88.2	21.9
ZH_Reste	533.4	303.8	3142.0	1512.2	1062.2	481.3	200.4	160.9	133.3	23.6
BE_Ober	279.7	152.0	942.6	827.2	574.0	251.4	99.6	75.9	68.1	13.8
BE_Jura	63.5	34.5	258.3	164.6	136.8	55.5	23.1	20.6	15.6	16.6
BE_Reste	524.5	298.2	2189.8	1529.4	1106.2	474.7	199.3	154.5	132.6	16.5
LU	273.0	162.5	1685.4	853.0	582.7	234.1	95.4	73.9	63.4	26.6
UR	61.0	32.3	194.7	164.1	134.6	53.9	22.0	17.1	15.0	13.0
SZ	183.1	103.7	641.5	561.4	391.1	155.3	63.5	53.1	44.5	14.4
OW	617.9	309.5	1761.8	1761.8	1162.6	494.8	207.4	178.5	178.5	9.9
NW	518.9	287.4	1339.9	1339.9	1096.4	462.9	163.6	159.8	159.8	8.4
GL	416.0	258.7	1394.9	1394.9	915.0	382.7	138.8	106.2	106.2	13.1
ZG	550.9	300.9	1924.4	1672.0	1129.4	478.1	192.4	175.4	171.4	11.2
FR	205.3	117.7	1034.8	605.1	421.8	169.0	68.5	61.0	46.2	22.4
SO	546.4	368.4	4368.6	1815.8	1084.1	487.0	203.0	163.0	135.0	32.4
BS	540.4	330.8	2461.7	1846.4	1120.5	495.8	201.6	171.6	150.6	16.4
BL	535.6	285.9	1967.8	1437.3	1063.4	491.5	204.8	171.1	151.3	13.0
SH	522.2	306.0	1743.3	1395.1	1162.3	473.8	204.2	181.8	151.3	11.5
AR	86.3	47.7	304.3	266.1	181.0	75.4	30.9	24.5	21.6	14.1
AI	372.2	204.8	767.8	767.8	765.2	346.2	142.6	121.2	121.2	6.3
SG	527.4	307.2	2619.0	1664.1	1119.2	439.4	178.2	131.9	118.4	22.1
GR	678.4	379.1	2084.3	1693.2	1484.0	617.8	232.4	185.3	159.0	13.1
AG	372.5	232.4	3999.2	1128.5	769.8	323.0	132.2	104.8	87.9	45.5
TG	530.6	305.3	1936.9	1589.0	1105.3	444.5	180.7	140.0	126.5	15.3
TI	174.6	104.5	987.2	550.5	365.9	151.4	63.1	48.9	41.9	23.5
VD	393.7	225.7	1826.4	1127.5	807.1	347.9	145.0	112.4	96.5	18.9
VS	224.9	137.0	1394.9	773.4	476.6	189.0	77.3	59.9	51.4	27.1
NE	207.2	121.2	950.4	637.7	419.3	186.9	78.0	60.4	51.8	18.3
GE	341.6	214.5	1444.7	1053.5	782.7	293.5	112.6	87.3	74.9	19.3
JU	109.2	58.0	343.7	293.7	214.0	97.0	38.4	33.0	26.3	13.1

T 18 Caractéristiques des poids finaux `Poidtrin_1854` (=`wght`) pour les données du questionnaire téléphonique par strate d'échantillonnage.

7.2 Pondération de l'enquête écrite

Parmi les individus ayant répondu à l'enquête téléphonique, tous n'ont pas accepté de répondre à l'enquête écrite. L'information collectée pour tous les individus répondant à l'enquête téléphonique a été utilisée pour modéliser la probabilité de réponse de la personne sélectionnée au questionnaire écrit.

Afin de corriger l'effet de la non-réponse, un calage (raking ratio) a été effectué sur les quatre différents types de marges suivantes, trois de ces séries de marges étant les mêmes que pour la pondération de l'enquête téléphonique :

- *Niveau lune* (InfoS) :
  - 48 marges informant sur la modélisation de la non-réponse au niveau individuel (voir section 7.1 et 5.2 ainsi que l'annexe A).
  - 71 marges informant sur la modélisation de la non-réponse au questionnaire écrit (au niveau individuel),définies par les sommes des poids pti_nrqi_k sur les 18760 individus répondants à l'enquête téléphonique (CATI[14]). Elles correspondent aux 71 GHR obtenus par l'arbre de segmentation décrit dans la section 5.3 ainsi que dans l'annexe B. Les valeurs de ces marges sont données dans le tableau 28.
- *Niveau étoile* (InfoU) :
  - 54 marges socio-démographiques définies dans le tableau 12.
  - 18 marges géographiques définies dans le tableau 15.

On note que le raking ratio produisant les poids finaux pour l'enquête écrite se différencie de celui utilisé pour produire les poids finaux de l'enquête téléphonique par l'ajout de 71 marges de calage supplémentaires au niveau de l'échantillon ainsi que par la restriction à un plus petit groupe de répondants. Remarquons également que la somme de chacune de ces 4 séries de marges vaut 6'186'711 et correspond à la taille de la population.
Le tableau 19 résume d'où vient et à quel niveau se trouve l'information auxiliaire utilisée. On doit donc mener les analyses par segmentation comme décrit dans les sections 5.2 et 5.3 afin d'identifier les GHR, i.e. définir les mécanismes de réponse à l'enquête écrite et à l'enquête téléphonique. La modélisation de la non-réponse à l'enquête téléphonique reste inchangée et on reprend donc les sommes des poids initiaux pti_init2_k (répondants + non-répondants à l'enquête CATI/CAPI) pour obtenir les marges de calage au premier niveau lune (voir tableau 27). Le deuxième niveau lune correspond à la modélisation de la non-réponse à l'enquête écrite. On obtient les marges associées en calculant les sommes des poids pti_nrqi_k pour les répondants à l'enquête CATI (= répondants + non-répondants à l'enquête écrite). Contrairement à la manière classique de faire (étapes 6b, 7b et 7c dans la figure 2) on n'utilise pas directement les poids corrigés pour la non-réponse pti_nrqi_k et pti_nrqe_k produits. Cependant on a besoin de la définition des GHR qui mène à leur production. Toute l'information auxiliaire au niveau lune, $\mathbf{X}_k^{\circ}$ ou *InfoS*, est contenue dans les marges (sommes de pti_init2_k et sommes des pti_nrqi_k) ainsi calculées.

Les marges de calage sociodémographiques et géographiques sont les mêmes que celles utilisée dans le calage décrit à la section précédante pour la pondération de l'enquête téléphonique. Leur source est la statistique ESPOP (état déc. 2006) corrigée pour refléter la population des ménages privés (donc sans les personnes en institutions, prisons, etc.). Les marges reflétant le mécanisme de réponse à l'enquête téléphonique sont également identiques dans le cas du calage produisant les poids finaux de l'enquête écrite. S'ajoutent ici les 71 marges reflétant le mécanisme de réponse à l'enquête écrite.
Pour produire ce deuxième jeu de poids finaux, on a aussi défini beaucoup plus de marges de calage que dans les pondérations de l'ESS02 et on a ainsi utilisé plus d'information auxiliaire ce qui tend, si elle est pertinente, à minimiser les biais lors de la production de résultats avec les données de l'ESS07.
Pour des remarques et commentaires sur les poids finaux obtenus se reporter à la section 7.3.

[14] Les proxys étant non-éligibles pour l'enquête écrite.

Type d'information	Vecteur auxiliaire $\mathbf{x}_k$	J	Input d'information **X**
InfoU	54 marges sociodémographiques + 18 marges géographiques	$J^* = 72$	voir les tableaux 12 et 15, totaux de l'ESPOP
InfoS	48 GHR selon le modèle de non-réponse à l'enquête tél., voir la section 5.2	$J^\circ = 119$	voir tableau 27 et annexe A, sommes sur les ménages documentés
	71 GHR selon le modèle de non-réponse à l'enquête écrite, voir la section 5.3		voir tableau 28 et annexe B, sommes sur les répondants CATI

T 19 Définitions et notations pour l'information auxiliaire dans le cas de la pondération de l'enquête écrite pour l'ESS07.

Strate de calage	Moy.	Ecart-type	Max	P99	P95	Méd.	P5	P1	Min	Max/Min
1	586.0	433.1	3956.4	2248.1	1418.2	500.3	153.9	124.5	79.6	49.7
2	436.2	438.6	4554.0	2060.4	1288.5	315.1	40.3	23.5	16.2	280.5
3	351.9	273.6	3766.1	1184.3	791.0	287.3	98.7	73.1	59.4	63.4
4	79.3	46.8	261.0	248.4	183.3	69.0	25.5	16.6	15.8	16.5
5	247.4	179.8	1961.6	947.3	548.9	201.2	76.4	55.8	49.6	39.6
6	266.5	183.7	1997.2	840.3	595.0	212.9	78.4	65.3	47.8	41.8
7	689.0	638.3	8711.4	2658.9	1552.8	566.1	209.9	173.9	160.6	54.2
8	110.3	79.3	762.8	454.8	233.6	89.0	34.2	28.2	22.5	34.0
9	487.6	378.8	4833.9	1804.8	1141.7	397.8	146.2	112.5	98.8	49.0
10	234.5	180.9	1978.6	870.0	581.7	179.9	66.7	50.1	37.8	52.4
11	504.1	355.5	3125.1	1913.8	1168.3	398.2	151.5	114.3	85.6	36.5
12	294.1	243.7	3115.5	1205.3	718.5	224.6	76.4	58.4	44.1	70.6
13	262.8	210.6	2694.1	1040.4	595.2	213.3	78.4	62.0	48.3	55.7
14	484.8	391.0	3217.1	1952.4	1292.0	356.7	125.8	99.3	73.4	43.8
15	140.7	95.1	785.1	496.8	305.5	113.9	40.7	34.0	28.3	27.8
16	672.5	456.4	3968.6	2180.2	1574.6	575.6	214.3	165.7	129.9	30.6
17	707.4	518.3	6706.9	2517.3	1730.9	558.8	202.4	148.8	111.1	60.4
18	687.5	412.0	2941.0	2055.3	1452.6	578.2	213.2	169.2	157.1	18.7

T 20 Caractéristiques des poids finaux `Poidtrie_1854_4871` (= `wecrit`) pour les données du questionnaire écrit par strate de calage.

Strate d'échan-tillon-nage	Moy.	Ecart-type	Max	P99	P95	Méd.	P5	P1	Min	Max/Min
ZH_Ville	431.1	349.4	3956.4	1842.3	1000.0	340.3	137.6	110.0	79.6	49.7
ZH_Reste	686.9	452.3	3386.6	2345.2	1528.0	577.7	215.1	178.9	137.1	24.7
BE_Ober	359.7	237.9	2674.6	1210.9	774.1	308.0	106.5	76.7	61.1	43.8
BE_Jura	86.0	69.1	810.4	321.2	196.9	66.6	25.0	19.3	16.2	49.9
BE_Reste	686.9	493.8	4554.0	2707.7	1547.4	555.6	209.6	167.6	118.3	38.5
LU	351.9	273.6	3766.1	1184.3	791.0	287.3	98.7	73.1	59.4	63.4
UR	79.3	46.8	261.0	248.4	183.3	69.0	25.5	16.6	15.8	16.5
SZ	247.4	179.8	1961.6	947.3	548.9	201.2	76.4	55.8	49.6	39.6
OW	755.7	399.8	2055.3	2055.3	1635.4	604.0	341.6	235.5	235.5	8.7
NW	645.8	393.9	1658.3	1658.3	1423.1	519.9	169.2	157.1	157.1	10.6
GL	516.6	330.0	1372.3	1372.3	1193.5	441.0	153.6	118.1	118.1	11.6
ZG	682.0	424.9	2941.0	2141.4	1452.6	578.2	214.3	192.0	171.3	17.2
FR	266.5	183.7	1997.2	840.3	595.0	212.9	78.4	65.3	47.8	41.8
SO	689.0	638.3	8711.4	2658.9	1552.8	566.1	209.9	173.9	160.6	54.2
BS	665.8	461.9	3245.1	2175.1	1640.2	580.2	214.3	179.4	155.9	20.8
BL	677.3	453.0	3968.6	2180.2	1532.7	564.0	215.6	150.4	129.9	30.6
SH	646.5	416.1	2040.7	2040.7	1721.5	550.4	232.0	211.6	211.6	9.6
AR	110.3	79.3	762.8	454.8	233.6	89.0	34.2	28.2	22.5	34.0
AI	406.6	219.5	942.8	942.8	942.8	392.1	137.6	137.6	137.6	6.9
SG	674.4	534.9	6706.9	2737.0	1468.6	529.6	199.8	145.9	111.1	60.4
GR	911.3	578.7	2866.8	2517.3	2067.4	758.1	252.6	177.3	161.6	17.7
AG	487.6	378.8	4833.9	1804.8	1141.7	397.8	146.2	112.5	98.8	49.0
TG	678.1	455.0	3096.3	2868.1	1474.8	549.9	206.2	148.8	130.0	23.8
TI	234.5	180.9	1978.6	870.0	581.7	179.9	66.7	50.1	37.8	52.4
VD	504.1	355.5	3125.1	1913.8	1168.3	398.2	151.5	114.3	85.6	36.5
VS	294.1	243.7	3115.5	1205.3	718.5	224.6	76.4	58.4	44.1	70.6
NE	262.8	210.6	2694.1	1040.4	595.2	213.3	78.4	62.0	48.3	55.7
GE	484.8	391.0	3217.1	1952.4	1292.0	356.7	125.8	99.3	73.4	43.8
JU	140.7	95.1	785.1	496.8	305.5	113.9	40.7	34.0	28.3	27.8

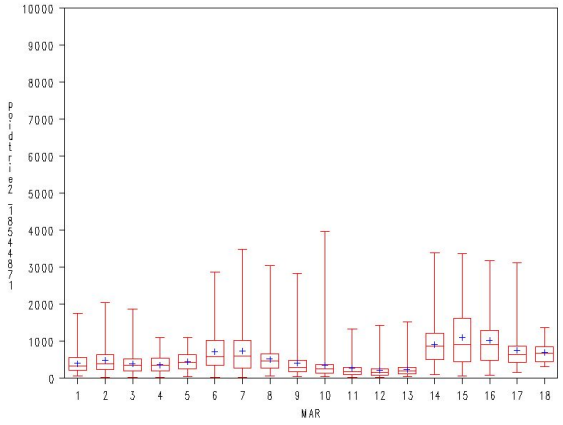
T 21 Caractéristiques des poids finaux `Poidtrie_1854_4871` (= `wecrit`) pour les données du questionnaire écrit par strate d'échantillonnage.

7.3 Commentaires sur les poids finaux obtenus

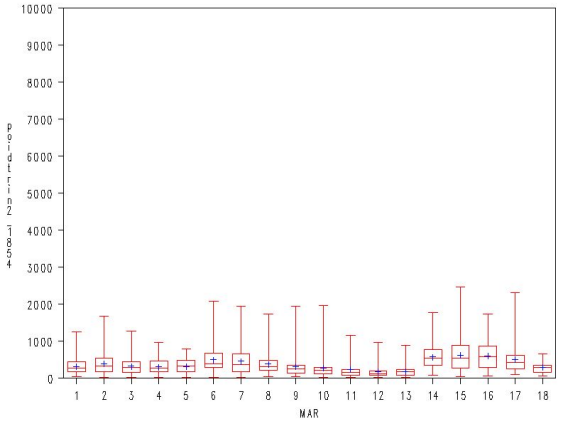
Pour pondérer cette enquête on a opté pour un calage permettant en une seule étape, de corriger pour la non-réponse (selon les modèles décrivant le mécanisme de réponse des individus à différents niveaux) et de respecter des totaux connus de la population (selon l'ESPOP). La manière qu'on a dénommée "conventionnelle" de faire est, comme on l'a exposé dans la figure 2, de d'abord procéder à une modélisation de la non-réponse, puis, dans une étape ultérieure et séparée, effectuer un calage sur de l'information auxiliaire connue au niveau de la population. La manière conventionnelle de faire présente l'inconvénient que le calage sur marges effectué au final détruit partiellement la correction pour la non-réponse effectuée auparavant.
Avec la méthodologie pour laquelle on a opté pour l'ESS07 (calage et correction de la non-réponse par calage en une seule étape), on aurait pu suspecter une plus grande variabilité dans les poids finaux

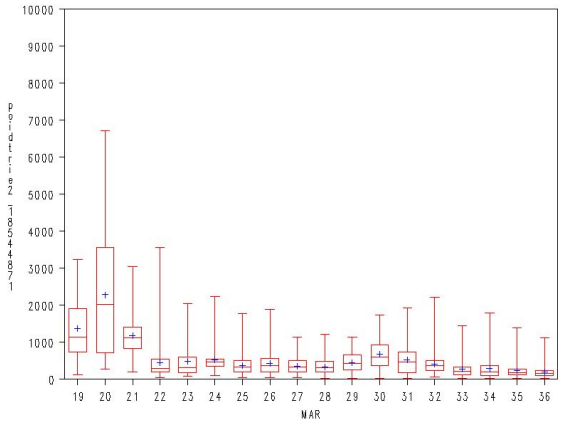
due au nombre plus élevé contraintes (marges de calage) à satisfaire, mais le tableau 16 montre que ça n'a pas vraiment été le cas. En effet l'écart-type et l'étendue des poids finaux *poidtrin*_1854$_k$ (`wght`) et des *poidtrie*_1854_4871$_k$ (`wecrit`) sont un peu plus grands mais restent dans le même ordre de grandeur que ceux des poids *poidtrin*$_k$ et *poidtrie*$_k$ calculés de manière conventionnelle (étapes 6b, 7b et 7c selon la figure 2). Le calage "lune et étoile" a permis une meilleure cohérence entre les poids finaux `wght` et `wecrit` pour les individus ayant participé à l'enquête écrite.
Les tableaux 17 et 18 donnent les caractéristiques de la distribution des poids finaux de l'enquête téléphonique par strate (géographique) de calage, et par strate d'échantillonnage. Les tableaux 20 et 21 présentent les mêmes statistiques pour les poids finaux de l'enquête écrite. On voit que la distribution des poids peut être assez différentes dans ces sous-échantillons.
Le graphique 4 compare la distribution des deux jeux de poids finaux dans les 54 marges (niveau étoile) `MAR` sociodémographiques ainsi que dans les 18 marges (niveau étoile) `STRAT_CALAGE` géographiques reflétant la calibration sur les nombres connus pour la population. Le graphique 5 expose les mêmes comparaisons pour les 48 marges (niveau lune) `RHG_2` reflétant la correction pour la non-réponse à l'enquête téléphonique. On constate que la dispersion par sous-groupe étudié est plus grande pour les poids finaux de l'enquête écrite que pour ceux de l'enquête téléphonique. Ce qui n'est pas surprenant, le calage produisant la pondération pour l'enquête écrite devant respecter 71 marges (2e niveau lune) `RHG_e` supplémentaires. Le graphique 6 présente pour chacun des GHR la dispersion des poids finaux `wecrit`.
Les poids de sondage sont au départ peu dispersés dans une strate géographique (mais le sont quand-même à cause de la taille des ménages). Les ajustements successifs dus à la non-réponse et au calage conduisent à augmenter la dispersion des poids. Cet effet est inégal selon les groupes de population, et l'analyste devra y prêter attention.
Enfin, dans le graphique 7, on compare les valeurs des poids téléphoniques et écrits pour les mêmes observations répondantes à l'enquête écrite. Les poids finaux de l'enquête écrite sont, comme on peut s'y attendre, en moyenne plus élevés et un peu plus dispersés que ceux de l'enquête téléphonique.

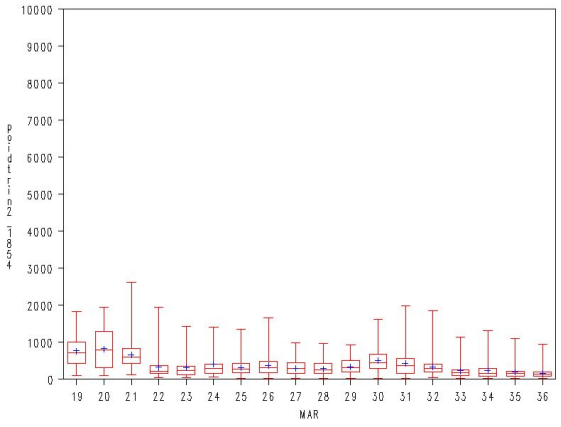
Pondération de l'enquête écrite (`wecrit`).
Poids finaux par MAR.

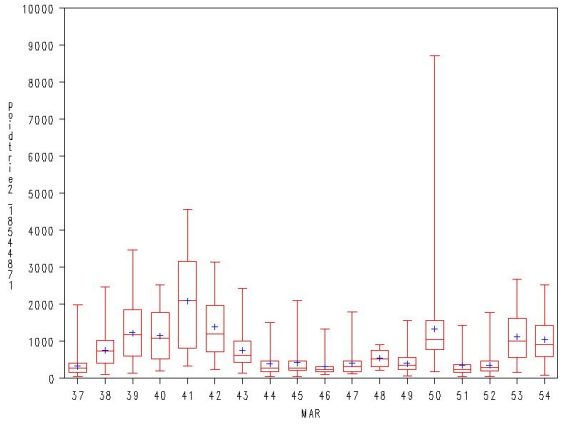


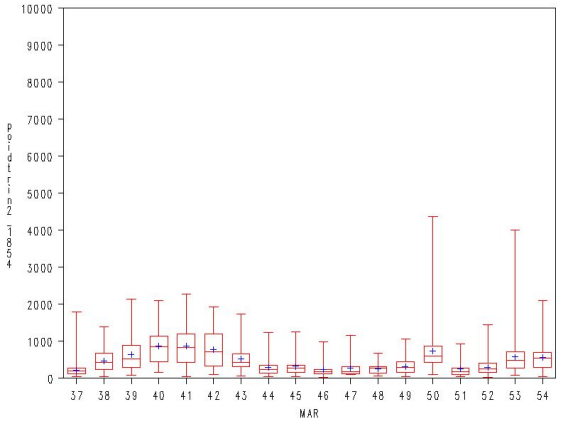
Pondération de l'enquête téléphonique (`wght`).
Poids finaux par MAR.



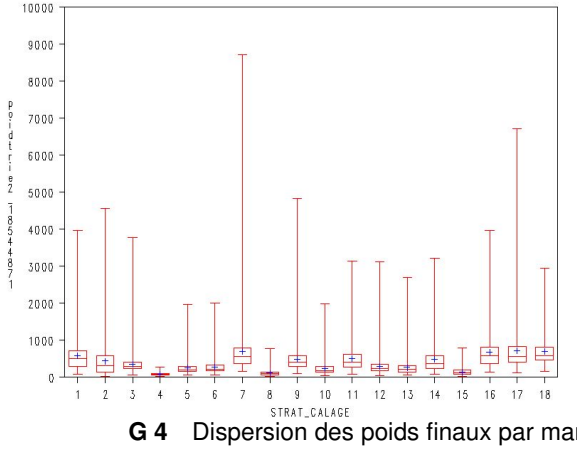




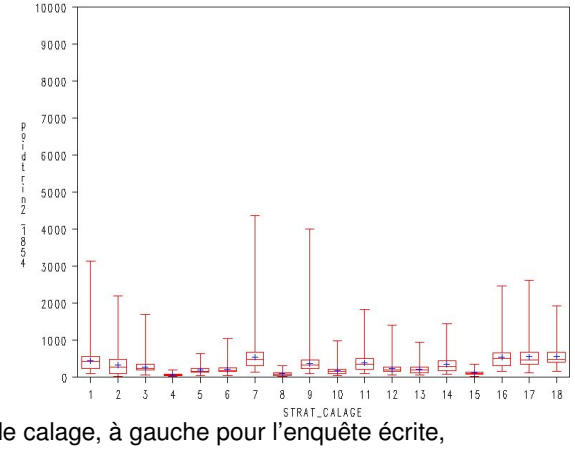




Poids finaux par STRATE_CALAGE.



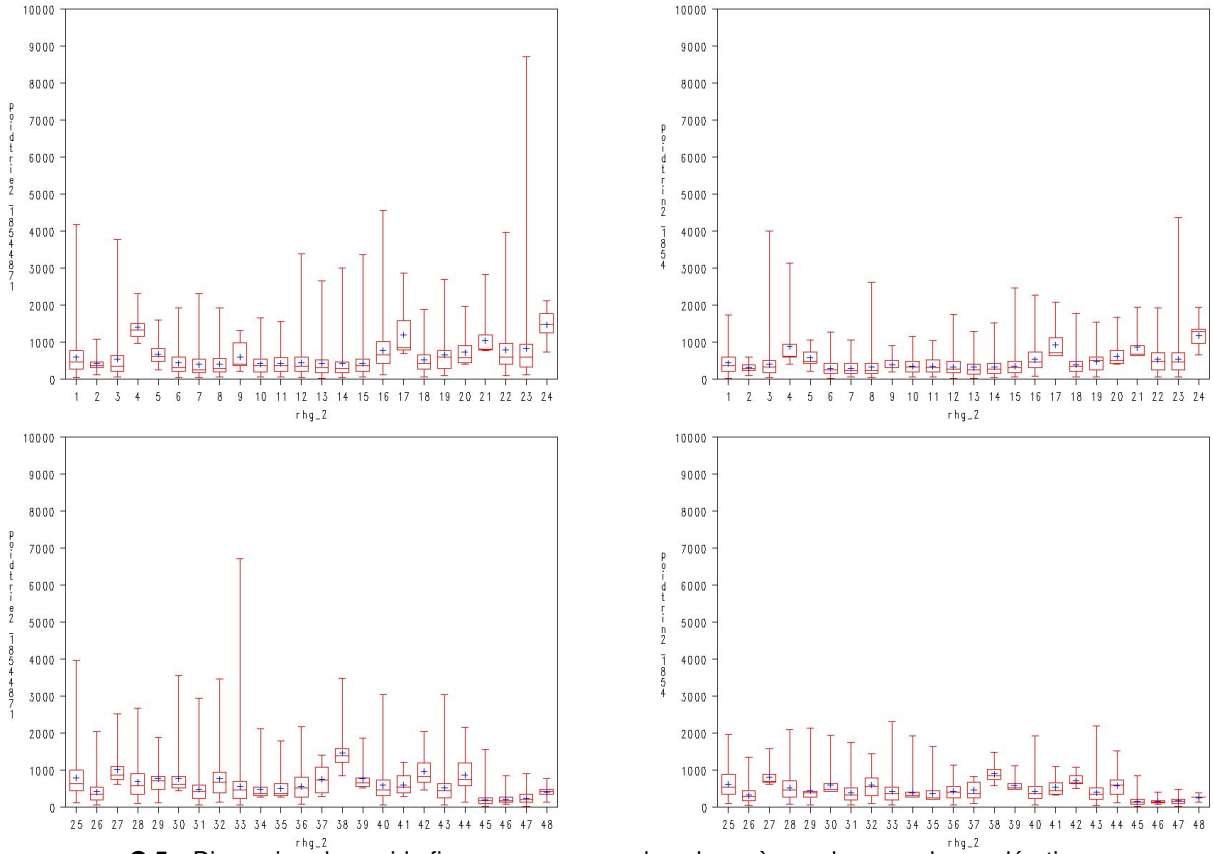
Poids finaux par STRATE_CALAGE.



G 4 Dispersion des poids finaux par marges de calage, à gauche pour l'enquête écrite, à droite pour l'enquête téléphonique. Le tableau 14 définit à quels sous-groupes de population correspondent chacune de ces marges.

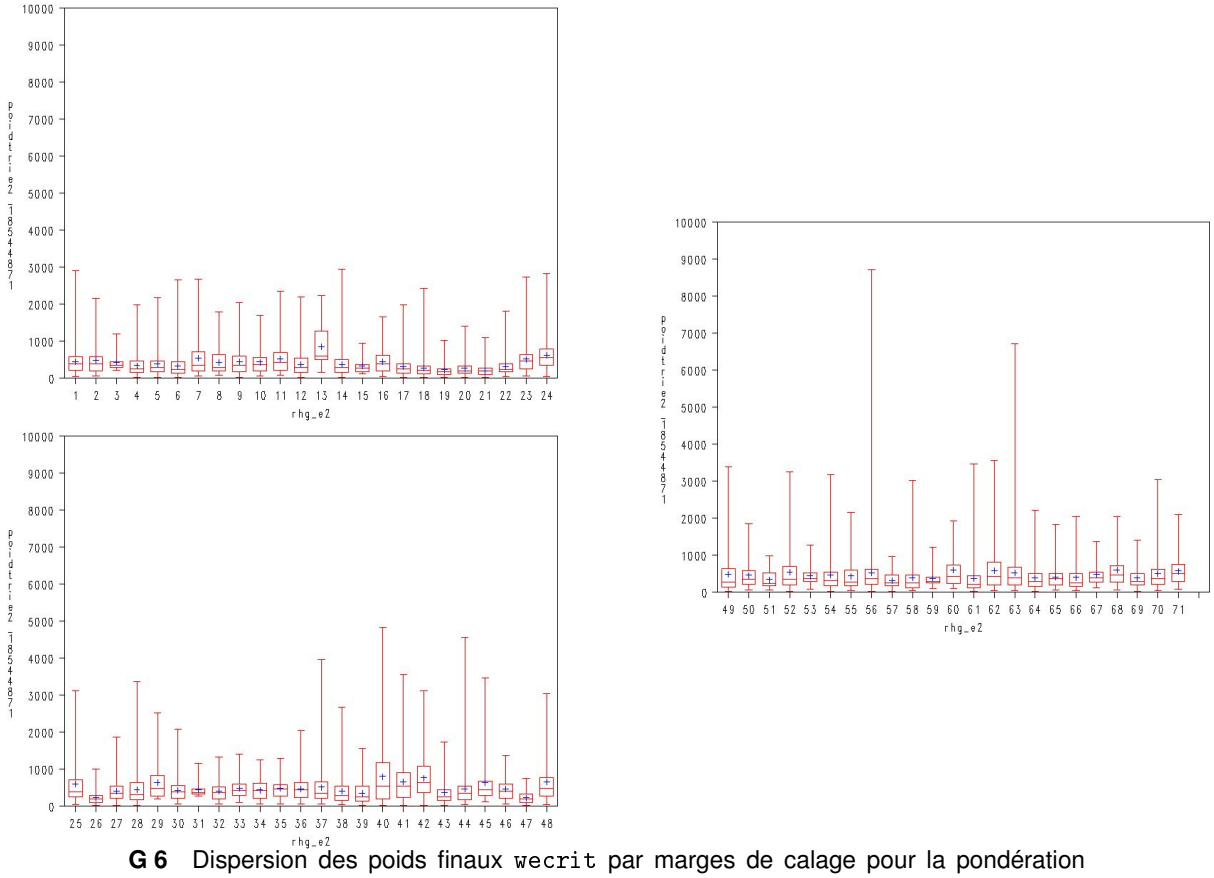
Pondération de l'enquête écrite (`wecrit`).
Poids finaux par RHG_2.

Pondération de l'enquête téléphonique (`wght`).
Poids finaux par RHG_2.

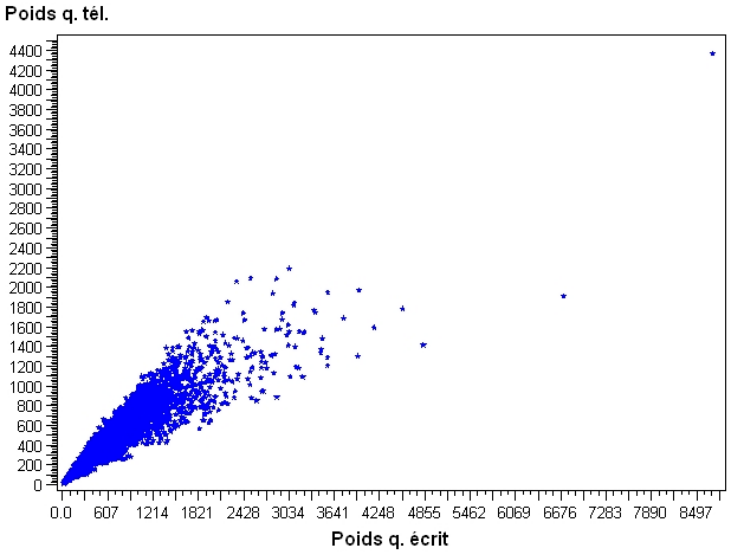


G 5 Dispersion des poids finaux par marges de calage, à gauche, pour la pondération de l'enquête écrite. Poids finaux par RHG_2. A droite, la même chose pour la pondération de l'enquête téléphonique.

Pondération de l'enquête écrite. Poids finaux par RHG_e2.



G 6 Dispersion des poids finaux `wecrit` par marges de calage pour la pondération de l'enquête écrite. Poids finaux par RHG_e2 (ajustement de la non-réponse au questionnaire écrit).



G 7 Poids finaux `wght` du questionnaire téléphonique comparés au poids finaux `wecrit` du questionnaire écrit pour les répondants au questionnaire écrit. On note deux valeurs extrêmes qu'on n'a pas traité (winsorisé par exemple), une pour chacun des deux jeux de poids.

Troisième partie

Analyses des données avec pondération

8 Documentation existante sur les échantillons de 1992 à 2007

Les pondérations des enquêtes suisses sur la santé de 1992 à 2002 ont été documentées dans les rapports suivants :
– Pour l'ESS92 se référer à [3], [13] et [15].
– Pour l'ESS97 se référer à [2] ([14] n'ayant pas été publié).
– Pour l'ESS02 se référer à [2].
Les pondérations de l'ESS07 fait l'objet du présent rapport.

8.1 Définition des strates d'échantillonnage

Le tableau 22 résume les définitions des strates d'échantillonnage utilisées pour les enquêtes suisses sur la santé de 1992 à 2007. Ces strates ont, à quelques détails près toujours été définies de manière analogue : les cantons trop petits en terme de nombre d'unités interviewées ont été regroupés ; les régions ayant financé des interviews supplémentaires ont été traitées comme des strates individuelles. Il est important de connaître les définitions de ces strates lorsqu'on veut calculer la précision de certains estimateurs (voir la section 9). Notons que la définition de ces strates repose sur les informations connues lors de l'échantillonnage. Autrement dit, une personne est ici classifiée dans la strate où son ménage a été sélectionné (selon les données du registre), ce qui ne correspond pas toujours à la strate (région) dans laquelle elle a été interrogée et vivait lors de l'interview (en raison des déménagements entre la sélection dans l'échantillon et l'interview effective). Les strates d'échantillonnage sont donc toujours définies à l'aide de la variable nommée `kanton_1` (= canton au tirage) et non par rapport à la variables `kanton` (= canton lors de l'interview).

8.2 Tailles des cadres de sondages

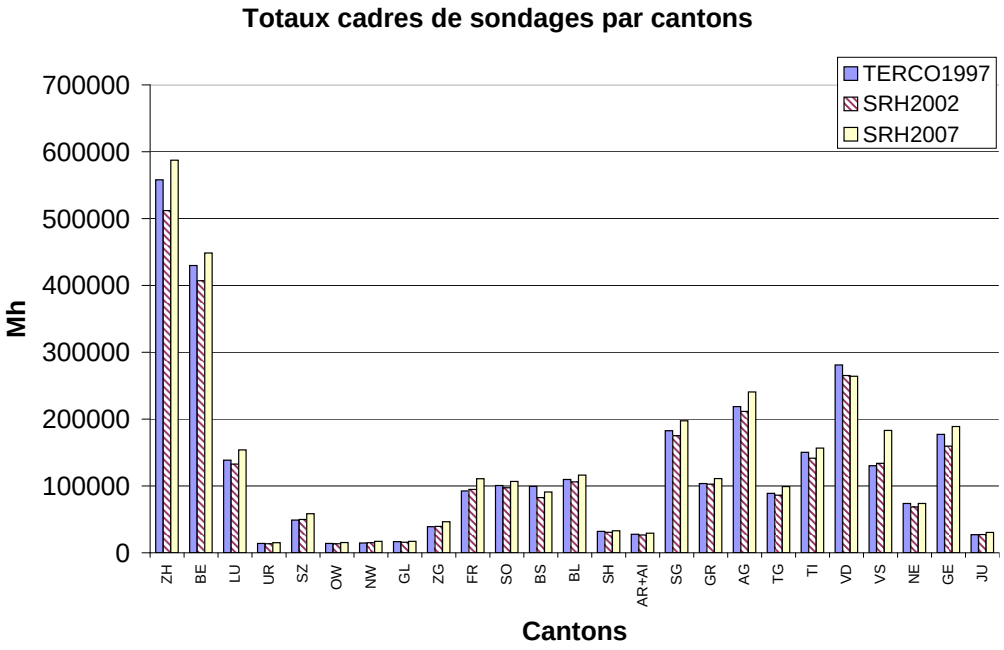
Le tableau 23 et le graphique 8 indiquent les tailles des différents cadres de sondage par strate d'échantillonnage. Cette information est nécessaire pour calculer la précision des estimateurs produits avec les données de l'enquête (voir section 9). Les tailles par strates d'échantillonnage du cadre de sondage de 1992 ne sont plus disponibles. Cependant ces tailles sont nécessaires afin de connaître les taux de sondage par strate d'échantillonnage et de pouvoir calculer des intervalles de confiance pour les estimateurs. Pour des calculs de précisions avec les données de l'ESS92, on pourra prendre les totaux du cadre de 1997.
En 1992 et 1997 les échantillons de l'ESS on été tirés dans le TERCO (annuaire téléphonique des PTT). Les échantillons de 2002 et 2007 sont issus du SRH (Stichproben Register für Haushaltserhebungen) de l'OFS.

ESS2007	ESS2002	ESS1997	ESS1992
AG	AG	AG	AG
AI	AI + AR	AI + AR	AI + AR
AR			
BE jura	BE	BE	BE
BE oberland			
BE reste			
BL	BL	BL	BL + BS
BS	BS	BS	
FR	FR	FR	FR
GE	GE	GE	GE
GL	GL	GL	GL
GR	GR	GR	GR
JU	JU	JU	JU
LU	LU	LU	LU
NE	NE	NE	NE
NW	NW	NW	NW + OW
OW	OW	OW	
SG	SG	SG	SG
SH	SH	SH	SH
SO	SO	SO	SO
SZ	SZ	SZ	SZ
TG	TG	TG	TG
TI	TI	TI	TI
UR	UR	UR	UR
VD	VD	VD	VD
VS	VS	VS	VS
ZG	ZG	ZG	ZG
ZH Stadt	ZH	ZH	ZH
ZH übrige Kt			
29 strates	25 strates	25 strates	23 strates

T 22 Définition des strates pour les plans d'échantillonnages des enquêtes suisse sur la santé de 1992 à 2007. Chaque enquête était formée de quatre échantillons (ou quatre vagues), un par saison, stratifiés comme exposé ici. Dans chacune des strates un échantillon aléatoire simple a été sélectionné.

STRATNUM	TERCO1997	SRH2002	SRH2007
11	557954	511854	183155
12			404284
21	429732	406913	98727
22			22896
23			326835
30	138648	132684	153894
40	13984	13601	15053
50	49049	49956	58376
60	13952	13608	15335
70	14565	15075	17220
80	16749	15820	17228
90	38966	39519	46405
100	92524	94848	110867
110	100608	97235	106838
120	99290	82625	90874
130	109787	106118	116255
140	32159	30628	33017
150	27703	26748	23507
160			5902
170	182562	175181	197457
180	103501	102818	111169
190	218732	211604	240658
200	88874	86269	99099
210	150277	141547	156603
220	281048	265187	264202
230	130226	133872	183159
240	73709	68620	73748
250	177232	159679	188833
260	27183	27284	30406
TOTAL	3169011	3009290	3391994
	25 strates	25 strates	29 strates

T 23 Tailles moyennes (M_h) sur l'année (moyenne sur les quatre trimestres) des cadres de sondage utilisés pour tirer les échantillons des enquêtes ESS de 1997, 2002 et 2007.



G 8 Tailles moyennes (M_h) sur l'année (moyenne sur les quatre trimestres) des cadres de sondage utilisés pour tirer les échantillons des enquêtes ESS de 1997, 2002 et 2007.

9 Comparaisons géographiques et temporelles

Beaucoup d'utilisateurs des données de l'ESS désirent faire des comparaisons cantonales ou régionales des résultats de leurs analyses. On aimerait aussi souvent comparer dans le temps avec les mêmes analyses faites sur les ESS précédentes. Nous abordons ici ce sujet en nous focalisant sur comment il faut utiliser les pondérations dans ces conditions.

9.1 Calculs de précisions

Le fait de vouloir comparer des estimations nous oblige à calculer la précision desdites estimations. Car en effet, il est crucial de savoir s'il s'agit de différences significatives ou pas. On doit donc calculer des grandeurs comme des variances, coefficients de variation, des intervalles de confiance,... . En effet ces mesures fournissent une indication de l'ampleur de la variabilité d'échantillonnage. Cependant, nous rappelons aux utilisateurs que les estimations de ces grandeurs fournies par les procédures classiques de la plupart des logiciels statistiques courants (SAS, SPSS, Stata, MS Excel,...) sont erronées lorsque les données analysées proviennent d'enquêtes comme l'ESS (voir la section 10.6 pour des exemples). Ces logiciels supposent entre autres qu'on a procédé à un échantillonnage aléatoire simple. C'est donc dire qu'ils ne tiennent pas compte des caractéristiques spéciales du plan de sondage de l'ESS comme la stratification, ni des corrections apportées pour la non-réponse, ni de la post-stratification (calage sur marge). Même si les procédures classiques de ces logiciels permettent le recours à la pondération dans la production d'estimations, elles ne prennent pas correctement en compte toutes les spécificités de l'enquête ESS lors du calcul des estimations de la variance, un élément essentiel pour la plupart des tests statistiques (voir section 10.6). Les calculs de ces logiciels ont tendance à sous-estimer ladite variance et par conséquent les intervalles de confiance liés à des estimations.
Malgré le fait que la variance est souvent sous-estimée à l'aide des logiciels cités, nous recommandons de la calculer tout de même en utilisant impérativement la pondération. Effectivement, dans le cas des intervalles de confiance autour d'estimations produites par exemple, ceux-ci auront tendance à être trop étroit. Mais si une différence ou une grandeur s'avère déjà être non-significative avec un tel calcul, il ne sera pas nécessaire d'investiguer davantage.
On a montré (voir [7]) qu'il est possible avec notamment les logiciels SAS et SPSS d'utiliser des procédures permettant de prendre en compte certaines spécificités du plan de sondage. Il s'agit des procédures SURVEY de SAS et des procédures du module COMPLEX SAMPLE de SPSS. Elles calculent la variance selon la méthode de linéarisation en série de Taylor et donnent des résultats très similaires à ceux produit par des méthodes de ré-échantillonnage plus lourdes à implémenter comme le bootstrap ou le jackknife. De ce fait, pour les utilisateurs de SAS ou de SPSS, dans la mesure du possible, nous recommandons d'utiliser ces procédures qui, assurément, fournissent des résultats bien plus corrects que ceux fournis par les procédures standards (même si ceux-ci restent des résultats approchés).
Lorsqu'on veut utiliser ces procédures il est nécessaire de connaître la définition des strates d'échantillonnage (voir section 8.1) ainsi que les tailles du cadre de sondage par strate (voir section 8.2) ou les taux de sondage par strate. Pour des exemples d'utilisation de ces procédures on peut se reporter à [7]. Des calculs de précision par les macros jackknife présentées dans le rapport des pondérations de l'ESS02, voir [2], sont bien sûr aussi une solution satisfaisante.

9.2 Comparaisons régionales

Par comparaisons régionales, nous désignons les comparaisons des mêmes statistiques calculées pour différents cantons ou strates ou d'autres sous-populations définies sur la base l'emplacement géographique des unités interrogées.
Typiquement, on désire comparer certaines statistiques calculées pour les cantons ayant financé une augmentation de leur échantillon net. S'il s'agit de pourcentages ou de totaux il faudra calculer les

intervalles de confiances autour de chaque estimation. Si les intervalles de confiance de deux pourcentages/totaux comparés deux à deux sont disjoints, on pourra conclure que la différence entre les deux pourcentages est significative (pour autant qu'on ait pas trop sous-estimé la variance...). On pourra aussi calculer des tests pour identifier des différences significatives en faisant cependant attention à la manière dont les pondérations sont prises en compte (voir section 10.6) et en se rappelant que la plupart des procédures implémentées dans les logiciels statistiques ont tendance à sous-estimer la variance (voir section 9.1).
D'un point de vue méthodologique, à part la taille des échantillons considérés qui a une influence sur la précision des estimateurs (voir section 10.5), rien n'empêche de comparer des statistiques calculées sur de plus petites zones géographique que le canton ou la strate, ou calculées pour des cantons n'ayant pas financé une augmentation de leur échantillon. Notons cependant que, pour le calcul des pondérations, ces cantons-là ont été regroupés avec les autres cantons de la même grande région[15] n'ayant pas financé d'augmentation de l'échantillon. Les pondérations ne permettent pas de retrouver la bonne population totale isolément pour chacun de ces cantons. On donc peut s'attendre à ce que de plus grands biais entachent les estimations basées sur ces cantons, surtout s'il s'agit de totaux. De plus l'estimation de la variance sera plus grossière si ces cantons autres zones géographiques ne correspondent pas à des strates de calage.

9.3 Comparaisons temporelles

Par comparaisons temporelles, nous comprenons les comparaisons des mêmes statistiques calculées sur les mêmes sous-populations, ou pour toute la population, à l'aide des données des différentes ESS de 1992 à 2007. Il s'agit de calculer la précision d'évolutions.
Lorsqu'on veut comparer deux statistiques calculées à partir de deux ESS différentes, on considère que les ESS constituent des échantillons indépendants de la population Suisse.
Sous l'hypothèse que les différents échantillons ESS sont indépendants, on pourra calculer des tests, intervalles de confiances, etc. pour les statistiques d'intérêt comme pour des comparaisons régionales.

[15] Selon la nomenclature NUTSII.

10 Questions fréquentes, conseils pour les utilisateurs

10.1 Faut-il utiliser les poids ?

La réponse est oui.
Les données des fichiers livrés ne sont pas pondérées. Les poids sont livrés dans des variables séparées. Il est du ressort des utilisateurs de données d'appliquer les poids appropriés compte tenu des analyses qu'ils veulent produire. On risque de forts biais si l'on n'applique pas les poids.

10.2 Quelle pondération utiliser : téléphonique/écrite ?

Si la ou les variable(s) étudiées sont toutes issues de l'interview CATI/CAPI[16], alors il faut utiliser la pondération de l'enquête téléphonique (variable `wght` dans les données livrées). Dès qu'une variable ou plus de l'analyse provient du questionnaire écrit[17], le simple fait de l'utiliser force l'utilisateur à se restreindre au sous-échantillon des répondants à l'enquête écrite et la bonne variable de poids à utiliser est celle de la pondération de l'enquête écrite (variable `wecrit` dans les données livrées). Il faudra faire particulièrement attention à la cohérence entre les tableaux calculés.

10.3 Quelle est la différence entre la strate et le canton ?

La section 8.1 répond entièrement à cette question. En effet, de l'ESS92 à l'ESS07, certains cantons ont parfois aussi été une strate, parfois un agglomérat de strates, parfois une partie de strate. La strate repose sur la définition du canton lors du tirage de l'échantillon (comme la variable `kanton_1`, non-livrée), mais lors de l'analyse on préfère souvent utiliser le canton où le ménage habitait effectivement lors de l'interview (variable `kanton`).

10.4 Faut-il tenir compte de la poststratification dans l'analyse ?

On en tient de toute façon compte dès lors que l'on applique les poids pour ce qui est de la valeur des estimations.
Pour ce qui est du calcul de la précision des mêmes estimations, il faut également en tenir compte si l'on veut effectuer des calculs de variances plus corrects que ceux fournis de manière standard par les procédures des logiciels statistiques (voir sections 9 et 10.6). Cependant cela devient vite plus difficile.

10.5 Les poids peuvent-ils être utilisés tels quels sur un sous-groupe de l'échantillon ?

Oui, et ils doivent être utilisés. En particulier, les poids sont valables pour les cantons ayant sur-échantillonné pris séparément, pour autant qu'il y ait assez d'observations.
De façon générale, plus l'échantillon est petit, plus la variabilité d'échantillonnage est grande. Par conséquent les estimations concernant de petits sous-groupes de population sont tendanciellement moins fiables que celles relatives à des (sous-)groupes importants.
Si une estimation repose sur moins d'une trentaine d'observations (nombre non pondéré d'observations), celle-ci ne devrait être diffusée qu'avec réserves, et une attention particulière devrait être prêtée à la valeur de son coefficient de variation. En effet, plus le groupe d'observations étudié est petit, plus le risque de biais élevé est grand.

[16] Les variables issues du questionnaire téléphonique ont toutes leur nom commençant par "t". Exemple : **t**gezu02b.
[17] Les variables issues du questionnaire écrit ont toutes leur nom commençant par "s". Exemple : **s**inan24.

10.6 Dans quelles conditions doit-on standardiser les poids ?

Par standardiser les poids, on entend créer un nouveau jeu de poids où l'on divise les poids population (w_i^{pop}) de chaque observation par le poids moyen. Ainsi la somme des poids ne reflète plus la taille de la population Suisse (dans le cas de l'ESS) mais plutôt la taille de l'échantillon s. On parle alors de poids échantillon (w_i^{std}) ou poids standardisés :

$$w_k^{std} = \frac{w_k^{pop}}{\overline{w}_k^{pop}},\ \forall k \in s \text{ où } \overline{w}_k^{pop} = \frac{\sum\limits_{k \in s} w_k^{pop}}{\sum\limits_{k \in s} \mathbf{1}_k}$$

Dans le cas de régressions logistique, tests du chi carré, etc., en principe les logiciels statistique courant permettent d'utiliser les poids pour calculer les paramètres, mais souvent ils ne les appliquent pas correctement pour en calculer la variance (ou les autres mesures de précision : écart-type, coefficient de variation, ods ratio, ...). Dans ces cas là la somme des poids est faussement interprétée comme étant égale au nombre d'observations dans l'échantillon, ce qui fausse les résultats (décisions lors de tests). Dans la suite nous illustrons ceci par trois exemples simples.
On utilise les poids standardisés dans les procédures classiques pour obtenir des statistiques de test moins erronées qu'avec les poids population. Ces standardisations doivent toutefois être considérées avec prudence car elles ne sont que des approximations très grossières.

10.6.1 Régression logistique

Prenons comme premier exemple la `proc logistic` de SAS. Soit le modèle suivant calculé sans poids :

```
proc logistic data= reglog;
    title "modele sans poids";
    model fumeur= AC1--AC7 homme KIDinHH TGEZU02b ALLEM;
run;
```

où :

fumeur	=1 si ttako01=1, 0 sinon
AC1	=1 si alter < 20, 0 sinon
AC2	=1 si 20 <= alter < 25, 0 sinon
AC3	=1 si 25 <= alter < 35, 0 sinon
AC4	=1 si 35 <= alter < 45, 0 sinon
AC5	=1 si 45 <= alter < 55, 0 sinon
AC6	=1 si 55 <= alter < 65, 0 sinon
AC7	=1 si 65 <= alter < 75, 0 sinon
homme	=1 si homme, 0 sinon
KIDinHH	=1 si NBKIDS>0, 0 sinon
TGEZU02b	poids en kg
ALLEM	=1 si KSPRACHE="D", 0 sinon

Le jeu de donnée `reglog` correspond aux données de l'enquête téléphonique de l'ESS07, restreintes aux valeurs non-manquantes pour `TGEZU02b` et `ttako01`. Il reste 18555 observations.
On calcule aussi le modèle avec les poids population :

```
proc logistic data= reglog;
    title "modele avec poids pop";
```

```
    model fumeur= AC1--AC7 homme KIDinHH TGEZU02b ALLEM;
    weight wght;
run;
```

...et le modèle avec les poids échantillon (standardisés) :

```
proc logistic data= reglog;
    title "modele avec poids pop";
    model fumeur= AC1--AC7 homme KIDinHH TGEZU02b ALLEM;
    weight wght_std;
run;
```

Les résultats sont présentés dans le tableau 24. On essaye d'expliquer le fait d'être fumeur ou pas. La variable ALLEM ne devrait pas être significative, le fait de parler allemand ou pas n'étant absolument pas corrélé avec le fait de fumer. Pour le modèle calculé sans poids, la variable n'est pas significative, tandis qu'elle le devient si l'on recalcule le modèle avec les poids population. Elle est à nouveau non-significative lorsqu'on applique les poids standardisés. Les estimations des paramètres ne changent pas que l'on utilise les poids population ou les poids échantillon (standardisés), par contre les calculs de précisions y sont très sensibles en raison de la mauvaise interprétation de la somme des poids par la procédure (comme expliqué plus haut).
L'utilisation de la `proc surveylogistic` nous économise la standardisation des poids et permet des calculs plus corrects de la variance. En effet, on peut alors renseigner la procédure avec les informations relatives au plan de sondage stratifié de l'ESS (strates et tailles du cadre par strate d'échantillonnage) et en tenir compte dans l'estimation de ladite variance (voir section 9.1). Le code utilisé est :

```
proc logistic data=reglog N=TaillesCadreSondage;
    title "modele avec poids pop proc survey";
    strata STRATNUM;
    model fumeur= AC1--AC7 homme KIDinHH TGEZU02b ALLEM;
    weight wght;
run;
```

Les fichier SAS `TaillesCadreSondage` contient les tailles du cadre de sondage par strate pour l'ESS07 données dans le tableau 23. La variable `STRATNUM` identifie les strates d'échantillonnage, elle doit figurer dans le jeu de données (ici `reglog`) et avec le même nom dans le fichier `TaillesCadreSondage`. La colonne du fichier `TaillesCadreSondage` contenant les tailles doit être nommée `_TOTAL_`.
On constate que les erreurs standard obtenues avec la `proc surveylogistic` sont supérieures à toutes leurs autres estimations décrites plus haut. La statistique du Chi2 change elle aussi mais pas au point de changer la liste des variables significatives à retenir dans le modèle (dans cet exemple).

Paramètre	Estimation		Erreur standard				Chi2				Pr>Chi2			
poids	-	wght & wght std & wght survey	-	wght	wght std	wght sur-vey	-	wght	wght std	wght sur-vey	-	wght	wght std	wght sur-vey
Intercept	1.87	2.17	0.12	0.01	0.13	0.16	225.75	86114.91	261.45	193.39	<.0001	<.0001	<.0001	<.0001
AC1	-1.47	-1.48	0.11	0.01	0.11	0.14	174.73	56785.18	172.4	118.98	<.0001	<.0001	<.0001	<.0001
AC2	-2.14	-2.12	0.11	0.01	0.11	0.13	396.16	118516.83	359.83	249.03	<.0001	<.0001	<.0001	<.0001
AC3	-1.91	-1.98	0.09	0.01	0.1	0.11	423.9	119139.63	361.72	323.41	<.0001	<.0001	<.0001	<.0001
AC4	-1.77	-1.73	0.09	0.01	0.1	0.11	369.6	89816.82	272.69	242.88	<.0001	<.0001	<.0001	<.0001
AC5	-1.75	-1.71	0.09	0.01	0.1	0.11	371.47	89962.4	273.13	240.15	<.0001	<.0001	<.0001	<.0001
AC6	-1.32	-1.34	0.09	0.01	0.11	0.11	210.39	53582.56	162.68	150.18	<.0001	<.0001	<.0001	<.0001
AC7	-0.84	-0.83	0.1	0.01	0.11	0.12	74.2	17481.09	53.07	50.64	<.0001	<.0001	<.0001	<.0001
homme	-0.51	-0.5	0.04	0	0.04	0.05	155.16	50269.66	152.62	90.39	<.0001	<.0001	<.0001	<.0001
KIDinHH	0.36	0.31	0.04	0	0.04	0.06	65.43	17579.12	53.37	29.11	<.0001	<.0001	<.0001	<.0001
TGEZU02B	0.01	0.01	0	0	0	0.00	46.02	7487.53	22.73	13.63	<.0001	<.0001	<.0001	0.0002
ALLEM	0.05	0.02	0.03	0	0.04	0.04	2.18	69.35	0.21	0.15	0.14	<.0001	0.65	0.7009

T 24 Comparaison des résultats de régressions logistiques avec et sans poids avec la procédure `proc logistic` de SAS pour les colonnees "-", "wght" et "wght std", et la `proc surveylogistic` pour les colonnes "wght survey".

10.6.2 Test binomial

Voyons un autre exemple où l'on voudrait tester par un test binomial si la proportion de femmes dans la population est significativement différente de 51.5%. On utilise la `proc freq` de SAS, d'abord sans les poids :

```
proc freq data = tests;
    title "sans poids";
    tables homme / binomial(p=.515);
run;
```

...ensuite avec les poids population :

```
proc freq data = tests;
    title "poids pop";
    tables homme / binomial(p=.515);
    weight wght;
run;
```

...puis avec les poids échantillon :

```
proc freq data = tests;
    title "poids ech";
    tables homme / binomial(p=.515);
    weight wght\_std;
run;
```

Les résultats sont présentés dans le tableau 25. Sans les poids on obtient une proportion de 55.10% et l'hypothèse H0 est acceptée (différence significative). Mais ce résultat est potentiellement biaisé car on n'a pas utilisé les poids. En effet, avec les poids population on s'aperçoit que la proportion estimée de femmes dans la population est de 51.15% ce qui correspond à la "vérité" vu que l'on a utilisé le sexe dans le calage. L'hypothèse H0 est aussi acceptée. Cependant la somme des poids est faussement interprétée comme correspondant à la taille de l'échantillon. On sous-estime très probablement la variance. On peut donc émettre des doutes quant à la validité du résultat. En utilisant les poids échantillon, l'hypothèse H0 est rejetée : la différence entre 51.15% (estimation) et 51.5% (H0) n'est pas significative sur la base des données (de l'échantillon) à disposition. Ce résultat est le plus acceptable et intuitif des trois.

10.6.3 Test du Chi-carré

Voyons un dernier exemple où l'on voudrait tester si les proportions observée pour la variable `fumeur`, considérée dans ce cas comme étant catégorielle, diffèrent significativement de proportions données. On utilise la `proc freq` de SAS, d'abord sans les poids, les proportions testées sont 72% de non-fumeurs et 28% de fumeurs :

```
proc freq data = reglog;
    title "sans poids";
    tables fumeur / chisq testp=(72 28);
run;
```

...ensuite avec les poids population :

poids	-	wght	wght_std
Proportion	0.5510	0.5115	0.5115
ASE	0.0036	0.0002	0.0036
95% lower conf limit	0.5438	0.5111	0.5044
95% upper conf limit	0.5581	0.5119	0.5187
Test of H0 : Prop =	0.515	0.515	0.515
ASE under H0	0.0036	0.0002	0.0036
Z	9.855	-17.2096	-0.9477
One-sided Pr > Z	<.0001	<.0001	0.1716
Two-sided Pr > |Z|	<.0001	<.0001	0.3433

T 25 Test de proportions binomiaux pour `homme=0` avec la procédure `proc freq` de SAS. Notons que bien qu'on utilise les poids dans les calculs relatés dans les deux dernières colonnes du tableau, les calculs de variances sous-jacents ne tiennent pas compte du plan de sondage ni des spécificités de l'enquête (correction de la non-réponse et calage).

```
proc freq data = reglog;
    title "poids pop";
    tables fumeur / chisq testp=(72 28);
    weight wght;
run;
```

...puis avec les poids échantillon :

```
proc freq data = reglog;
    title "poids ech";
    tables fumeur / chisq testp=(72 28);
    weight wght_std;
run;
```

Les résultats sont résumés dans le tableau 26. Sans poids, l'hypothèse H0 est rejetée, la différence serait significativement différente. Ce résultat est surprenant car les pourcentages obtenus sont proches de ceux de l'hypothèse nulle. Avec les poids population, on continue de rejeter l'hypothèse nulle, pourtant on obtient des proportions de 72.06% et 27.94% qui sont encore plus proches de celles testées. Il semble clair que la variance est sous-estimée dans les deux cas. Dans le premier cas (sans poids) on a probablement en plus un problème de biais. Avec les poids échantillon on accepte enfin H0, la différence entre 72.06% / 27.94% et 72% / 28% est non-significative. Ce résultat est en accord avec notre intuition.

10.6.4 Conclusions

Dès que des calculs de variances sont en jeu il s'agit d'être prudent sur la manière dont les poids seront traités par les procédures utilisées. A moins que l'on veuille estimer des totaux, il est en principe sage de standardiser les poids.
Par ailleurs, précisons que tous les calculs de variance présentés dans ce paragraphe ne tiennent pas compte du plan d'échantillonnage ni des calage sur marges et corrections de non-réponse effectués. Il existe des procédures plus complexes permettant de mieux tenir compte des spécificités de l'enquête (`proc survey` de SAS, module complex samples de SPSS, etc., voir la section 9 ainsi que la section 10.4), cependant leur utilisation est un peu moins aisée et elles offrent moins d'options.

non-fumeurs	13504 (72.78%)	4403820 (72.06%)	13370.31 (72.06%)
fumeurs	5051 (27.22%)	1707695 (27.94%)	5184.685 (27.94%)
Chi2	5.5742	10.1087	0.0307
DF	1	1	1
Pr > ChiSq	0.0182	0.0015	0.8609
Sample Size	18555	6111514.7823	18555

T 26 Test du Chi carré pour des proportion d'une variable catégorielle, H0 : 72% de non-fumeurs et 28% de fumeurs, avec la procédure `proc freq` de SAS. Notons que bien qu'on utilise les poids dans les calculs pour les deux dernières colonnes du tableau, les calculs de variances sous-jacents ne tiennent pas compte du plan de sondage ni des spécificités de l'enquête (correction de la non-réponse et calage).

10.7 La somme des poids ne correspond pas aux nombres publiés dans l'ESPOP

Les marges de calages géographiques (tableau 15) et sociodémographiques (tableau 12) utilisées dans les étapes finales de la production des pondérations correspondent aux totaux pour la population issus d'une version modifiée la statistique ESPOP, état de décembre 2006.
En effet, la population cible de l'enquête suisse sur la santé sont les personnes résidant dans des ménages privés (voir section 1.2). Or la statistique ESPOP standard dénombre toute la population Suisse, y compris les personnes résidantes en homes, institutions, internats, prisons, ménages collectifs, etc.. C'est pourquoi la section EKL de l'OFS produit une version corrigée pour que les totaux (nombres de personnes) correspondent au mieux à la population résidant dans des ménages *privés*. Cette ESPOP corrigée est plus adaptée car elle correspond à la population cible de l'ESS.
Les totaux utilisés dans le calage peuvent donc légèrement différer par rapport aux nombres publiés par l'ESPOP standard de la même époque. Les différences sont un peu plus marquées pour les jeunes et pour les personnes âgées.
D'autre part, la pondération se base sur la variable `KANTON_1`, i.e. la définition du canton lors de l'échantillonnage (variable non-livrée), et non sur la variable `KANTON` (canton lors de l'interview, variable livrée). Les totaux sont calibrés avec `KANTON_1` comme définition du canton. Donc, quand on somme les poids par rapport à `KANTON`, on observe de légères différences en raison des quelques déménagements entre le moment du tirage de l'échantillon et celui de l'interview effective. Ces différences reflètent une évolution de la population entre l'ESPOP06 et l'ESS07.

Conclusions

L'objectif principal de l'ESS07 est de fournir, à partir des données récoltées sur un échantillon aléatoire stratifié de personnes résidant en Suisse et âgées de 15 ans et plus, des résultats sur l'état de santé de la population et sur les comportements en matière de santé ou le recours aux prestations du système de santé.
Le but de ce rapport est de décrire les travaux méthodologiques conduisant aux résultats définitifs de l'ESS07. Que ce soit dans les estimations ponctuelles, les estimations de variance ou les analyses, les poids et le plan d'échantillonnage doivent être pris en compte.
Ce document est divisé en trois parties :
Partie I : Le première section de ce rapport est consacré au rappel de la délimitation de la population, du cadre de sondage et de l'unité d'échantillonnage. La section 2 détaille le plan d'échantillonnage et les taux de réponse obtenus.
Partie II : Pour obtenir des résultats extrapolés à la population, une pondération différente a été élaborée selon le type de questionnaire (téléphonique ou écrit). Les paragraphes de cette partie détaillent toutes les étapes de la construction des poids ainsi que les fondements théoriques de la méthodologie utilisée. La section 4 expose le calcul des poids initiaux, la section 5 traite de la modélisation de la non-réponse. Enfin les sections 6 et 7 sont consacrées à la théorie et à l'application pratique dans le cadre de l'ESS des calage sur marges utilisés pour produire les poids finaux.
Partie III : cette partie aborde de manière non-exhaustive quelques questions et problèmes pratiques courants rencontrés lors de l'utilisation des données de l'ESS dans des analyses. On insiste aussi sur le fait qu'il est impératif de toujours utiliser les pondérations et on justifie cette recommandation par des exemples. On essaye aussi dans la section 10 de répondre de manière brève à certaines questions récurrentes quant à l'utilisation des pondérations.
L'estimation de la précision des estimateurs basés sur cette enquête est aussi brièvement abordée. On conseille dans la mesure du possible, d'utiliser des méthodes de ré-échantillonnage (par exemple jackknife (voir [2]) ou l'utilisation des procédures "survey" de SAS ou module "complex samples" de SPSS qui tous deux se basent sur des linéarisation en série de Taylor de la variance. Dans le cas de l'utilisation de procédures qui ne permettent pas de prendre en compte les spécificités de l'enquête (plan de sondage, correction de la non-réponse, calage) on recommande la plus grande prudence, et de bien étudier la manière dont ces procédures tiennent compte des poids pour les calculs de précision. L'estimation de la précision des résultats obtenus dans cette enquête est particulièrement importante pour établir la fiabilité des résultats et poser ainsi les limites d'exploitation de cette enquête.

Références

[1] Deville, J.-C. and Särndal, C.-E. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418) :376–382, 1992.

[2] Graf, E., Renfer, J.-P. *Enquête suisse sur la santé 2002 : plan d'échantillonnage, pondération et estimation de la précision*. OFS, Neuchâtel, 2005.

[3] Hulliger, B. & Peters, R. *La technique de pondération des données : application à l'enquête suisse sur la santé*. OFS, Berne, 1994.

[4] Kalton, G. et Kasprzyk, D. Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12 :1–17, 1986.

[5] Kass, G.V. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29 :119–127, 1980.

[6] Nakache, J.-P. et Confais, J. *Statistique explicative appliquée : analyse discriminante, modèle logistique, segmentation par arbre*. Editions Technip, 2003.

[7] Plaza, S., Graf, E. *Recommandations et exemples pratiques concernant l'application des pondérations. Panel Suisse de Ménages.* Panel Suisse de Ménages, 2007.

[8] Sautory, O. La macro calmar : redressement d'un échantillon par calage sur marges. Technical report, INSEE, 1993.

[9] Section de la santé. Conception générale de l'enquête suisse sur la santé 1992/93. Technical report, OFS, Berne, 1991.

[10] Särndal, C. E., Sixten Lundström. *Estimation in Survey with Nonresponse*. Wiley, 2005.

[11] Särndal, C. E., Swensson, B. and Wretman, J. *Model Assisted Survey Sampling, Springer Series in Statistics*. Springer, 1992.

[12] Tambay, J. L., Schiopu-Kratina, I., Mayda, J., Stukel, D. et Nadon, S. Traitement de la non-réponse du cycle de deux enquêtes sur la santé de la population. *Techniques d'enquête*, 24 :159–169, 1998.

[13] Vonlanthen, C. *Pondération de l'enquête suisse sur la santé 1992/93*. OFS, Berne, 1994.

[14] Vonlanthen, C. Méthodes statistiques de l'enquête suisse sur la santé. Non paru, 1997.

[15] Vonlanthen, C. *Méthodologie statistique de l'enquête suisse sur la santé 1992/93*. OFS, Berne, 1997.

Annexes

A Modélisation de la non-réponse à l'enquête CATI/CAPI

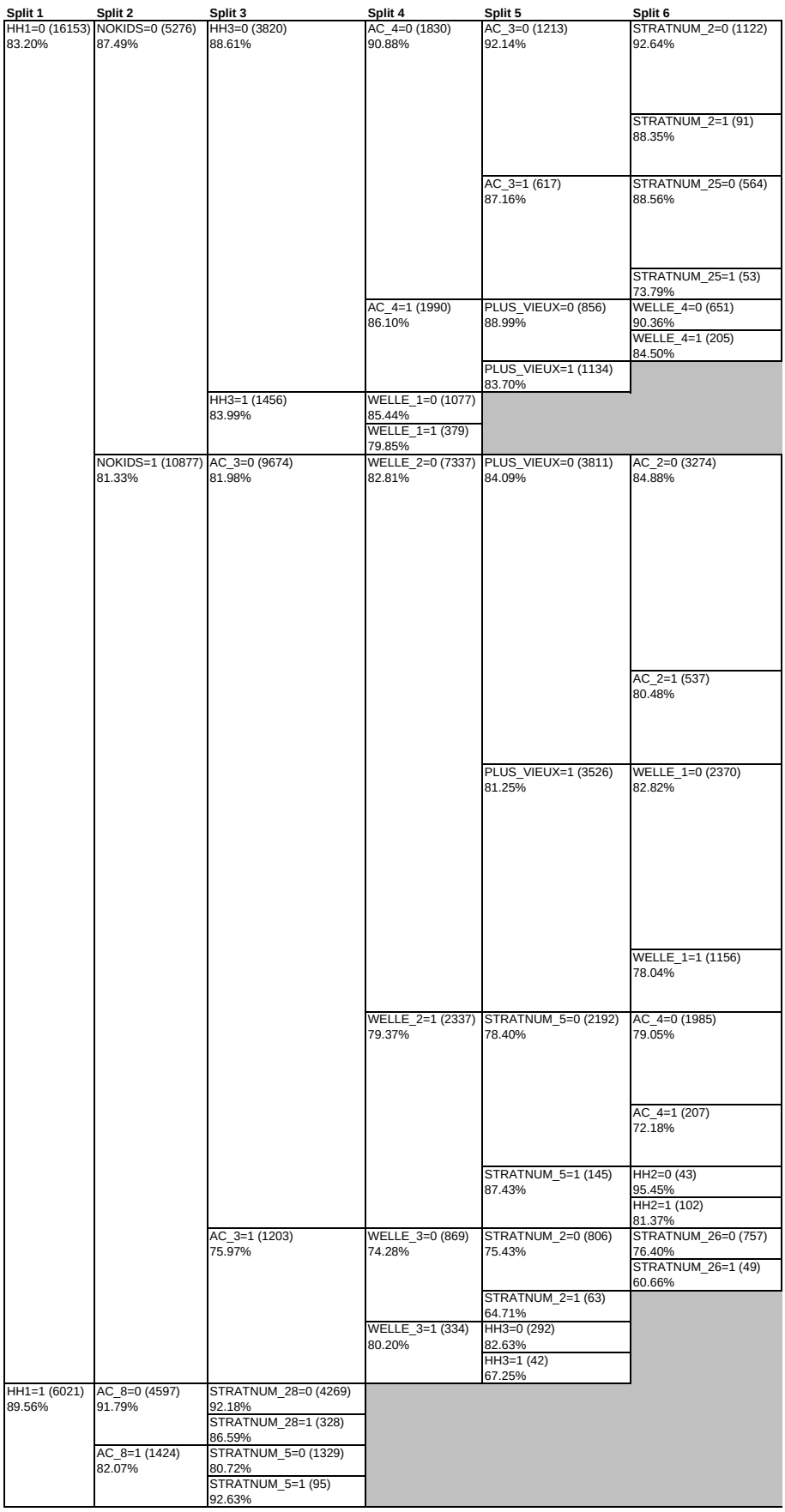
Parmi les ménages ayant été recensés lors du screening téléphonique, toutes les personnes sélectionnées n'ont pas accepté de répondre à l'enquête téléphonique. L'information collectée au niveau de tous les ménages a été utilisée pour modéliser a posteriori la probabilité de réponse de la personne sélectionnée.

L'arbre (voir les figures 9 et 10) modélisant la non-réponse a été obtenu par une méthode de segmentation. Dans chacune des cases de l'arbre on trouve :
– la valeur correspondante de la variable dichotomique la définissant,
– entre parenthèses, le nombre de répondants correspondant à la case concernée,
– le taux de réponse des répondants ayant les caractéristiques de la case.
Les critères minimaux devant être satisfaits pour chaque case de l'arbre de segmentation ont été fixés dans ce cas précis à 50% de taux de réponse et 30 observations.
Les "feuilles" de l'arbre, i.e. les cases finales de chaque ligne du tableau, définissent les groupes homogènes de réponse (GHR). On fait l'hypothèse que la non-réponse est uniforme dans chaque GHR. Pour chacun de ces GHR on calcule alors la somme des poids initiaux des individus (répondants ET non-répondants) y correspondant. Ces sommes définissent nos marges de correction pour la non-réponse. Les poids calés (=poids finaux après correction de la non-réponse) des répondants sommeront donc au même total que celui auquel sommaient les poids initiaux des répondants et des non-répondants. Les marges sont données dans le tableau 27.

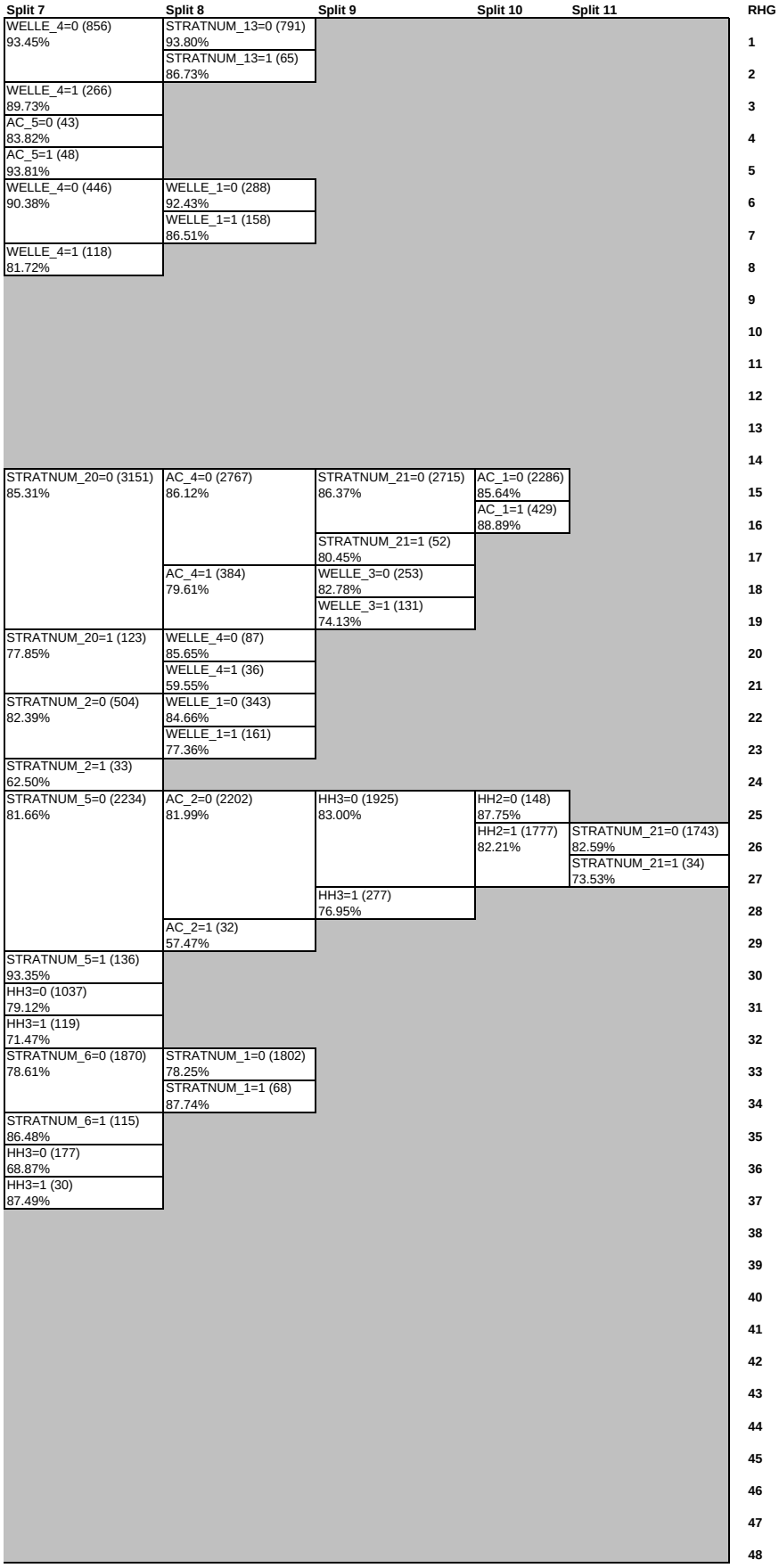
RHG_1 316842	RHG_2 16559	RHG_3 91653	RHG_4 30782	RHG_5 25576	RHG_6 76560	RHG_7 40693	RHG_8 31127
RHG_9 15527	RHG_10 202104	RHG_11 62013	RHG_12 318357	RHG_13 290213	RHG_14 101726	RHG_15 691118	RHG_16 199993
RHG_17 39517	RHG_18 83396	RHG_19 48099	RHG_20 46237	RHG_21 19689	RHG_22 152143	RHG_23 68666	RHG_24 23539
RHG_25 79726	RHG_26 460531	RHG_27 20204	RHG_28 111132	RHG_29 9299	RHG_30 75342	RHG_31 318990	RHG_32 52454
RHG_33 587380	RHG_34 23206	RHG_35 36158	RHG_36 55319	RHG_37 11957	RHG_38 36717	RHG_39 48638	RHG_40 239200
RHG_41 15621	RHG_42 30782	RHG_43 96448	RHG_44 18115	RHG_45 622427	RHG_46 45616	RHG_47 176669	RHG_48 22650

T 27 Totaux des marges de correction pour la non-réponse au niveau individuel. On a ici les sommes pour chacun des 48 GHR des pti_init2_k sur les 22174 ménages documentés. La somme totale redonne le nombre d'individus dans la population selon l'ESPOP, soit 6'186'711.

Les variables dichotomiques explicatives de la non-réponse utilisées par l'algorithme de segmentation pour construire l'arbre sont définies dans le tableau 6. L'arbre de segmentation est donné dans les figures 9 et 10.



G 9 Arbre de segmentation pour la modélisation de la non-réponse au questionnaire CATI/CAPI (partie de gauche).



G 10 Arbre de segmentation pour la modélisation de la non-réponse au questionnaire CATI/CAPI (partie de droite).

B Modélisation de la non-réponse à l'enquête écrite

Parmi les individus ayant répondu à l'enquête téléphonique, tous n'ont pas accepté de répondre à l'enquête écrite (questionnaire papier envoyé par la poste une fois l'interview téléphonique complétée). L'information collectée pour tous les individus répondant à l'enquête téléphonique a été utilisée pour modéliser, a posteriori, la probabilité de réponse au questionnaire écrit de la personne sélectionnée.

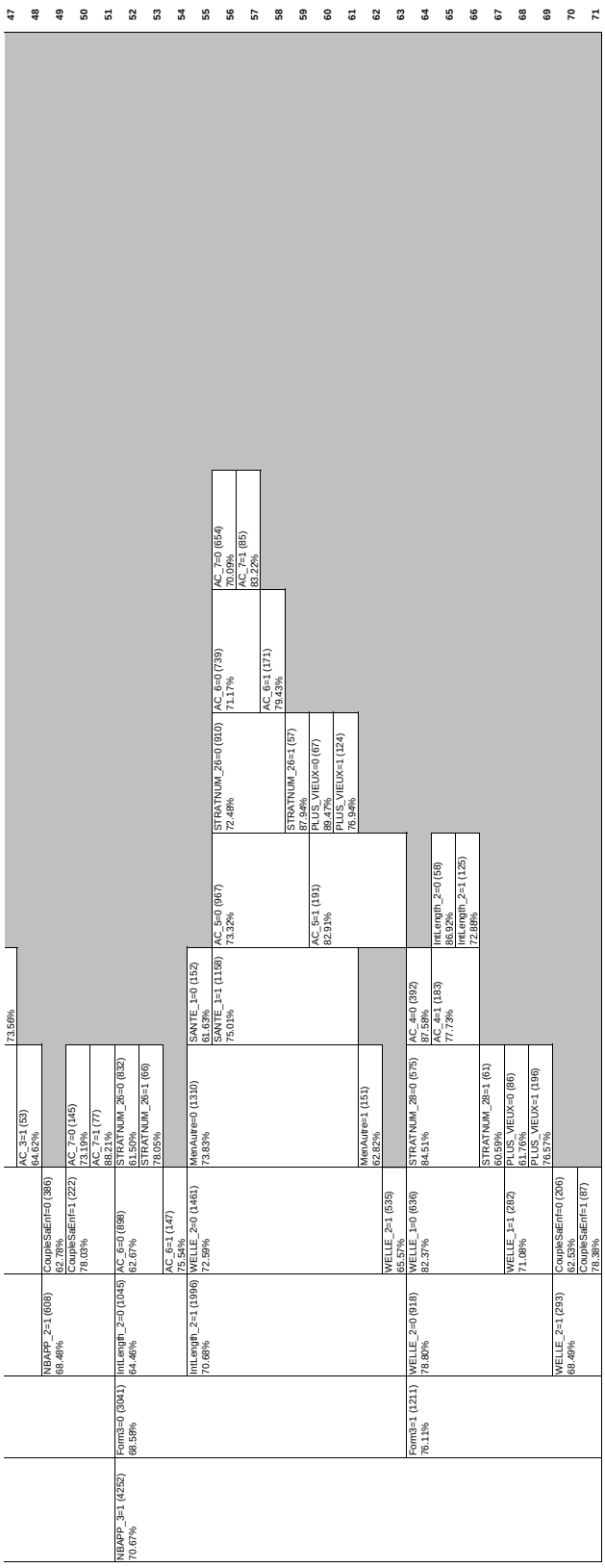
L'arbre (voir les figures 11 et 12) modélisant la non-réponse a été obtenu par une méthode de segmentation. Dans chacune des cases de l'arbre on trouve :
– la valeur correspondante de la variable dichotomique la définissant,
– entre parenthèses, le nombre de répondants correspondant à la case concernée,
– le taux de réponse des répondants ayant les caractéristiques de la case.
Les critères minimaux devant être satisfaits pour chaque case de l'arbre de segmentation ont été fixés dans ce cas précis à 60% de taux de réponse et 50 observations.
Les "feuilles" de l'arbre, i.e. les cases finales de chaque ligne du tableau définissent les groupes homogènes de réponse (GHR). On fait l'hypothèse que la non-réponse est uniforme dans chaque GHR. Pour chacun de ces GHR on calcule alors la somme des poids initiaux des individus (répondants ET non-répondants) y correspondant. Ces sommes définissent nos marges de correction pour la non-réponse. Les poids calés (=poids finaux) des répondants sommeront donc au même total que celui auquel sommaient les poids initiaux des répondants et des non-répondants. Les marges sont données dans le tableau 28. On y lit les sommes à un facteur près, pour chacun des 71 GHR, des pti_nrqi2_k sur les $18760 - 829 = 17931$ ménages répondants à l'enquête téléphonique. En effet, les sommes par GHR des pti_nrqi2_k sont légèrement inférieures à celles données ici du fait que les 829 interviews PROXY ont été déclarées comme inéligibles. On a donc multiplié les sommes par un facteur pour que le total des $pti_nrqi2_k \cdot f_k^{aj}$ sur tous les GHR redonne le nombre d'individus dans la population selon l'ESPOP, soit 6'186'711. Cette adaptation est nécessaire pour effectuer les calages lune/étoile décrits dans la section 7 : tous les types de marges doivent arriver à la même somme totale. Elle revient à faire une hypothèse de non-réponse ignorable à propos des interviews proxys.

RHG_1 227614	RHG_2 64967	RHG_3 29685	RHG_4 128777	RHG_5 66476	RHG_6 261727	RHG_7 43037	RHG_8 20445
RHG_9 39124	RHG_10 48462	RHG_11 52320	RHG_12 162470	RHG_13 34150	RHG_14 427289	RHG_15 22997	RHG_16 31941
RHG_17 57091	RHG_18 78807	RHG_19 27065	RHG_20 40867	RHG_21 38586	RHG_22 76664	RHG_23 135020	RHG_24 60984
RHG_25 50393	RHG_26 46091	RHG_27 65859	RHG_28 217806	RHG_29 30371	RHG_30 708568	RHG_31 33597	RHG_32 52579
RHG_33 38823	RHG_34 20779	RHG_35 20328	RHG_36 75888	RHG_37 36367	RHG_38 18742	RHG_39 18107	RHG_40 62389
RHG_41 187431	RHG_42 301113	RHG_43 29315	RHG_44 203537	RHG_45 27946	RHG_46 43196	RHG_47 75086	RHG_48 20291
RHG_49 119620	RHG_50 48446	RHG_51 23023	RHG_52 279804	RHG_53 21342	RHG_54 48482	RHG_55 39022	RHG_56 242806
RHG_57 21787	RHG_58 49743	RHG_59 18233	RHG_60 33770	RHG_61 37196	RHG_62 56095	RHG_63 186842	RHG_64 126534
RHG_65 19789	RHG_66 37488	RHG_67 18038	RHG_68 34564	RHG_69 58805	RHG_70 64968	RHG_71 39145	

T 28 Totaux des marges de correction pour la non-réponse au niveau individuel dans le cadre de la modélisation de la non-réponse au questionnaire écrit.

Les variables dichotomiques explicatives de la non-réponse utilisées par l'algorithme de segmentation pour construire l'arbre sont définies dans le tableau 8. L'arbre de segmentation est donné dans les figures 11 et 12.

G 11 Arbre de segmentation pour la modélisation de la non-réponse au questionnaire écrit (partie du haut).



G 12 Arbre de segmentation pour la modélisation de la non-réponse au questionnaire écrit (partie du bas).

Methodenberichte des Dienstes Statistische Methoden des BFS
Rapports de méthodes du Service de méthodes statistiques de l'OFS
Methodology reports published by the SFSO's Statistical Methods Unit

Graf, E. (2010). Enquête suisse sur la santé 2007. Plan d'échantillonnage, pondérations et analyses pondérées des données. Numéro de commande : 338-0061

Eichenberger P., Hulliger B., Potterat J. (2010). Describing the Anticipated Accuracy of the Swiss Population Survey. Order number : 338-0060

Graf, E. (2010). Étude empirique de l'attrition du Panel Suisse de Ménages : vers une caractérisation du profil du non-répondant. Numéro de commande : 338-0059

Graf, E. (2009). Weightings of the Swiss Household Panel : SHP_I wave 9, SHP_II wave 4, SHP_I et SHP_II combined. Order number : 338-0058

Graf, E. (2009). Pondérations du Panel Suisse de Ménages : PSM_I vague 9, PSM_II vague 4, PSM_I et PSM_II combinés. Numéro de commande : 338-0057-05

Qualité, L., Tillé, Y. (2009). Estimation de la précision d'évolutions dans l'enquête sur la valeur ajoutée. Numéro de commande : 338-0056

Renaud, A., Panchard, C. et Potterat, J. (2008). Statistique de l'emploi. Révision 2007 : méthodes d'estimation. Numéro de commande : 338-0055

Graf, E. (2008). Pondérations du PSM. PSM_I vague 8, PSM_II vague 3, PSM_I et PSM_II combinés. Numéro de commande : 338-0054

Andrade, B., Graf, M. (2008). Enquête suisse sur la structure des salaires 2006. Aspects méthodologiques du modèle des salaires "Salarium". Numéro de commande : 338-0053

Renaud, A. (2008). Statistique de l'emploi. Révision 2007 : cadre de sondage et échantillonage. Numéro de commande : 338-0052

Graf, E. (2008). Pondérations du SILC pilote. SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Numéro de commande : 338-0051

Kilchmann, D. (2008). Statistik der sozialmedizinischen Institutionen 1999-2004 und Krankenhausstatistik 1999-2002. Einsetzungen für fehlende Daten. Bestellnummer : 338-0050

Renaud, A. (2008). Technologies de l'information et de la communication. Estimations sur la base de la statistique de la valeur ajoutée. Numéro de commande : 338-0049

Assoulin, D. (2007). Wertschöpfungsstatistik. Einsetzungsversuche für fehlende Antworten grosser Unternehmen. Bestellnummer : 338-0048

Kilchmann, D. (2007). Beherbergungsstatistik Campingplätze. Stichprobenrahmen und Schätzverfahren 2005/06. Bestellnummer : 338-0047

Gabler, S., Häder, S. (2007). Haushalts- und Personenerhebungen. Machbarkeit von Random Digit Dialing in der Schweiz. Bestellnummer : 338-0046

Ferrez, J., Graf, M. (2007). Enquête suisse sur la structure des salaires. Programmes R pour l'intervalle de confiance de la médiane. Numéro de commande : 338-0045

Renaud, A. (2007). Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles. Numéro de commande : 338-0044

Potterat, J. (2007). Betriebszählung 2005. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0043

Hulliger, B. (2006). Umweltschutzausgaben der Unternehmen 2003, Stichprobenplan, Datenaufbereitung und Schätzverfahren. Bestellnummer : 338-0042

Renfer, J.-P. (2006). Enquête sur les chiffres d'affaires du commerce de détail. Plan d'échantillonnage et méthodes d'estimation. Numéro de commande : 338-0041

Salamin, P.-A. (2006). Statistique de l'aide sociale dans le domaine de l'asile. Plan de sondage et extrapolations pour l'enquête pilote 2005. Numéro de commande : 338-0040

Renaud, A. (2006). Statistique suisse des bénéficiaires de l'aide sociale. Pondération des communes 2004. Numéro de commande : 338-0039

Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample : Analysis and precision assessment of wage components. Order number : 338-0038

Potterat, J. (2006). Pensionskassenstatistik 2004. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. Bestellnummer : 338-0037

Potterat, J. (2006). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2004. Stichprobenplan, Gewichtung und Schätzverfahren. Bestellnummer : 338-0036

Kilchmann, D. (2006). Vierteljährliche Wohnbaustatistik. Stichprobenplan, statistische Datenaufarbeitung und Schätzverfahren 2005. Bestellnummer : 338-0035

Kilchmann, D. (2006). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2004. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer : 338-0034

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 2. Bestellnummer : 338-0033

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 1. Bestellnummer : 338-0032

Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro de commande : 338-0031

Graf, E., Renfer, J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro de commande : 338-0030

Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. Bestellnummer : 338-0029

Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. Bestellnummer : 338-0028

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Order number : 338-0027

Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. Bestellnummer : 338-0026

Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro de commande : 338-0025

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro de commande 338-0024

Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro de commande : 338-0023

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande : 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer : 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer : 338-0020.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande : 338-0019

Hulliger, B. (2003). Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Erhebung über F+E in der schweizerischen Privatwirtschaft 2000. Bestellnummer : 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande : 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer : 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande : 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer : 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des é lèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande : 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d' échantillonnage. Numéro de commande : 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer : 338-0011

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande : 338-0010

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande : 338-0009

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer : 338-0008

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande : 338-0007

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande : 338-0006

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande : 338-0005

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer : 338-0004

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande : 338-0003

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer : 338-0002

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande : 338-0001

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen : Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer : 338-0000

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active : rapport de pondération pour 1996.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse : établissement du plan d'échantillonnage.

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Hulliger, B. (1996). Haushalterhebung Familie 1994 : Stichprobenplan, Stichprobenziehung und Reservestichproben.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994 : plan d'échantillonnage et pondération.

Hulliger, B. (1996). Gütertransportstatistik 1993 : Schätzverfahren mit Kompensation der Antwortausfälle.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Peters, R. (1995). Enquête de structure sur les loyers : établissement d'un plan d'échantillonnage stratifié.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung : Stichprobenplan und Schätzverfahren.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Peters, R., Hulliger, B. (1994). La technique de pondération des données : application à l'enquête suisse sur la santé.

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse : stratégie d'échantillonnage pour le secteur privé.

Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statistische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln

Programme des publications de l'OFS

En sa qualité de service central de statistique de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public.

L'information statistique est diffusée par domaine (cf. verso de la première page de couverture); elle emprunte diverses voies:

Diffusionsmittel	*Kontakt* *N° à composer*	*Moyen de diffusion*
Individuelle Auskünfte	032 713 60 11 info@bfs.admin.ch	Service de renseignements individuels
Das BFS im Internet	www.statistik.admin.ch	L'OFS sur Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch	Communiqués de presse: information rapide concernant les résultats les plus récents
Publikationen zur vertieften Information (zum Teil auch als Diskette/CD-Rom)	032 713 60 60 order@bfs.admin.ch	Publications: information approfondie (certaines sont disponibles sur disquette/CD-Rom)
Online Datenrecherche (Datenbanken)	www.statdb.bfs.admin.ch	Données interactives (banques de données, accessibles en ligne)

Nähere Angaben zu den verschiedenen Diffusionsmitteln im Internet unter der Adresse www.statistik.admin.ch→ Dienstleistungen→Publikationen Statistik Schweiz

Informations sur les divers moyens de diffusion sur Internet à l'adresse www.statistique.admin.chRServices RLes publications de Statistique suisse.

Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodology Reports by the Statistical Methods Unit

Die Methodenberichte beschreiben die mathematischen und statistischen Methoden, die den Resultaten und Analysen der öffentlichen Statistik zu Grunde liegen. Sie enthalten ausserdem die Evaluation und Entwicklung von neuen Methoden im Hinblick auf eine zukünftige Anwendung. Diese Publikationen sollen einerseits die verwendeten Methoden dokumentieren, um Transparenz und Wissenschaftlichkeit sicher zu stellen, und sie sollen andererseits die Zusammenarbeit mit den Hochschulen und der Wissenschaft fördern.

Zur Illustration der beschriebenen mathematischen Konzepte, werden im Bericht numerische Resultate aufgeführt. Diese sind allerdings nicht als offizielle Resultate der betreffenden Erhebungen zu verstehen. Ebenfalls können die tatsächlich angewendeten Methoden leicht von den hier beschriebenen abweichen.

Die Methodenberichte sind auf der Internetseite des BFS in elektronischer Form verfügbar.

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le déve-loppement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes uti-lisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site internet de l'OFS.

L'ESS07 a pour objectif principal de fournir, à partir d'un échantillon de près de 19'000 personnes résidant en Suisse et âgées de 15 ans et plus, des résultats sur l'état de santé de la population, les comportements en matière de santé, l'assurance-maladie et le recours aux prestations du système de santé.

Ce rapport décrit les méthodes utilisées pour établir le plan d'échantillonnage en abordant les questions relatives à la délimitation de la population, au cadre de sondage et aux unités d'échantillonnage. Le tirage, la gestion et l'érosion des échantillons sont également abordés dans la première partie.

L'ESS se compose en fait de deux enquêtes: la première est une enquête téléphonique alors que la seconde, sous forme écrite, ne s'adresse qu'aux personnes ayant répondu à la première. La pondération relative à chacune de ces enquêtes est décrite dans la deuxième partie de ce rapport.

Enfin, la troisième partie tente d'apporter des réponses aux questions les plus courantes reçues par le service des méthodes statistiques à propos de l'utilisation et de l'exploitation des données de l'ESS à l'aide des pondérations.

N° de commande
338-0061

Commandes
Tél.: 032 713 60 60
Fax: 032 713 60 61
E-mail: order@bfs.admin.ch

Prix
gratuit

ISBN 978-3-303-00436-4