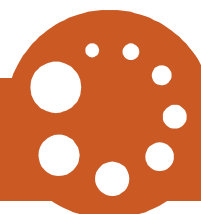




Enquête sur le budget des ménages (EBM)

Révision de la pondération 2019

EXPERIMENTAL STATISTICS



Neuchâtel, 2022

Éditeur:	Office fédéral de la statistique (OFS)	Concept de mise en page:	Section DIAM
Renseignements:	habe@bfs.admin.ch	Téléchargement:	www.experimental.bfs.admin.ch
Rédaction:	Rémy Mariétan, Anne Massiani, Daniel Kilchmann	Copyright:	OFS, Neuchâtel 2022
Domaine:	00 Bases statistiques des généralités		La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée
Langue du texte original:	Français		

Table des matières

1	Introduction et déroulement de l'enquête	2
1.1	Enquête	2
1.2	Mode de relevé	2
2	Population, échantillon, non-réponse, calage et extrapolation	2
2.1	Population et échantillon	2
2.2	Non-réponse	3
2.2.1	Taux de réponse	3
2.2.2	Variables auxiliaires	3
2.3	Calage	4
2.3.1	Marges	4
2.3.2	Variables auxiliaires	4
2.3.3	Objectifs du calage	4
2.4	Extrapolation	4
2.5	Résumé des spécificités et l'enquête	4
3	Pondération : variantes de solutions et discussion du choix	5
3.1	Non-réponse	5
3.1.1	Variantes possibles et choix	5
3.1.2	Description de la variante retenue	6
3.1.3	Poids corrigés pour la non-réponse	6
3.1.4	Analyse du modèle et des poids	6
3.2	Calage	7
3.2.1	Choix des variables	7
3.2.2	Choix de la fonction de distance	8
3.2.3	Analyse du modèle et des poids	8
3.2.4	Adaptations ultérieures des poids	10
3.3	Extrapolation et calcul de la précision	10
3.3.1	Indicateurs	10
3.3.2	Résultats	13
3.3.3	Évolution des estimations	13
4	Conclusions	15
5	Références	17
A	Appendice	18
B	Appendice	19

1 Introduction et déroulement de l'enquête

1.1 Enquête

L'enquête sur le budget des ménages (EBM, HABE en allemand) de la section EKL permet de connaître les dépenses et les revenus des ménages en Suisse. Un échantillon d'environ 12'000 ménages est tiré aléatoirement pour participer à l'enquête. Il est demandé aux participants de remplir pendant le mois d'enquête un cahier contenant tous les revenus et toutes les dépenses du ménage. Renseigner les dépenses alimentaires représentent une tâche conséquente pour les participants. C'est pourquoi ces dépenses ne sont demandées que sur la première moitié respectivement la deuxième moitié du mois. Les dépenses plus conséquentes comme l'achat d'une voiture sont quant à elles aussi indiquées même si ces achats sont faits plusieurs mois avant le mois d'enquête. Puisque la tâche demandée aux participants est conséquente, le taux de réponse de cette enquête est plus faible que dans la plupart des enquêtes de l'OFS. De plus, comme présenté plus loin dans le document, l'EBM d'une année utilise des échantillons tirés dans 5 cadres. Cela soulève des questions quant aux totaux à utiliser pour le calage.

En se basant sur les données de 2016 à 2018, la section METH a développé une nouvelle méthodologie de pondération pour l'enquête 2019. En effet, la disponibilité d'informations supplémentaires est exploitée au moment de la correction de la non-réponse ainsi que de la procédure de calage. Ces informations supplémentaires proviennent notamment des cadres de tirage de l'OFS et des données de la caisse de compensation relatives aux revenus des personnes (CDC).

1.2 Mode de relevé

Les ménages sélectionnés dans l'échantillon brut sont tous contactés par une lettre officielle de l'OFS, puis ensuite par téléphone, s'il est disponible à l'OFS. Les participants reçoivent un appui avant, durant et après le mois d'enquête pour les aider à remplir les questionnaires afin d'améliorer la qualité des données. De plus, un effort important est investi pour améliorer le taux de réponse.

2 Population, échantillon, non-réponse, calage et extrapolation

2.1 Population et échantillon

La population cible correspond aux ménages privés contenant au moins un résident permanent âgé de 13 ans ou plus en Suisse. Le ménage privé se définit comme un groupe de personnes qui partagent le même logement, mettent en commun une partie ou la totalité de leurs revenus et de leur patrimoine, et consomment collectivement certains biens et services. Une personne n'appartient qu'à un seul ménage dans lequel elle vit la majorité de la semaine et contribue aux dépenses communes du ménage.

Pour bénéficier de la meilleure actualisation des cadres possibles, le tirage est effectué à partir des données de 5 cadres de tirage d'enquêtes auprès des personnes et des ménages, SRPH (Stichprobenregister für Personen- und Haushaltserhebungen). Ainsi, pour une année d'enquête, 5 échantillons différents (un par cadre) sont tirés. Le premier échantillon ne portera que sur les 2 premiers mois et proviendra d'un premier cadre. Les deuxième, troisième et quatrième échantillons porteront chacun sur 3 mois et proviendront respectivement d'un deuxième, troisième et quatrième cadre. Enfin le dernier mois est couvert par un échantillon provenant d'un dernier cadre. La table 1 illustre cette configuration pour l'enquête 2016. La table 2 présente, pour 2016, 2017 et 2018, les tailles des cadres, échantillons bruts (ménages sélectionnés) et échantillons nets (ménages répondants). On constate une augmentation de la taille des cadres qui correspond bien à l'évolution démographique en Suisse. Lorsqu'un ménage est contacté et accepte de participer à l'enquête, il se peut que la composition réelle du ménage diffère de celle du cadre. C'est par exemple le cas d'un ménage qui a perdu un

membre (décès, déménagement...) ou qui comprend un membre supplémentaire (naissance, emménagement...). De plus, il existe des situations où le concept de composition de ménages n'est pas identique entre les cadres et l'enquête. C'est le cas par exemple d'une fille au pair qui n'est pas membre du même ménage dans le cadre. En se basant sur la définition d'un ménage ci-dessus, si un individu partage une partie du budget du ménage et y réside la majorité de la semaine (Ce qui est le cas en général pour la jeune fille au pair), il est considéré comme faisant partie du ménage.

Table 1 Relation entre mois d'enquête et cadre de tirage pour l'enquête 2016.

Année d'enquête	Mois d'enquête	Num. Cadre utilisé	Date du cadre
2016	1	152	30.06.2015
2016	2	152	30.06.2015
2016	3	153	30.09.2015
2016	4	153	30.09.2015
2016	5	153	30.09.2015
2016	6	154	31.12.2015
2016	7	154	31.12.2015
2016	8	154	31.12.2015
2016	9	161	31.03.2016
2016	10	161	31.03.2016
2016	11	161	31.03.2016
2016	12	162	30.06.2016

Table 2 Taille des 5 cadres, N_1 , N_2 , N_3 , N_4 et N_5 . Tailles des échantillons bruts et nets annuels, n et m . Taux de réponse, *taux*.

	N_1	N_2	N_3	N_4	N_5	n	m	<i>taux</i>
2016	3'614'112	3'626'193	3'636'059	3'639'029	3'657'477	10'570	3'269	31%
2017	3'657'477	3'668'052	3'683'616	3'696'359	3'702'659	10'470	3'217	31%
2018	3'702'659	3'711'419	3'721'768	3'731'683	3'742'095	11'360	3'364	30%

2.2 Non-réponse

2.2.1 Taux de réponse

Comme dit plus haut, le taux de réponse de l'enquête (EBM) d'environ 30% est plus faible que dans d'autres enquêtes de l'OFS, voir table 2. Ce taux pose un défi pour la pondération et nécessite d'exploiter au mieux les informations contenues dans les variables auxiliaires disponibles.

2.2.2 Variables auxiliaires

Les variables explicatives à disposition regroupent toutes les informations sur les personnes et les ménages de la statistique de la population et des ménages (STATPOP) appartenant au SRPH, ainsi que les données de la CDC datées approximativement d'un an et demi avant le mois enquêté, des variables de typologie des ménages construites par EKL (à partir de STATPOP) et des variables caractérisant les logements (nombre d'appartements, nombre de pièces, nombre d'étages) provenant de la statistique des bâtiments et logements (StatBL). À cela s'ajoutent des variables connues uniquement sur l'échantillon brut telles qu'une variable indiquant la présence d'un numéro de téléphone

pour contacter le ménage, le mois d'enquête et une variable indiquant que le ménage relève ses dépenses alimentaires durant la première ou la deuxième partie du mois.

2.3 Calage

2.3.1 Marges

Les marges de calages sont des totaux de variables auxiliaires connus sur la population. Ces marges sont calculées à partir des marges des 5 cadres de tirages augmentés des informations de la CDC et de la StatBL. Elles sont des moyennes des totaux calculés sur les cinq cadres, pondérées par le nombre de mois où le cadre est utilisé (2,3,3,3,1, cf. table 1).

2.3.2 Variables auxiliaires

Les variables auxiliaires disponibles sont les mêmes que pour la modélisation du mécanisme de non-réponse, cf. 2.2.2 à l'exception bien entendu des variables connues exclusivement sur l'échantillon telles que la présence d'un numéro de téléphone, le mois d'enquête ou la vague.

2.3.3 Objectifs du calage

Nous effectuons un calage pour plusieurs raisons.

- Premièrement, le risque de biais devrait ainsi diminuer.
- De plus, ayant à disposition des variables relativement corrélées avec certaines variables publiées, nous espérons un gain de précision.
- Ensuite, la stabilité au fil des années devrait être augmentée. En effet, le taux de réponse faible et la petite taille de l'échantillon pourraient induire des fluctuations conséquentes.
- Enfin, le calage permettra une certaine cohérence entre des statistiques publiées à différentes échelles.

2.4 Extrapolation

Les résultats produits se basent sur une pondération qui attribue à chaque ménage un poids lors de la construction d'estimateurs. Le développement de la nouvelle pondération s'est basé sur les données de 2016, 2017 et 2018. La section EKL a alors proposé d'évaluer la procédure sur 4 variables clés enquêtées au niveau suisse :

- Revenu brut ;
- A50m : total mensuel des dépenses ;
- A51m : dépenses mensuelles pour l'alimentation et les boissons (sans alcool) ;
- A57m : dépenses mensuelles pour le logement et l'énergie.

Ainsi la recherche du modèle de calage est effectuée dans le but d'améliorer les résultats au niveau suisse en réduisant le risque de biais et en augmentant la précision.

2.5 Résumé des spécificités et l'enquête

Il y a quelques spécificités dans l'enquête EBM qui doivent être prises en considération.

- Malgré un taux de réponse très bon au vu de l'effort demandé aux participants, le pourcentage de réponse plus faible que la moyenne de la majorité des enquêtes de l'OFS soulève des difficultés particulières quant à la correction de la non-réponse.

- L'EBM utilise plusieurs cadres de tirage au fil des mois. Cette spécificité implique qu'il faut être consistant dans l'utilisation des variables auxiliaires et considérer leurs valeurs au moment du tirage, en tenant compte du fait qu'il y a plusieurs cadres de tirage. De plus, l'utilisation de plusieurs cadres de tirage doit également être prise en compte lors du calcul des marges de calage.

3 Pondération : variantes de solutions et discussion du choix

Les ménages de l'échantillon brut se voient attribuer un poids initial construit par la section EKL. La section METH adapte ensuite la pondération en deux étapes :

- 3.1 Correction de la non-réponse ;
- 3.2 Calage.

Divers traitements de la non-réponse et différents calages ont été étudiés en détail afin d'élaborer la pondération finale.

3.1 Non-réponse

3.1.1 Variantes possibles et choix

Pour le traitement de la non-réponse, une première variante avec un traitement de la non-réponse par un modèle logistique a été analysée avant d'être rejetée. En effet le taux de réponse faible créait quelques poids très grands qui mettaient en péril la stabilité des résultats. Une adaptation de cette méthode consistant à tronquer à 10% les probabilités de réponse trop faibles a été analysée et comparée à une segmentation bornant les probabilités de réponse à 10%, 15% et 20%. Pour comparer les approches nous observons les concordants et les discordants décrits dans [SAS Institute Inc. \(2017, p. 5575\)](#). Les méthodes par segmentation bornant la probabilité de réponse à 15% et 20% se sont avérées avoir une concordance trop faible comparée aux deux autres approches. La segmentation bornée à 10%, qui permet un compromis entre différents critères, a finalement été retenue pour les raisons ci-dessous.

- Les performances de la régression logistique tronquée et de la segmentation, avec un taux minimum de réponse fixé à 10% dans les groupes de réponse homogènes, sont équivalentes en termes de taux de concordants / discordants.
- Les évolutions dans le temps des estimateurs des variables clés basées sur la pondération après traitement de non-réponse sont similaires entre les deux approches.
- La segmentation permet de fixer "proprement" une borne inférieure à 10% pour les probabilités de non-réponse. Avec la régression logistique, ce seuil est franchi par quelques probabilités de réponse et implique de tronquer les probabilités de réponse. Borner "proprement" les probabilités de réponse est un avantage permis par la segmentation qui serait d'autant plus important si le taux de réponse devait baisser dans le futur. Un tel comportement du taux de réponse entraînerait une troncature plus excessive pour la régression logistique.
- Les poids après calage obtenus par une segmentation contiennent moins de très gros poids que ceux obtenus par une approche logistique. Ce qui laisse penser que la segmentation sera plus stable.
- La segmentation, qui fournit des performances et des résultats équivalents à ceux de la régression logistique pour les années 2016, 2017 et 2018, ne semble donc pas moins stable que la régression logistique en l'absence de changements brutaux dans le comportement de réponse. La segmentation pourrait néanmoins être plus adaptative en cas de changement dans le mécanisme de réponse, par exemple en lien avec la pandémie du COVID-19. La segmentation

permettrait en effet de créer automatiquement des croisements plus pertinents pour la nouvelle situation.

3.1.2 Description de la variante retenue

La variante retenue est une segmentation utilisant 40 variables binaires listées ci-dessous et décrites plus en détail en appendice A :

1	Auslaendertyp0	11	Geschlechterttyp0	21	Strate017	31	cdc_classe9
2	Auslaendertyp1	12	Geschlechterttyp1	22	cdc_classe0	32	indice_ind
3	Auslaendertyp2	13	Geschlechterttyp2	23	cdc_classe1	33	indice_nbrev0
4	Auslaendertyp3	14	Paartyp	24	cdc_classe2	34	indice_nbrev1
5	Familiientyp08_SRP230	15	Strate011	25	cdc_classe3	35	indice_nbrev2
6	Familiientyp08_SRP400	16	Strate012	26	cdc_classe4	36	indice_nbrev3
7	Familiientyp08_SRP500	17	Strate013	27	cdc_classe5	37	indice_rente
8	FamiliientypAlter08_SRP105	18	Strate014	28	cdc_classe6	38	mAnzahlWohnungen151
9	FamiliientypAlter08_SRP106	19	Strate015	29	cdc_classe7	39	mitTelefonnummer0
10	FamiliientypAlter08_SRP205	20	Strate016	30	cdc_classe8	40	indiceap8

Cet ensemble de variables dit "affiné" est une combinaison des variables retenues durant les 5 premiers splits par la segmentation appliquée avec toutes les variables disponibles, mais aussi des variables retenues par l'approche logistique. Ce choix est motivé par le souhait de considérer un ensemble de variables suffisamment riche pour ne pas amoindrir le pouvoir explicatif du modèle, sans pour autant conserver toutes les variables à disposition, afin d'apporter une certaine stabilité au modèle. Les marges de l'échantillon brut ont par ailleurs été comparées à celles obtenues à partir de l'échantillon net utilisant les poids corrigés pour la non-réponse et ce pour toutes les marges construites à partir des variables auxiliaires à disposition. Cette vérification a mis en évidence que les plus de 80 ans ne sont pas bien pris en compte par le modèle de non-réponse. Nous avons donc ajouté la variable indiceap8 qui indique si une personne de plus de 80 ans est présente dans le ménage. Cet ajout a amélioré le résultat des plus de 80 ans.

3.1.3 Poids corrigés pour la non-réponse

Les poids corrigés pour la non-réponse s'obtiennent alors en multipliant les poids initiaux, w_i^{ini} produits par la section EKL et reflétant l'inverse de la probabilité de tirage (cf. [Cornali Schweingruber et al. \(2007\)](#)), avec l'inverse de $\hat{\theta}_i$, la probabilité de réponse estimée au sein des groupes homogènes de réponse, RHG, résultant de la segmentation :

$$w_i^{NR} = w_i^{ini} \frac{1}{\hat{\theta}_i}, i \in R,$$

$$\hat{\theta}_i = \frac{m_l}{n_l} \text{ pour } i \in RHG_l,$$

où m_l est le nombre de ménages répondant dans RHG_l et n_l le nombre de ménages de l'échantillon dans RHG_l . L'arbre résultant de la segmentation de l'année 2017 est visible sur la figure 5 en appendice B et divise la population en 32 RHG. Les segmentations appliquées à d'autres années produisent des résultats assez similaires en terme de nombre de RHG et de variables retenues dans les premiers "splits" de la segmentation. L'ordre d'utilisation des variables dans l'arbre peut en revanche différer.

3.1.4 Analyse du modèle et des poids

Pour les trois années, nous obtenons les taux de concordance et de discordance de la segmentation choisie dans la table 3. Les taux de concordance et discordance restent relativement stables pour les années 2016 à 2018. Néanmoins, on observe d'une part une légère dégradation de la concordance au

fil du temps ainsi qu'un taux de non classifiées (tied/égaux) qui croît légèrement au détriment des concordants.

Table 3 Taux de concordance-discordance du modèle de non-réponse pour 2016, 2017 et 2018.

	2016	2017	2018
Concordant	64.96	62.92	60.72
Discordant	31.09	32.4	31.44
Tied	3.95	4.68	7.84

Les quantiles des poids avant et après traitement de la non-réponse se basant sur la segmentation choisie sont présentés dans la table 4. Pour tous les quantiles, on observe certaines fluctuations entre les années aussi bien pour les poids initiaux que pour les w^{NR} . Néanmoins, l'étendue des deux poids reste limitée et les distributions des poids restent tout à fait comparables au fil des années.

Table 4 Distributions des poids avant et après traitement de la non-réponse.

année	nom	m	total	P0	P1	P5	P10	P25	P50	P75	P90	P95	P99	P100	P100/P0
2016	Poids initiaux	3269	1139255.6	123.2	125.5	131.5	144.1	347.9	372.1	387.8	402.5	412.3	430.5	441.8	3.6
	Poids après traitement NR		3662059.4	228.4	267.0	392.6	633.5	771.7	973.3	1270.0	1912.7	2425.0	3585.4	4199.9	18.4
2017	Poids initiaux	3217	1134621.1	127.1	127.1	134.4	137.7	354.5	379.5	399.2	412.6	417.6	424.2	437.3	3.4
	Poids après traitement NR		3667306.8	263.6	278.7	372.5	679.2	811.5	1014.8	1306.0	1912.6	2256.1	2958.7	3378.4	12.8
2018	Poids initiaux	3364	1095456.6	98.5	98.5	127.9	138.1	285.4	357.8	394.8	414.1	424.2	445.6	461.1	4.7
	Poids après traitement NR		3678145.4	190.7	248.7	363.1	537.9	765.7	988.3	1265.0	1730.8	2092.8	3579.8	4259.1	22.3

3.2 Calage

Le développement du calage consiste à choisir des variables de calage parmi les variables auxiliaires, puis à caler les marges de ces variables auxiliaires sur leurs marges obtenues par pondération (2,3,3,3,1) des marges des 5 cadres de calages utilisés. Le choix des variables de calage se base sur l'étude de la corrélation entre les variables auxiliaires et les variables clés afin de tenter de maximiser le potentiel gain de précision permis par le calage.

3.2.1 Choix des variables

Les variables retenues pour le calage ont un certain pouvoir explicatif des variables clés proposées par la section EKL. Il s'agit de (cf. appendice A) :

1	nbre_homme	Nombre d'hommes dans le ménage
2	nbre_suisse	Nombre de suisses dans le ménage
3	nbre_age3_0-nbre_age3_8	Nombre de personnes dans le ménage entre : 0-10, 10-20,..., 70-80 et 80-
4	strate01	Grandes régions
5	Familiotyp08_SRPB	Typologie ménage proposée par EKL
6	cdc_classe	Classe CDC total du ménage. Répartition par quantile de 10%
7	mAnzahlWohnungen15_1_2	Indicateur d'un logement individuel, à deux appartements ou autre
8	Familiotyp08_SRPB*commul_cdc_rev_trunc_099	Croisement de la typologie ménage avec le revenu CDC du ménage tronqué à 0.99%

Pour choisir ces variables, nous avons tenté d'expliquer les 4 variables clés avec les variables auxiliaires à disposition. Le choix des variables s'est fait en deux étapes.

— Optimisation automatique d'un modèle

Premièrement, le nombre de variables explicatives est réduit par backward sélection et cross-validation réalisé avec la commande SAS proc glmselect documentée dans [SAS Institute Inc. \(2015\)](#). Un modèle retenu permet d'expliquer chaque variable clé de chaque année. Les coefficients de détermination R^2 qui en découlent sont utilisés comme référence dans les développements qui suivent. Néanmoins le choix des variables retenues n'est pas toujours facilement explicable/communicable.

— Construction d'un modèle compact

Nous sélectionnons ensuite les variables listées ci-dessus en nous basant sur 4 critères :

- Premièrement, le choix d'une variable doit être communicable. Ainsi, ne retenir que 3 classes de revenus CDC sur les 10 semble difficile à expliquer.
- D'un autre côté, nous ne voulons conserver que des variables dont la contribution au R^2 pour au moins une variable clé est pertinente. L'impact sur le gain de variance après calage est alors lui aussi pertinent. Par exemple, la variable *paartyp*, étant déjà incluse lors du traitement de la non-réponse et de contribution modeste au R^2 , n'est pas retenue.
- Ensuite, certaines variables ont été supprimées ou agrégées sur conseil de la section EKL. C'est le cas de la variable sur les âges qui regroupe les âges par tranche de 10 ans et fusionne les personnes plus âgées que 80 ans dans une catégorie, ou encore de la variable sur la langue de communication du ménage qui ne contribuait que peu au calage et a été exclue.
- Enfin, nous souhaitons conserver des R^2 comparables aux R^2 de référence construits sans les contraintes énumérées dans ce paragraphe. Les R^2 du modèle compact retenu sont proches des R^2 du modèle de référence tout en limitant la saturation du modèle. Le modèle retenu est alors un modèle à 8 variables décrites ci-dessus.

3.2.2 Choix de la fonction de distance

Le calage a été effectué à l'aide de CALMAR2 en appliquant la méthode linéaire tronquée. Les g-weights ont ainsi été bornés entre 0.4 et 2 afin d'éviter des poids extrêmes et pour renforcer la stabilité des résultats. Nous obtenons ainsi les poids calés w^C :

$$w_i^C = \frac{g_i w_i^{ini}}{\hat{\theta}_i} = g_i w_i^{NR}, \quad i \in R,$$

où w_i^{ini} sont les poids initiaux de l'unité i , $\hat{\theta}_i$ sont les probabilités de réponse estimées pour l'unité i , w_i^{NR} sont les poids corrigés pour la non-réponse et g_i sont les g-weights.

3.2.3 Analyse du modèle et des poids

Les coefficients de détermination R^2 obtenus par régression linéaire pondérée des 4 variables clés sur les variables de calage sont présentés dans la table 5 pour les 3 années. Plus précisément, nous nous intéressons pour chaque variable clé Y à la relation linéaire suivante,

$$Y_i = \beta \mathbf{X}_i + \epsilon_i, \quad i \in U,$$

où $\mathbf{X}$ désigne les variables de calage. La régression appliquée à l'échantillon des répondants R est alors pondérée par les poids corrigés pour la non-réponse, w^{NR} , et β est estimé par la méthode des moindres carrés pondérés. On remarque dans la table 5 qu'une part non négligeable des dépenses et des revenus est explicable par les variables auxiliaires. Comme présenté dans sa forme simplifiée dans Tillé (2019, p. 179), on peut espérer que les variables auxiliaires aient un effet bénéfique sur l'estimation de la variance des variables clés d'un facteur multiplicateur d'environ $1 - R^2$. La distribution des poids à chaque étape se trouve dans la table 6, qui reprend et complète la table 4. La table indique aussi que le ratio entre les poids maximums et les poids minimums peut atteindre 56. Néanmoins il s'agit de petits pourcentages des données et la dispersion est moins forte sur le 98% ($P99/P1$) de l'échantillon des répondants. La table montre aussi que le pourcentage d'unités dont les poids atteignent les bornes est très faible. Les histogrammes des poids au fil des étapes de la pondération sont présentés sur la figure 1 pour l'année 2017. Premièrement on note que la bi-modalité de la distribution des poids initiaux est atténuée par les traitements. Bien que le calage augmente la dispersion des poids, les distributions avant et après calage sont proches et les estimateurs qui en découlent devraient être généralement proches.

Table 5 Coefficient R^2 de la régression pondérée des variables clés sur les variables de calage pour toutes les années.

	2016	2017	2018
A50m	0.36	0.38	0.27
A57m	0.17	0.24	0.23
A51m	0.36	0.36	0.34
Bruttoeinkommen08	0.57	0.53	0.51

Table 6 Distributions des Poids au fil des étapes.

année	nom	m	total	P0	P1	P5	P10	P25	P50	P75	P90	P95	P99	P100	P100/P0
2016	Poids initiaux	3269	1139255.6	123.2	125.5	131.5	144.1	347.9	372.1	387.8	402.5	412.3	430.5	441.8	3.6
	Poids après traitement NR		3662059.4	228.4	267.0	392.6	633.5	771.7	973.3	1270.0	1912.7	2425.0	3585.4	4199.9	18.4
	g-weight		3222.3	0.40	0.57	0.71	0.76	0.86	0.97	1.08	1.23	1.34	1.60	2.00	5.00
	Poids après calage		33636075.8	158.3	250.0	392.2	533.4	702.6	923.9	1291.1	1908.8	2573.5	3873.4	4906.0	31.0
2017	Poids initiaux	3217	1134621.1	127.1	127.1	134.4	137.7	354.5	379.5	399.2	412.6	417.6	424.2	437.3	3.4
	Poids après traitement NR		3667306.8	263.6	278.7	372.5	679.2	811.5	1014.8	1306.0	1912.6	2256.1	2958.7	3378.4	12.8
	g-weight		3153.6	0.40	0.55	0.68	0.75	0.85	0.96	1.09	1.23	1.33	1.62	2.00	5.00
	Poids après calage		3683906.3	135.2	218.3	332.5	517.1	733.1	969.0	1368.3	2045.3	2634.6	3436.9	6007.3	44.4
2018	Poids initiaux	3364	1095456.6	98.5	98.5	127.9	138.1	285.4	357.8	394.8	414.1	424.2	445.6	461.1	4.7
	Poids après traitement NR		3678145.4	190.7	248.7	363.1	537.9	765.7	988.3	1265.0	1730.8	2092.8	3579.8	4259.1	22.3
	g-weight		3330.2	0.40	0.50	0.64	0.72	0.84	0.97	1.11	1.27	1.40	1.76	2.00	5.00
	Poids après calage		3723454.9	101.9	173.6	287.7	428.6	689.6	962.7	1345.8	1906.0	2381.7	3869.2	5750.6	56.4

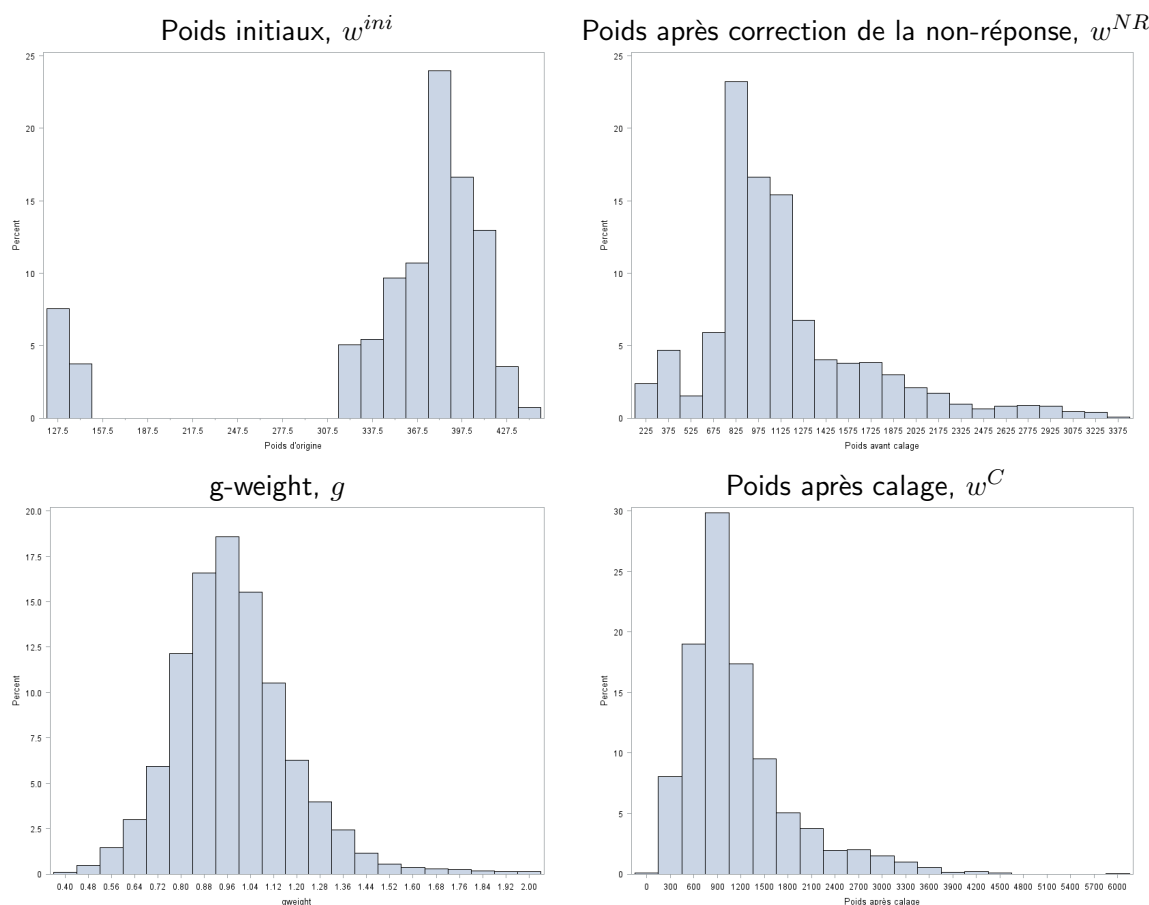


Figure 1 Histogramme des poids initiaux, des poids après correction de la non-réponse, des poids après calage et des g-weights pour l'année 2017.

Remarque Notez que la bi-modalité des poids initiaux est une conséquence de la surreprésentation, voulue par la section EKL, de la plus petite "Grande Région". Le but étant que les tailles des échantillons dans les différentes "Grandes Régions" soient proches.

3.2.4 Adaptations ultérieures des poids

Suite aux difficultés de caler en conservant les bornes 0.4 et 2, la section EKL a renoncé aux bornes. De plus un facteur de correction a été appliqué aux poids pour que les totaux du nombre de personnes correspondent au nombre de personnes dans les ménages économiques.

3.3 Extrapolation et calcul de la précision

3.3.1 Indicateurs

Nous estimons essentiellement des moyennes, des ratios et des quantiles. Pour chacun de ces paramètres, nous considérons un estimateur basé sur les poids calés ainsi que différents estimateurs de variance. Nous ne présentons dans ce rapport que les estimations de la variance obtenues par la formule de Deville qui sont ici très proches de celles obtenues par `proc surveymeans` ou par une formule exacte. Ci-dessous nous développons toutes les formules utilisées pour construire les estimateurs et estimer les variances.

Estimateur

Total :

Pour estimer un total Y sur la population nous utilisons l'estimateur

$$\hat{Y} = \sum_{i \in R} w_i^C Y_i,$$

où R est l'échantillon des répondants et w_i^C sont les poids calés obtenus en combinant les poids initiaux w_i^{ini} , les probabilités de non-réponse estimées $\hat{\theta}_i$ et les g-weights du calage g_i :

$$w_i^C = \frac{g_i w_i^{ini}}{\hat{\theta}_i} = g_i w_i^{NR}.$$

La pondération avant calage est alors $w_i^{NR} = \frac{w_i^{ini}}{\hat{\theta}_i}$.

Ratio :

Pour estimer un ratio Y/X nous utilisons l'estimateur

$$\hat{Y}/\hat{X} = \frac{\sum_{i \in R} w_i^C Y_i}{\sum_{i \in R} w_i^C X_i}.$$

Moyenne :

Une moyenne est un cas particulier du ratio avec $X_i = 1$.

Remarque La moyenne peut aussi être vue comme un total divisé par la taille de la population N que nous connaissons d'après nos registres. Ces deux estimateurs sont équivalents au niveau suisse (puisque les poids sont calés sur N). De même les deux estimateurs sont aussi équivalents pour tous les domaines pris en compte dans le calage. Néanmoins, ce n'est pas le cas pour un domaine qui n'est pas pris en compte dans le calage. Dans ce cas, diviser par la somme des poids plutôt que par un "vrai" total fournit un estimateur ayant une variance plus faible. De plus, des moyennes peuvent

également être calculées pour des domaines dont la "vraie" taille est inconnue (domaines non basés sur les variables présentes dans les registres comme par exemple la formation).

Quantile :

Soit une variable Y_i avec $i \in U$. On définit premièrement $q_{p,Y}$, le quantile p de la variable Y . Soit R un échantillon de taille m de la population U . On définit $\hat{q}_{p,Y}$ comme l'estimateur pondéré (w^C) de $q_{p,Y}$ comme ci-dessous.

Soient $Y_{(1)}, Y_{(2)}, \dots, Y_{(m)}$ la variable Y ordonnée par ordre croissant sur R .

$$\hat{q}_{p,Y} = \begin{cases} Y_{(j)} & \text{Si } (\sum_{i \in R} w_i^C) p = \sum_{i \in R} \mathbf{1}_{\{Y_i \leq Y_{(j)}\}} w_i^C \\ \frac{Y_{(j)} + Y_{(j+1)}}{2} & \text{Si } (\sum_{i \in R} w_i^C) p \neq \sum_{i \in R} \mathbf{1}_{\{Y_i \leq Y_{(j)}\}} w_i^C \end{cases}$$

et j est tel que $\sum_{i \in R} \mathbf{1}_{\{Y_i \leq Y_{(j-1)}\}} w_i^C < (\sum_{i \in R} w_i^C) p \leq \sum_{i \in R} \mathbf{1}_{\{Y_i \leq Y_{(j)}\}} w_i^C$.

Variance

Total :

Nous voulons calculer la variance de l'estimateur pondéré $\hat{Y}$,

$$\text{Var}(\hat{Y}) = \text{Var}\left(\sum_{i \in R} w_i^C Y_i\right) = \text{Var}\left(\sum_{i \in R} \frac{g_i w_i^{ini}}{\hat{\theta}_i} Y_i\right).$$

Nous supposons que Y est explicable en partie par une régression linéaire des variables auxiliaires décrites en 3.2.1 que nous nommons ici $\mathbf{X}$,

$$Y_i = \beta \mathbf{X}_i + \epsilon_i, \quad i \in U.$$

Les erreurs ϵ_i sont alors les résidus obtenus en appliquant les moindres carrés sur toute la population. Nous désignons par $\hat{\epsilon}$ les résidus observés de la régression par moindres carrés (pondéré par $w_i^{NR} = \frac{w_i^{ini}}{\hat{\theta}_i}$) sur l'échantillon des répondants R ,

$$Y_i = \hat{\beta} \mathbf{X}_i + \hat{\epsilon}_i, \quad i \in R,$$

où $\hat{\beta}$ est l'estimateur pondéré de β . D'après [Deville et Särndal \(1992\)](#), nous avons alors

$$\text{Var}(\hat{Y}) \approx \text{Var}\left(\sum_{i \in R} w_i^{NR} \epsilon_i\right) = \text{Var}\left(\sum_{i \in R} \frac{w_i^{ini}}{\hat{\theta}_i} \epsilon_i\right)$$

que l'on peut approcher avec

$$\widehat{\text{Var}}\left(\sum_{i \in R} w_i^{NR} \hat{\epsilon}_i\right),$$

La variance est estimée de trois manières

1. La commande SAS proc surveymeans est documentée en référence [SAS Institute Inc. \(2016\)](#). Les poids utilisés sont w_i^{NR} et la variable sur laquelle s'applique le calcul est $\hat{\epsilon}$.

2. La formule de Deville : Soit $w_i^{NR} = \frac{w_i^{ini}}{\theta_i}$ les poids avant calage. On définit $\hat{\pi}_i^{NR} = 1/w_i^{NR}$ comme la probabilité d'inclusion d'ordre 1 dans l'échantillon R des répondants. Alors on a,

$$\begin{aligned}\widehat{\text{Var}}(\hat{Y}) &\approx \widehat{\text{Var}}\left(\sum_{i \in R} w_i^{NR} \hat{e}_i\right) \\ &= \sum_{i \in R} \frac{c_i}{\hat{\pi}_i^{NR}} (\hat{e}_i - \hat{e}_i^*)^2,\end{aligned}$$

où

$$\begin{aligned}\hat{e}_i^* &= \hat{\pi}_i^{NR} \frac{\sum_{j \in R} c_j \hat{e}_j / \hat{\pi}_j^{NR}}{\sum_{j \in R} c_j}, \\ c_j &= (1 - \hat{\pi}_j^{NR}) \frac{n}{n - 1}.\end{aligned}$$

Le développement de cette formule est décrit dans [Tillé \(2019\)](#).

3. La formule exacte : Pour un tirage de Poisson avec probabilité d'inclusion π_i , nous pouvons utiliser la formule exacte pour la variance d'un total. L'échantillon R est alors obtenu par un tirage de Poisson avec probabilité d'inclusion $\pi_i \theta_i$ où θ_i est la probabilité de réponse de l'unité i sachant qu'il est sélectionné. La variance peut alors être estimée par Horvitz-Thompson :

$$\widehat{\text{Var}}(\hat{Y}) \approx \widehat{\text{Var}}\left(\sum_{i \in R} w_i^{NR} \hat{e}_i\right) = \sum_{i \in R} \hat{e}_i^2 \frac{1 - \pi_i \hat{\theta}_i}{(\pi_i \hat{\theta}_i)^2},$$

car

$$\widehat{\text{Var}}\left(\sum_{i \in R} w_i^{NR} \hat{e}_i\right) \Big| \theta_i = \sum_{i \in R} \hat{e}_i^2 \frac{1 - \pi_i \theta_i}{(\pi_i \theta_i)^2}.$$

Nous disposons d'une pondération initiale dont l'inverse des poids w_i^{ini} est proche de la probabilité d'inclusion de l'unité i . En assimilant le tirage à un plan de Poisson de probabilité d'inclusion $\hat{\pi}_i = \frac{1}{w_i^{ini}}$, nous remplaçons π_i par $\hat{\pi}_i$ dans la formule.

Ratio :

Pour calculer la variance d'un ratio $\hat{Y}/\hat{X}$, nous linéarisons l'estimateur comme décrit dans [Tillé \(2019\)](#). La linéarisation d'un estimateur consiste à construire une variable $\text{lin}_{Y/X,i}$ pour chaque unité i telle que la variance du total de cette variable puisse estimer la variance de l'estimateur.

$$\begin{aligned}\text{Var}(\hat{Y}/\hat{X}) &\approx \text{Var}\left(\frac{1}{\hat{X}} \left(\hat{Y} - \frac{Y}{\hat{X}} \hat{X}\right)\right) \\ &= \text{Var}\left(\sum_{i \in R} \frac{1}{\hat{X}} \left(Y_i - \frac{Y}{\hat{X}} X_i\right) w_i^C\right)\end{aligned}$$

Nous posons $\text{lin}_{Y/X,i} = \frac{1}{\hat{X}} (Y_i - \frac{Y}{\hat{X}} X_i)$. Il faut à présent estimer la variance d'un total comme décrit ci-dessus pour $\text{lin}_{Y/X}$.

Moyenne :

Comme dit précédemment, une moyenne est un ratio $\hat{Y}/\hat{X}$ avec $X_i = 1$. Comme ci-dessus, nous pouvons calculer sa linéarisée puis estimer la variance du total de la linéarisée.

Quantile :

Pour calculer la variance d'un quantile $\hat{q}_{p,Y}$, nous linéarisons l'estimateur comme décrit dans Tillé (2019) ou Osier (2009) dont nous utilisons la macro en SAS.

$$\text{lin}_{\hat{q}_{p,Y},i} = \frac{p - \mathbf{1}_{\{Y_i \leq \hat{q}_{p,Y}\}}}{\hat{N} \hat{f}_Y(\hat{q}_{p,Y})},$$

où $\hat{N} = \sum_R w_i^C$ et $\hat{f}_Y$ est l'estimation de la densité lissée de Y obtenue par un lissage pondéré utilisant un noyau gaussien et une largeur de bande fixée par SNR (Simple Normal Reference, $h \propto \sqrt{\widehat{\text{Var}}(Y_i)/N^{1/5}}$) comme décrit dans le manuel de proc kde, SAS Institute Inc. (2013, p. 4087).

3.3.2 Résultats

Nous présentons enfin les résultats des variables clés pour les différentes années dans les tables 7, 8, 9 et 10. Les variances après calage présentées sont toujours obtenues par la formule de Deville qui fournit un estimateur très proche des deux autres approches décrites dans 3.3.1. Les variances avant calage sont elles aussi obtenues par la formule de Deville comme décrit dans Tillé (2019). Comme attendu, le gain sur la variance, entre avant le calage et après le calage, est environ de R^2 , le coefficient de détermination présenté en 3.2.3.

$$\text{Gain sur la variance} = \frac{\text{Var. avant calage} - \text{Var. après calage}}{\text{Var. avant calage}} \approx R^2$$

En effet, la table 7 montre que les gains, "gain var", sont souvent proches des R^2 de la table 5. La variance après calage est ainsi proche de la variance avant calage multipliée par un facteur $(1 - R^2)$. La table 8 présente le gain sur les publications de ratios, la table 9 présente le gain sur les quantiles, la table 10 présente les gains sur la moyenne des dépenses totales par type de ménage pour une année. La réduction de variance fluctue entre 8% et plus de 50% en fonction des indicateurs et des années. Cela se traduit par un gain sur la longueur de l'intervalle de confiance entre environ 4% et 30%.

Remarque

- La table 8 présente les gains sur des ratios. Nous précisons que les ratios correspondent à la part des dépenses (A50, A51, A57) cumulées sur toute la population et divisée par le cumul des revenus de la population. Cela ne représente pas la part moyenne du budget que les ménages dépensent.

$$\frac{\sum_{i \in R} Y_i w_i}{\sum_{i \in R} X_i w_i^C} \text{ est la part cumulée des dépenses (Y) divisés par le cumul des revenus (X).}$$
$$\frac{\sum_{i \in R} \frac{Y_i}{X_i} w_i}{\sum_{i \in R} w_i^C} \text{ est la part moyen du budget (X) que les ménages dépensent (Y).}$$

- Les tables 7, 8, 9 et 10 présentent des estimations de dépenses et de revenus. Pour les revenus, le nombre indiqué est positif alors que pour les dépenses, le nombre indiqué est négatif.

3.3.3 Évolution des estimations

Les évolutions des variables clés au fil des 3 années étudiée dans les tables 7, 8 et 9 sont présentée sur les figures 2, 3 et 4. Sur ces Figures, nous comparons l'évolution estimée avant et après calage

Table 7 Comparaison des estimateurs des moyennes avant et après calage. Les notations Est, Et et Var sont utilisées pour respectivement l'estimateur, son écart-type et sa variance. Les colonnes après calage présentent les caractéristiques de l'estimateur avec les poids calés w^C et les colonnes avant calage décrivent l'estimateur avec les poids de correction de non-réponse w^{NR} . La colonne gain var présente le gain sur la variance et ratio intervalle est le ratio des écart-types après et avant calage.

année	variable	Est. après calage	Et. après calage	Var. après calage	Est. avant calage	Et. avant calage	Var. avant calage	gain var	ratio intervalle
2016	A50m moyenne	-5049.22	44.63	1991.47	-5093.17	55.92	3126.93	0.36	0.80
	A51m moyenne	-612.10	5.85	34.27	-611.61	7.44	55.41	0.38	0.79
	A57m moyenne	-1429.15	14.24	202.88	-1432.01	15.55	241.86	0.16	0.92
	Bruttoeinkommen08 moyenne	9444.77	80.83	6534.80	9556.51	126.40	15978.84	0.59	0.63
2017	A50m moyenne	-5064.38	42.85	1836.31	-5091.93	55.12	3038.05	0.40	0.78
	A51m moyenne	-620.81	5.81	33.70	-621.68	7.26	52.72	0.36	0.80
	A57m moyenne	-1420.06	12.46	155.27	-1420.11	14.63	213.92	0.27	0.85
	Bruttoeinkommen08 moyenne	9406.91	88.27	7792.32	9403.08	126.74	16063.08	0.51	0.70
2018	A50m moyenne	-5072.29	55.40	3068.92	-5129.70	66.81	4464.22	0.31	0.83
	A51m moyenne	-613.74	6.21	38.60	-623.10	7.66	58.70	0.34	0.81
	A57m moyenne	-1424.66	12.71	161.43	-1421.05	14.72	216.81	0.26	0.86
	Bruttoeinkommen08 moyenne	9557.30	94.96	9016.88	9730.51	132.12	17456.09	0.48	0.72

Table 8 Comparaison des estimateurs des ratios avant et après calage. Les colonnes sont décrites plus précisément dans la table 7.

année	variable	Est. après calage	Et. après calage	var. après calage	Est. avant calage	Et. avant calage	var. avant calage	gain var	ratio intervalle
2016	A50m/Bruttoeinkommen08	-0.5346	0.0050	0.000025	-0.5330	0.0056	0.000032	0.21	0.89
	A51m/Bruttoeinkommen08	-0.0648	0.0008	0.000001	-0.0640	0.0009	0.000001	0.34	0.82
	A57m/Bruttoeinkommen08	-0.1513	0.0017	0.000003	-0.1498	0.0021	0.000004	0.32	0.82
2017	A50m/Bruttoeinkommen08	-0.5384	0.0053	0.000028	-0.5415	0.0057	0.000032	0.15	0.92
	A51m/Bruttoeinkommen08	-0.0660	0.0008	0.000001	-0.0661	0.0009	0.000001	0.25	0.86
	A57m/Bruttoeinkommen08	-0.1510	0.0017	0.000003	-0.1510	0.0020	0.000004	0.25	0.86
2018	A50m/Bruttoeinkommen08	-0.5307	0.0064	0.000041	-0.5272	0.0067	0.000045	0.10	0.95
	A51m/Bruttoeinkommen08	-0.0642	0.0008	0.000001	-0.0640	0.0009	0.000001	0.23	0.88
	A57m/Bruttoeinkommen08	-0.1491	0.0017	0.000003	-0.1460	0.0020	0.000004	0.30	0.84

Table 9 Comparaison des estimateurs des quantiles avant et après calage. Les colonnes sont décrites plus précisément dans la table 7.

année	variable	Est. après calage	Et. après calage	var. après calage	Est. avant calage	Et. avant calage	var. avant calage	gain var	ratio intervalle
2016	Bruttoeinkommen08 quantile 20	4527.26	94.69	8966.06	4671.29	121.88	14853.82	0.40	0.78
	Bruttoeinkommen08 quantile 40	6708.57	79.91	6385.98	6909.49	114.49	13109.05	0.51	0.70
	Bruttoeinkommen08 quantile 50	7888.14	92.00	8463.44	8166.56	131.70	17344.72	0.51	0.70
	Bruttoeinkommen08 quantile 60	9408.38	108.58	11789.68	9640.01	152.44	23238.48	0.49	0.71
	Bruttoeinkommen08 quantile 80	13090.92	131.62	17323.70	13363.97	180.66	32637.59	0.47	0.73
2017	Bruttoeinkommen08 quantile 20	4443.25	84.73	7179.10	4565.64	110.21	12145.69	0.41	0.77
	Bruttoeinkommen08 quantile 40	6765.12	77.64	6028.10	6910.02	108.17	11699.74	0.48	0.72
	Bruttoeinkommen08 quantile 50	7927.17	83.15	6913.60	8058.64	115.90	13433.79	0.49	0.72
	Bruttoeinkommen08 quantile 60	9347.71	99.53	9907.11	9500.28	137.68	18955.50	0.48	0.72
	Bruttoeinkommen08 quantile 80	12904.63	123.23	15185.29	12876.70	170.95	29223.49	0.48	0.72
2018	Bruttoeinkommen08 quantile 20	4615.00	74.18	5503.27	4780.17	99.74	9948.05	0.45	0.74
	Bruttoeinkommen08 quantile 40	6729.57	85.59	7326.42	7007.33	121.12	14670.94	0.50	0.71
	Bruttoeinkommen08 quantile 50	8014.99	88.53	7837.10	8264.37	125.72	15804.48	0.50	0.70
	Bruttoeinkommen08 quantile 60	9305.92	110.19	12141.35	9657.10	152.88	23371.54	0.48	0.72
	Bruttoeinkommen08 quantile 80	13482.81	127.02	16133.03	13642.89	172.73	29836.13	0.46	0.74

Table 10 Comparaison des estimateurs de la moyenne de A50m ventilé par type de ménage. Les colonnes sont décrites plus précisément dans la table 7.

année	variable	ventilation	niveau	Est. après calage	Et. après calage	var. après calage	Est. avant calage	Et. avant calage	var. avant calage	gain var	ratio intervalle
2018	Bruttoeinkommen08 moyenne	Familientyp08	110	7036.88	123.81	15329.15	7290.75	161.16	25971.36	0.41	0.77
			130	4399.57	132.86	17653.03	4428.46	138.59	19207.65	0.08	0.96
			210	11771.98	196.19	38489.59	12358.24	283.70	80485.25	0.52	0.69
			230	8472.12	219.45	48158.56	8278.37	234.54	55007.94	0.12	0.94
			300	13700.13	318.98	101747.67	13705.92	359.24	129051.70	0.21	0.89
			400	8358.35	307.21	94376.00	8215.23	368.85	136053.90	0.31	0.83
			500	11705.99	730.64	533832.57	12137.17	788.64	621958.35	0.14	0.93
			900	10700.36	513.69	263876.47	11462.14	685.25	469570.40	0.44	0.75

avec les intervalles de confiance. Le calage ajuste légèrement les résultats pour certains estimateurs, en réduisant notablement l'intervalle de confiance.

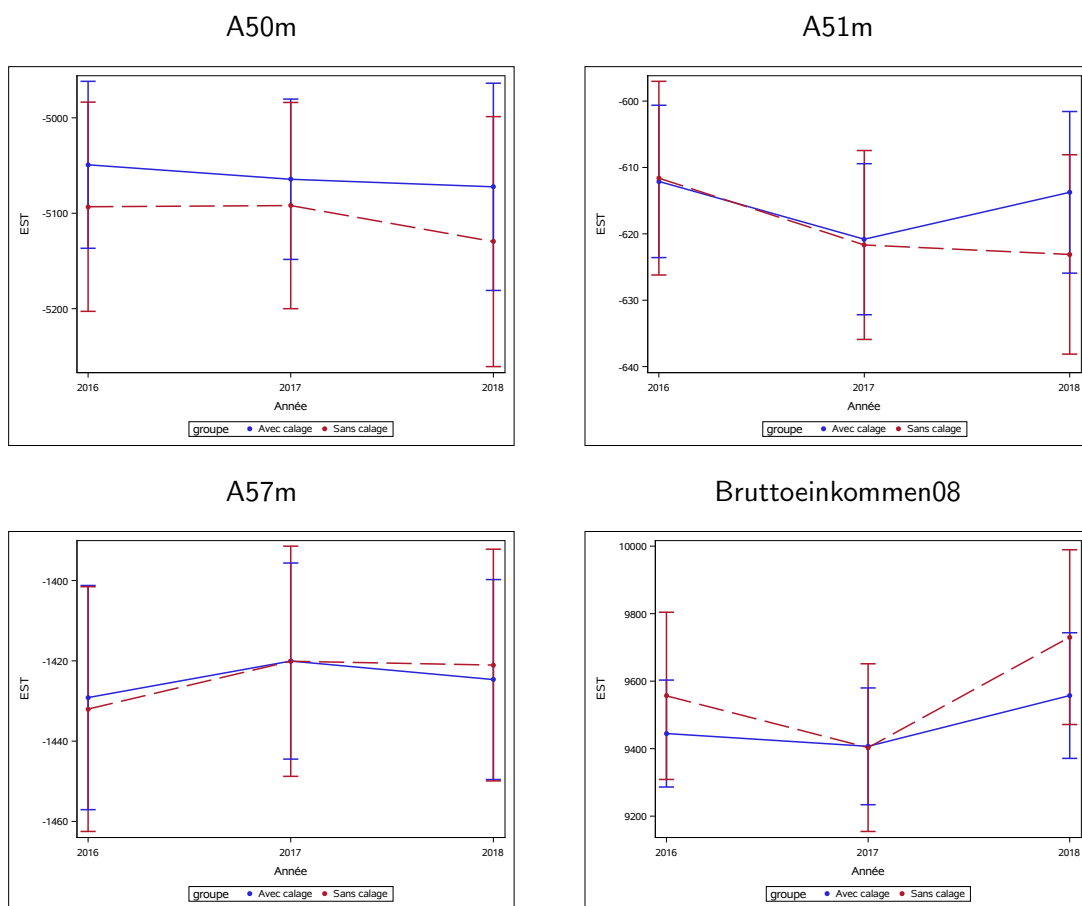


Figure 2 Evolution des moyennes des quatre variables clés estimées entre 2016 et 2018 avant et après calage.

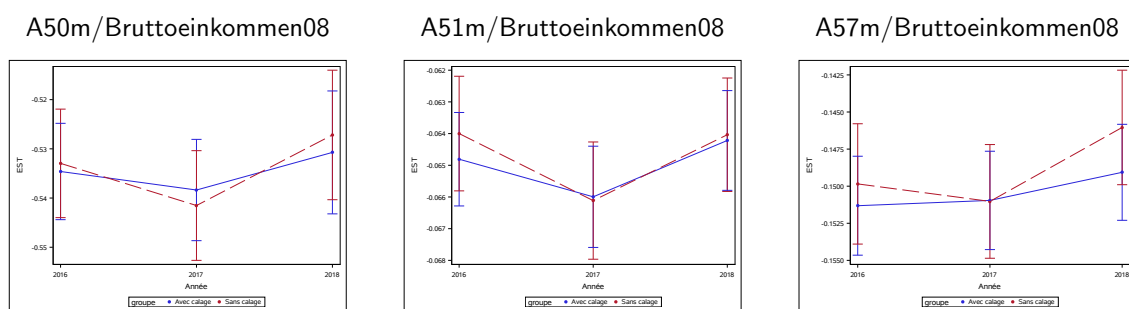


Figure 3 Evolution des parts du revenu qui passent dans les dépenses clés estimées entre 2016 et 2018 avant et après calage.

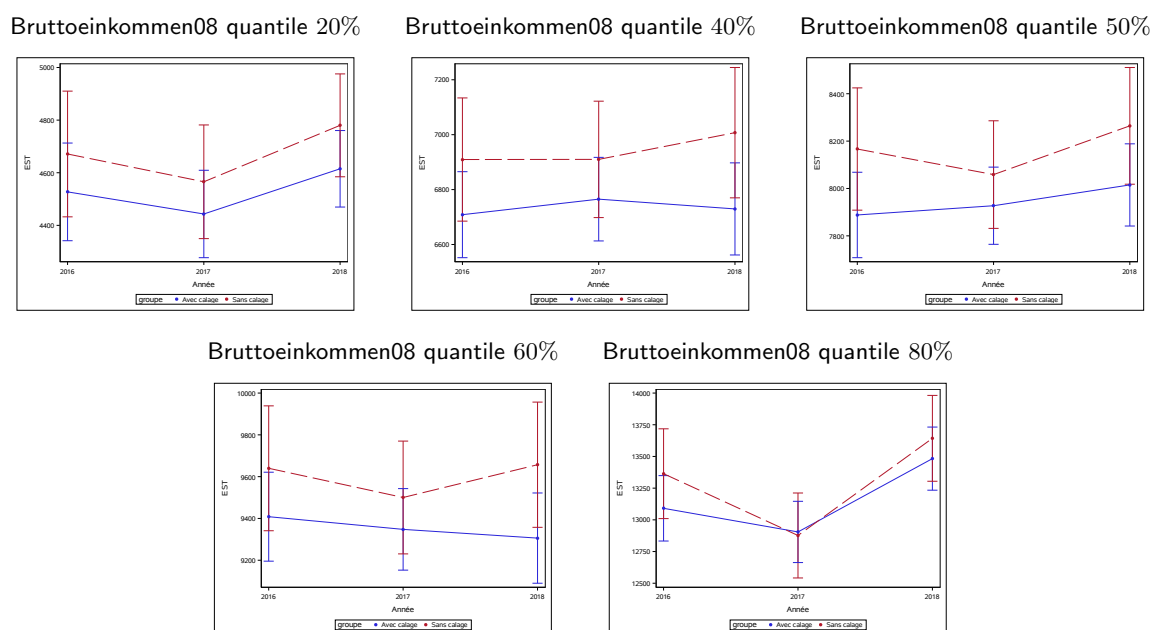


Figure 4 Evolution des quantiles de Bruttoeinkommen08 estimés entre 2016 et 2018 avant et après calage.

4 Conclusions

Dans ce travail, la section METH a développé une nouvelle pondération de l'EBM avec un nouveau traitement de la non-réponse et un nouveau calage. Afin d'exploiter au mieux cette nouvelle pondération, un calcul de variance adapté a été développé. Les traitements exploitent les informations auxiliaires disponibles pour ces travaux telles que les données de la CDC ou des variables de typologie de ménages. L'utilisation de ces nouvelles variables auxiliaires permet d'augmenter la qualité des estimateurs produits d'une part en réduisant le risque de biais et d'autre part en augmentant la précision des estimateurs.

- La réduction du risque de biais est obtenue premièrement par un traitement de non-réponse plus puissant qui capture mieux les mécanismes de non-réponse. Etant donné que le taux de réponse de l'enquête est d'environ 30%, cette modification peut être vue comme une amélioration importante. Dans un second temps, le risque de biais est encore réduit par le calage au moyen de variables corrélées avec les variables publiées.
- Les gains de précision obtenus par le nouveau calage sont non-négligeables. En fonction de l'indicateur publié et de l'année, le gain sur la variance varie entre 8% et plus de 50% (cf. table 10) ce qui équivaut à une précision que l'on obtiendrait avec un échantillon de taille d'environ 8% à 50% plus grand (pour un échantillon de 3000 unités, cela équivaut à des échantillons de 3240 à 4500 unités).

La section METH a dans ce travail uniquement travaillé sur le traitement de la non-réponse et le calage. Comme dans les autres enquêtes, le principal risque de biais réside dans le taux de réponse. Le travail effectué pour améliorer ce taux est déjà conséquent et des réflexions visant à encore l'améliorer sont en cours.

5 Références

- Anne Cornali Schweingruber, Ruedi Epple, Ueli Oetliker et Sylvie Rochat : Une nouvelle méthode pour l'enquête sur les revenus et la consommation : Nouveau modèle de pondération, résultats 2000-2003 et étude sur la prévoyance vieillesse. Rapport technique, Office fédéral de la statistique (OFS), Neuchâtel, 2007.
- Jean-Claude Deville et Carl-Erik Särndal : Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.
- Guillaume Osier : Variance estimation for complex indicators of poverty and inequality using lineari-
zation techniques. *Survey Research Methods*, 3:167–195, 2009.
- SAS Institute Inc. : *SAS/STAT 13.1 User's Guide, The KDE Procedure*. Cary, NC, 2013. URL <https://support.sas.com/documentation/onlinedoc/stat/131/kde.pdf>.
- SAS Institute Inc. : *SAS/STAT 14.1 User's Guide, The GLMSELECT Procedure*. Cary, NC, 2015. URL <https://support.sas.com/documentation/onlinedoc/stat/141/glmselect.pdf>.
- SAS Institute Inc. : *SAS/STAT 14.2 User's Guide, The SURVEYMEANS Procedure*. Cary, NC, 2016. URL <https://support.sas.com/documentation/onlinedoc/stat/142/surveymeans.pdf>.
- SAS Institute Inc. : *SAS/STAT 14.3 User's Guide, The LOGISTIC Procedure*. Cary, NC, 2017. URL <https://support.sas.com/documentation/onlinedoc/stat/143/logistic.pdf>.
- Yves Tillé : *Théorie des sondages : Échantillonnage et estimation en populations finies*. 2e Édition. Dunod, 2019.

A Appendice

Variables	Modalités Descriptifs
Familientyp08_SRP	110 Ménages d'une personne ayant strictement moins de 65 ans
Familientyp08_SRP	130 Ménages d'une personne ayant plus de 65 ans (inclus)
Familientyp08_SRP	210 Ménages de deux adultes (strictement moins que 65 ans) sans enfants
Familientyp08_SRP	230 Ménages de deux adultes (plus de 65 ans inclus) sans enfants
Familientyp08_SRP	300 Ménages avec deux adultes et des enfants (jusqu'à 25 ans inclus)
Familientyp08_SRP	400 Personnes seules avec enfants (jusqu'à 25 ans inclus)
Familientyp08_SRP	500 Ménages avec au moins 3 adultes et avec enfants (jusqu'à 25 ans inclus)
Familientyp08_SRP	900 Autres ménages
FamilientypAlter08_SRP	105 Ménages d'une personne ayant entre 65 et 74 ans (inclus)
FamilientypAlter08_SRP	106 Ménages d'une personne ayant plus de 75 ans (inclus)
FamilientypAlter08_SRP	205 Ménages de deux adultes ayant entre 65 et 74 ans (inclus)
Paartyp	0 Ménages sans couples mariés
Paartyp	1 Ménages avec couples mariés
Geschlechtertyp	0 Ménages mixtes avec des hommes et des femmes
Geschlechtertyp	1 Ménages avec uniquement des hommes
Geschlechtertyp	2 Ménages avec uniquement des femmes
Auslaendertyp	0 Ménages avec uniquement des suisses
Auslaendertyp	1 Ménage avec au moins un suisse et au moins un étranger
Auslaendertyp	2 Ménages avec uniquement des étrangers dont au moins un d'un pays voisin
Auslaendertyp	3 Ménages avec uniquement des étrangers de pays non voisins
Strate01	1 GE, VD, VS
Strate01	2 BE, FR, JU, NE, SO
Strate01	3 AG,BS, BL
Strate01	4 ZH
Strate01	5 AR, AI, GL, GR, SG, SH, TG
Strate01	6 LU, NW, OW, SZ, UR, ZG
Strate01	7 TI
indice_rente	0 Aucune rente dans le ménage
indice_rente	1 Indique une rente dans le ménage
indice_ind	0 Pas de revenu indépendant
indice_ind	1 Le ménage perçoit un revenu indépendant
mitTelefonnummer	0 Aucun numéro de téléphone fixe du ménage est connu
mitTelefonnummer	1 Un numéro de téléphone fixe du ménage est connu
mAnzahlWohnungen15_1_2	1 1 appartements
mAnzahlWohnungen15_1_2	2 2 appartements
mAnzahlWohnungen15_1_2	3 autre
indice_nombre_rev	0 Ménages sans revenu
indice_nombre_rev	1 Ménages percevant un revenu
indice_nombre_rev	2 Ménages percevant deux revenus
indice_nombre_rev	3 Ménages percevant 3 revenus ou plus
cdc_classe	0 Ménages dont le total des revenus est entre le quantile 0 et 10%
cdc_classe	1 Ménages dont le total des revenus est entre le quantile 10% et 20%
cdc_classe	2 Ménages dont le total des revenus est entre le quantile 20% et 30%
cdc_classe	3 Ménages dont le total des revenus est entre le quantile 30% et 40%
cdc_classe	4 Ménages dont le total des revenus est entre le quantile 40% et 50%
cdc_classe	5 Ménages dont le total des revenus est entre le quantile 50% et 60%
cdc_classe	6 Ménages dont le total des revenus est entre le quantile 60% et 70%
cdc_classe	7 Ménages dont le total des revenus est entre le quantile 70% et 80%
cdc_classe	8 Ménages dont le total des revenus est entre le quantile 80% et 90%
cdc_classe	9 Ménages dont le total des revenus est entre le quantile 90% et 100%
cumule_cdc_rev_trunc_099	Revenu cdc sommé sur le ménage et tronqué à 99%
indiceap8	Indicateur de la présence d'une personne plus vieille que 80 ans dans le ménage
nbre_homme	Nombre d'hommes dans le ménage
nbre_suisse	Nombre de suisses dans le ménage
nbre_age3_	0 Nombre de personnes entre 0 et 10 ans exclus
nbre_age3_	1 Nombre de personnes entre 10 et 20 ans exclus
nbre_age3_	2 Nombre de personnes entre 20 et 30 ans exclus
nbre_age3_	3 Nombre de personnes entre 30 et 40 ans exclus
nbre_age3_	4 Nombre de personnes entre 40 et 50 ans exclus
nbre_age3_	5 Nombre de personnes entre 50 et 60 ans exclus
nbre_age3_	6 Nombre de personnes entre 60 et 70 ans exclus
nbre_age3_	7 Nombre de personnes entre 70 et 80 ans exclus
nbre_age3_	8 Nombre de personnes au-delà de 80 ans inclus

B Appendice

	Split 1	Split 2	Split 3	Split 4	Split 5	Split 6	Split 7	Split 8	Split 9	Split 10	Split 11
1	intTéléphonnum0-q (7432)	indEca-g0 (6831)	Au1sander typ3-q (6442)	Pratryp0-q (1003)	Geductive typ2-q (1179)	Strat02-1-q (1440)	indEca_ind0-q (1320)	indEca_class0-q (1138)	Famintyp08_SPH12-0-q (1044)	indEca_rbre_rev0-q (892)	indEca_class0-q (750)
2	13.86%	35.60%	36.94%	32.68%	30.18%	28.47%	25.55%	30.06%	31.61%	32.96%	18.99%
3											indEca_rbre_rev0-q (132)
4											23.00%
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											

Figure 5 Arbre résultant de la segmentation appliquée à l'année 2017 créant 32 groupes homogènes de réponse, RHG. Le nombre d'unités dans chaque groupe est indiqué entre parenthèses et un pourcentage indique le taux de réponse dans chaque groupe.