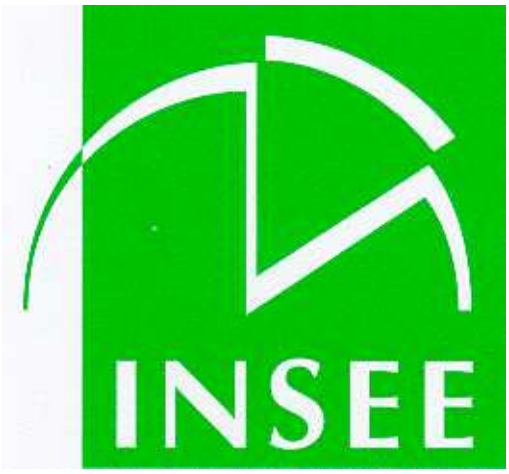
Direction des Statistiques Démographiques et Sociales

N° F1503

PRÉCISION DE L'ENQUÊTE PATRIMOINE 2010

Pierre LAMARCHE - Laurianne SALEMBIER

Document de travail



Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des Documents de Travail
de la
DIRECTION DES STATISTIQUES DÉMOGRAPHIQUES ET SOCIALES

N°F1503

PRÉCISION DE L'ENQUÊTE PATRIMOINE 2010

AUTEURS : PIERRE LAMARCHE et LAURIANNE SALEMBIER

(DIVISION REVENUS ET PATRIMOINE DES MÉNAGES*)

Document de travail

mai 2015

(*au moment de la rédaction)

Remerciements:

Nous tenons à remercier Guillaume Chauvet pour ses conseils et ses remarques extrêmement précieux, ainsi qu'Eric Lesage, David Haziza et les autres membres du laboratoire du CREST-LSE pour le temps qu'ils ont bien voulu nous accorder lors de la réunion à l'Ensai en août 2012. Toute notre gratitude va également bien entendu à la division Sondages, et en particulier Emmanuel Gros, ainsi que les membres de la division Revenus et Patrimoine des Ménages, Simon Beck, Rosalinda Coppoletta et Kim-Hoa Luu Kim pour leurs commentaires et leurs remarques très utiles, ainsi que leur soutien précieux. Par ailleurs, une partie des calculs réalisés dans ce document ont été rendus possibles par les travaux réalisés sur l'enquête Patrimoine par Karim Moussallam au sein de la division Sondage (que nous désignons dans l'ouvrage par son unité de rattachement, l'Unité des Méthodes Statistiques à l'époque où ces travaux ont été réalisés). Nous restons bien évidemment seuls responsables des erreurs et imprécisions restantes.

PRÉCISION DE L'ENQUÊTE PATRIMOINE 2010

Résumé

Ce document de travail présente les résultats en termes de précision pour les principaux indicateurs calculés sur l'enquête Patrimoine 2010. Cette enquête a bénéficié d'un certain nombre d'innovations méthodologiques qui ont permis de mieux capter la forte concentration du patrimoine dans le haut de la distribution. Ce travail offre une évaluation de ces innovations en termes de gain de précision et analyse les écarts observés entre 2004 et 2010. Par ailleurs, la précision de l'enquête a été calculée de manière analytique, par une décomposition des sources d'imprécision due aux différents degrés de sondage. Un calcul par poids répliqués a également été réalisé, ce qui permet de mettre à disposition des utilisateurs de l'enquête l'information liée à la précision des estimateurs. Ce calcul est validé empiriquement en comparant les résultats obtenus par le calcul analytique et ceux donnés par les poids répliqués. Enfin, l'information contenue dans les poids répliqués est combinée avec les imputations multiples réalisées dans le cadre de la production de l'enquête afin d'obtenir une estimation de l'incertitude globale.

Mots-clés : Bootstrap, poids répliqués, variance, plan de sondage, patrimoine.

Abstract

In this paper we show results for the precision for the main indicators estimated on the French Wealth Survey 2010 ("Enquête Patrimoine 2010"). The 2010 wave of this survey benefited from several significant innovations that aimed at improving the measurement of high concentrated wealth at the top of the distribution. These innovations are therefore assessed with respect to the gain of precision and the statistical significance of the gaps between waves 2004 and 2010 of the French Wealth Survey is evaluated. Moreover this work exhibits two different methods for estimating the precision of the survey: the first one is obtained thanks to an analytical approach involving confidential information. The second one is given thanks to replicate weights which may be disseminated among users. The results with replicate weights are compared with the ones given by the analytical method, thereby validating the bootstrap method used for computing the replicate weights. Finally, information provided through replicate weights in terms of variance due to sampling is combined with multiple imputations in order to evaluate the global uncertainty, similarly to what is done on the Household Finance and Consumption Survey, the European counterpart of the French Wealth Survey.

Keywords: Bootstrap, replicate weights, variance, sampling design, wealth.

Table des matières

1 Motivation

Le but de ce document de travail est multiple. Tout d'abord, il vise à présenter de manière synthétique les choix méthodologiques qui ont guidé la constitution de l'échantillon de la dernière enquête Patrimoine dont la collecte s'est déroulée entre la fin de l'année 2009 et le début de l'année 2010. Ces choix ont été guidés par un certain nombre d'innovations motivées à la fois par les expériences à l'étranger et les résultats des précédentes vagues de l'enquête française. Il a été en particulier possible de sur-représenter les ménages appartenant au haut de la distribution de patrimoine, ce qui doit théoriquement mener à une amélioration de la précision de la mesure. Dans un second temps, nous présentons les principaux résultats en matière de précision obtenus pour cette enquête. Nous nous concentrons en particulier sur les indicateurs les plus usités dans l'enquête et cherchons à la fois à déterminer leur précision et d'évaluer l'impact des modifications méthodologiques sur cette précision. Ces résultats sont établis à l'aide d'une méthode analytique basée sur la décomposition de la variance entre les différentes étapes du sondage. L'implémentation de cette méthode nécessite que l'on ait sa disposition les données relatives au plan de sondage, qui en général ne sont pas diffusées pour des raisons de confidentialité. En effet, fournir une telle information rendrait potentiellement identifiables les ménages ayant répondu à l'enquête. Dans un troisième temps, nous développons une méthode de bootstrap *ad hoc* qui doit permettre l'évaluation de la précision de l'enquête à l'aide de poids dit répliqués. Cette méthode permet ainsi de mettre à disposition de l'utilisateur les données de l'enquête relative à la précision sans toutefois dévoiler d'informations confidentielles. Dans un dernier temps, nous utilisons ces poids répliqués ainsi que différents jeux d'imputation des mêmes observations pour traiter de manière globale la question de l'incertitude de la mesure, qui recouvre ainsi non seulement l'imprécision due à l'échantillonnage et à la non-réponse totale, mais également l'incertitude liée à la non-réponse partielle.

2 Plan de sondage

L'échantillon de l'enquête Patrimoine a été tiré dans les fichiers de la Taxe d'Habitation ainsi que d'autres fichiers fiscaux pour les ménages vivant en France métropolitaine et à la Réunion, ainsi que pour les ménages « hauts patrimoines » des Antilles, et dans les Enquêtes Annuelles de Recensement pour les autres ménages habitant dans les Antilles.

2.1 Tirage dans la Taxe d'Habitation

2.1.1 Tirage de 1er degré

Le tirage réalisé pour la partie France métropolitaine et Réunion est un tirage à deux degrés. Le premier degré concerne la sélection des Zones d'Action Enquêteur (ZAE), qui est commun à l'ensemble des enquêtes ménages de l'Insee. Ce tirage est un tirage proportionnel à la taille, stratifié par région [1]. Certaines ZAE ont une probabilité de tirage supérieure ou égale à 1, ce qui est le cas de 37 d'entre elles (toutes sont évidemment de grandes agglomérations regroupant un grand nombre de logements principaux) ; elles sont alors automatiquement sélectionnées et le tirage n'est réalisé que sur les unités primaires ayant une probabilité de tirage strictement inférieure à 1. Par ailleurs, le tirage des ZAE est équilibré sur un certain nombre de totaux régionaux :

- nombre de résidences principales de la ZAE
- revenu fiscal total de la ZAE

1. sauf pour l'Ile-de-France, pour laquelle une stratification petite couronne/grande couronne a été retenue

- en Ile-de-France : âge de la personne de référence, type de ménage, habitat collectif ou individuel, statut d'occupation, nombre d'étrangers...

Au total, 525 ZAE ont été sélectionnées en France métropolitaine, ce qui représente 11,6 millions de ménages sélectionnables au moment du second degré de tirage. La Réunion a été considérée comme une seule et unique ZAE.

2.1.2 Tirage de 2nd degré

Pour le second degré du tirage, la population a été séparée en deux strates distinctes : les « hauts patrimoines » [2] (pour laquelle le tirage sur les fichiers de la Taxe d'Habitation est réalisé sur l'ensemble du territoire, c'est-à-dire métropole, Réunion et Antilles), et les ménages dits « standards ». Au sein de chacune des strates et pour chaque ZAE, un tirage systématique a été réalisé après allocation par ZAE selon la taille de l'échantillon définie pour chaque strate. Pour ce qui concerne l'échantillon dit « standard », 6 strates ont été définies en France métropolitaine :

- la strate Agriculteurs : ménages dont au moins un des membres est exploitant agricole.
- la strate Hauts indépendants : ménages dont la personne de référence a 65 ans ou moins, déclarant au moins 5 000 euros de revenus annuels d'indépendants non agricoles et n'appartenant pas à la strate Agriculteurs.
- la strate Cadres : ménages dont la personne de référence a 65 ans ou moins, déclarant soit des gains de levées d'option, soit de hauts salaires (plus de 36 000 euros de salaires annuels pour une personne seule, 72 000 pour un couple) et n'appartenant pas aux strates précédentes.
- la strate Revenus du patrimoine : ménages dont la personne de référence a 65 ans ou moins, dont un type de revenus financiers déclaré est conséquent, à savoir plus de 1 000 euros annuels de revenus de valeurs mobilières (soumises ou non à prélèvement libératoire) et n'appartenant pas aux strates précédentes.
- la strate Âgés : ménages dont la personne de référence a 60 ans ou plus, et qui ne relèvent pas des strates précédentes.
- la strate Reste : tous les autres ménages.

À la Réunion, les agriculteurs réunionnais ont été reversés dans la strate « Reste ». Pour les autres ménages, les strates définies pour la France métropolitaine ont été conservées.

Pour la partie « hauts patrimoines », 4 strates, communes à la France entière, ont été définies de la façon suivante :

- la strate Riches Urbains : ménages résidants dans l'une des 37 ZAE exhaustives et dont le patrimoine brut est supérieur à 3 millions d'euros.
- la strate Patrimoine à dominante mobilière : ménages dont le patrimoine brut est supérieur à 1,8 million d'euros et dont le patrimoine mobilier est à la fois strictement supérieur à leur patrimoine immobilier et supérieur à 200 000 euros.
- la strate Patrimoine à dominante immobilière : ménages dont le patrimoine brut est supérieur à 1,1 million d'euros et dont le patrimoine immobilier est à la fois strictement supérieur à leur patrimoine mobilier et supérieur à 200 000 euros.
- la strate Patrimoine plus faible : tous les autres ménages de la strate « hauts patrimoines ».

2.2 Tirage dans les Enquêtes Annuelles de Recensement

Pour la Guadeloupe, 1 500 ménages ont été sélectionnés dans les Enquêtes Annuelles de Recensement. Il s'agit d'un tirage systématique, pour assurer la représentativité des micro-régions

2. Pour plus de précision sur la construction et la motivation de cette stratification, se reporter à la partie 2.3.

Strates		France métropolitaine		Réunion	
		Nombre	Taux de sondage	Nombre	Taux de sondage
Échantillon standard	Agriculteurs	1 150	4,48 %	-	-
	Hauts indépendants	1 141	0,36 %	153	1,40 %
	Cadres	796	0,17 %	107	0,93 %
	Revenus du patrimoine	1 310	0,24 %	113	0,93 %
	Âgés	4 762	0,18 %	416	0,70 %
	Reste	5 071	0,04 %	711	0,47 %

Table 1 – Nombre de ménages échantillonnés et taux de sondage pour la strate « standard »

Strates		France entière	
		Nombre	Taux de sondage
Échantillon	Riches urbains	497	2,37 %
« Hauts Patrimoines »	Patrimoine à dominante mobilière	916	0,89 %
	Patrimoine à dominante immobilière	921	0,89 %
	Patrimoine plus faible	902	0,29 %

Table 2 – Nombre de ménages échantillonnés et taux de sondage pour la strate « hauts patrimoine »

guadeloupéennes. La stratification adoptée pour cet échantillonnage s'inspire de celle adoptée pour l'échantillon « standard » de la métropole et de la Réunion :

- la strate Âgés : ménage dont la personne de référence a 60 ans ou plus
- la strate Indépendants Actifs : ménages dont la personne de référence a moins de 60 ans et est un indépendant actif
- la strate Cadres : ménages dont la personne de référence a moins de 60 ans et est cadre
- la strate Reste : tous les autres ménages.

L'échantillon martiniquais est d'une taille beaucoup plus modeste : 80 ménages ont été sélectionnés pour cette partie du territoire français. La stratification adoptée est alors beaucoup plus frustre, puisque seules deux strates ont été retenues pour la constitution de l'échantillon :

- la strate Cadres, non salariés et retraités de ces professions
- la strate Reste, c'est-à-dire tous les autres ménages

Strates		Nombre	Taux de sondage
Guadeloupe	Âgés	486	0,93 %
	Indépendants actifs	241	2,36 %
	Cadres	145	2,66 %
	Reste	628	0,76 %
Martinique	Cadres, non salariés et retraités de ces professions	27	0,15 %
	Reste	53	0,04 %

Table 3 – Nombre de ménages échantillonnés et taux de sondage pour les Antilles

2.3 Pourquoi sur-représenter les hauts patrimoines ?

La sur-représentation des hauts patrimoines dans l'enquête fait partie des nouveautés méthodologiques introduites en 2009. L'idée de la sur-représentation de ménages les mieux dotés en patrimoine repose sur le constat de l'extrême concentration du patrimoine dans le haut de la distribution.

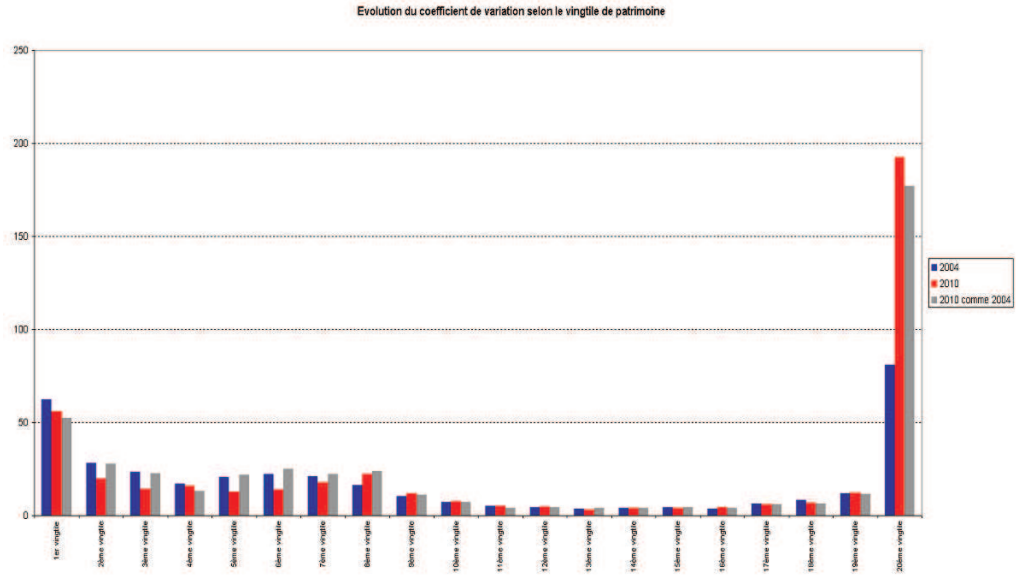


FIGURE 1 – Coefficient de variation du patrimoine selon le vingtile de patrimoine brut, en 2004 et en 2010

La sur-représentation des hauts patrimoines est une méthodologie qui a par ailleurs fait ses preuves, en particulier aux États-Unis, pour l'enquête américaine SCF [3]. La figure 1 montre explicitement que les parties les plus hétérogènes se situe dans le haut de la distribution ; en effet, le haut de la distribution présente déjà naturellement une variance élevée, par un simple effet d'échelle. Une fois que l'on contrôle de cet effet d'échelle, grâce au calcul d'un coefficient de variation, les 5 % des ménages les mieux dotés en patrimoine présentent toujours une hétérogénéité très élevée, révélatrice de la très forte concentration du patrimoine. L'observation faite en 2004 de la distribution du patrimoine au sein de la population des ménages montrait déjà une forte concentration du patrimoine ; la figure 1 semble suggérer un accroissement de cette concentration, et partant un accroissement des inégalités de patrimoine. Pour autant, il n'est pas complètement possible de conclure, puisque la mesure de l'hétérogénéité du patrimoine dans le haut de la distribution a été substantiellement améliorée grâce à la sur-représentation des hauts patrimoines ainsi que les différentes améliorations du processus d'estimation des montants. Il est possible de neutraliser l'effet dû à l'amélioration de l'estimation des montants en estimant des montants pour 2010 avec une méthodologie comparable à celle employée en 2004. C'est ce que représentent les barres grises du graphique 1. On constate ainsi qu'une fois que l'on tient compte de l'effet de cette modification, la mesure de la concentration du patrimoine dans le haut de la distribution reste très élevé. Cependant, cet accroissement peut être dû soit à un accroissement

3. Pour plus de précision sur la sur-représentation des hauts patrimoines dans l'enquête américaine, on pourra se reporter au lien suivant `http://www.federalreserve.gov/econresdata/scf/files/isi2007.pdf`

de la concentration, soit à une meilleure mesure consécutive à la sur-représentation des hauts patrimoines. Il faut par conséquent réaliser un calcul de précision pour distinguer ce qui rélève de l'amélioration de la mesure, et ce qui rélève d'un réel accroissement des inégalités.

3 Non-réponse

3.1 Qui sont les répondants ?

Les ménages considérés comme répondants à l'enquête Patrimoine 2010 sont les ménages ayant accepté de répondre au questionnaire et ayant répondu à un nombre minimal de questions : on estime que les ménages ayant répondu *a minima* au Tronc Commun des Ménages et à la partie de recensement des actifs sont des ménages répondants. Il aurait été possible de réaliser le calage et le traitement de la non-réponse en même temps. L'autre option, retenue ici, était de réaliser d'abord le traitement de la non-réponse, puis de caler dans un second temps. Ce choix a été fait en particulier parce que la base de sondage contient un grand nombre de variables qui permettent de construire des modèles satisfaisants pour expliquer la non-réponse.

On considère que la non-réponse est le résultat d'un processus poissonnien, pour lequel la probabilité de répondre varie en fonction des caractéristiques observables des ménages. La méthode des scores a été retenue pour corriger la non-réponse : dans un premier temps, on modélise la probabilité de répondre grâce à un modèle logit, puis on construit des groupes homogènes de réponse. Dans chacun de ces groupes ainsi constitués, on multiplie alors le poids des répondants par l'inverse du taux de réponse du groupe.

De manière globale, le taux de réponse se situe à 69 % des ménages sélectionnés. Cependant, ce taux de réponse peut varier de manière assez significative d'une strate à l'autre, comme le montrent les chiffres du tableau 4.

Strate	Taux de réponse
Agriculteurs	79 %
Hauts indépendants	65 %
Cadres	63 %
Revenus du patrimoine	71 %
Âgés	69 %
Reste	74 %
Riches urbains	46 %
Patrimoine à dominante mobilière	55 %
Patrimoine à dominante immobilière	60 %
Patrimoine plus faible	65 %

TABLE 4 – Taux de réponse en fonction de la strate d'appartenance des ménages

Les comportements de réponse varient également en fonction de l'âge des personnes interrogées. Le taux de réponse se situe globalement autour de 70 % pour les ménages de moins de 75 ans, il diminue ensuite pour atteindre 61 % pour les ménages plus âgés. Le taux de réponse peut également varier en fonction de la composition du ménage : les personnes seules sont les plus susceptibles de ne pas répondre à l'enquête, avec un taux de réponse de l'ordre de 63 %.

Par ailleurs, il existe des différences selon le type de territoire dans lequel vivent les ménages sélectionnés. Ainsi, les zones rurales présentent des taux de réponse importants, de l'ordre de 72 %. À l'inverse, l'unité urbaine de Paris est la zone dans laquelle il est le plus difficile d'obtenir

des réponses : le taux de réponse y est de 57 %. Enfin, on constate une forte hétérogénéité dans le comportement de réponse selon le type de logement dans lequel vivent les ménages. Comme le montre le tableau 5, les taux de réponse sont plus faibles pour les ménages vivant dans des immeubles de grande taille ; l'accès à ces immeubles est souvent plus difficile, et cette difficulté peut expliquer, au moins en partie, les différences constatées selon la taille d'unité urbaine.

Strates	Taux de réponse
Une ferme, un pavillon ou une maison indépendante	72 %
Une maison de ville mitoyenne, jumelée...	70 %
Un appartement dans un immeuble de deux logements	65 %
Un appartement dans un immeuble de trois à neuf logements	67 %
Un appartement dans un immeuble de 10 logements ou plus	62 %
Une habitation précaire	83 %
Un autre type de logement	58 %

TABLE 5 – Taux de réponse en fonction du type de logement occupé par le ménage

3.2 Modélisations de la non-réponse

Les tableaux 14 et 15 en annexe A montrent les résultats obtenus pour expliquer le mécanisme de non-réponse dans les strates « standard » et « hauts patrimoines ». Ces résultats apportent un certain nombre d'éclairage sur les déterminants de l'obtention de la réponse de la part des ménages, mais ils présentent cependant certaines limites. En effet, les caractéristiques disponibles dans le fichier concernent les ménages au moment de la constitution de la base de sondage. Ces derniers peuvent avoir déménagé, et dans ce cas de figure, le ménage les ayant remplacés est interrogé. Ceci étant dit, les données employées pour l'échantillonnage sont suffisamment récentes pour garantir que ce phénomène reste relativement mineur. En ce sens, les variables doivent être vues comme des variables proxy. Cependant, les résultats semblent indiquer que l'âge et la composition du ménage sont des éléments déterminants qui expliquent le phénomène de non-réponse dans l'enquête. En particulier, les personnes seules sont, à caractéristiques identiques, moins susceptibles d'avoir répondu à l'enquête : en effet, la présence de plusieurs personnes dans le ménage augmente les possibilités de contact pour l'enquêteur. Enfin, la variable renseignant la région de gestion de l'enquêté est également pertinente pour expliquer le mécanisme de non-réponse à l'enquête.

3.2.1 Résultats propres à l'échantillon standard

Par ailleurs, pour l'échantillon standard, les spécificités du protocole de collecte peuvent expliquer en partie les écarts de probabilité de réponse à l'enquête. La taille d'unité urbaine peut également constituer un élément explicatif du processus de réponse, pour les raisons invoquées ci-dessus. Cependant, il n'est pas possible d'isoler un effet uniforme de la variable, bien que l'hypothèse de nullité globale des coefficients soient clairement rejetée. Par ailleurs, à taille d'unité urbaine donnée, le type de logement (maison ou appartement) ne semble pas avoir d'effet sur le comportement de réponse, ce qui invalide l'hypothèse formulée précédemment selon laquelle le type de logement pouvait avoir un impact sur l'accessibilité du ménage par l'enquêteur. En revanche, et ce à revenu constant,la taille du logement influe positivement sur la probabilité de réponse du ménage, captant ainsi probablement en partie l'effet attendu.

La strate de tirage a un effet significatif sur la probabilité de répondre à l'enquête, ce qui reflète une hétérogénéité entre les ménages dont permet de tenir compte la stratification. Ainsi, par rapport aux ménages appartenant à la strate dite « Reste », et qui peut s'interpréter comme la strate regroupant les ménages ne présentant pas de caractéristique particulière du point de vue du patrimoine, les agriculteurs ont une probabilité plus élevée de répondre à l'enquête, quand les cadres et les indépendants sont pour leur part moins susceptibles de répondre. Ce résultat est cohérent avec les statistiques données par le tableau 4.

Enfin, les variables de revenu peuvent expliquer également la non-réponse, sans qu'il soit possible de tirer des conclusions claires de la modélisation. En effet, du côté des revenus d'activité, les effets ne sont pas monotones : ainsi, les ménages appartenant au milieu de la distribution des revenus d'activité ont une probabilité plus forte de répondre à l'enquête que les ménages les moins aisés. En revanche, les ménages appartenant aux 20 % les mieux dotés n'ont pas une probabilité statistiquement plus grande de répondre par rapport à leurs homologues du bas de la distribution.

3.2.2 Résultats propres à l'échantillon « hauts patrimoines »

Du côté des ménages « hauts patrimoines », le type de logement a un effet significatif sur le processus de réponse : les ménages vivant dans une maison sont plus accessibles que les ménages vivant dans des immeubles. En revanche, si les revenus d'activité ont un effet positif sur la probabilité de répondre à l'enquête, les variables de patrimoine ne semblent pas significativement influer les comportements de réponse au sein de la strate « Hauts Patrimoines ». Ceci peut s'expliquer par le fait que la stratification est très fortement corrélée au niveau de patrimoine brut, par ailleurs présent comme variable explicative dans le modèle.

3.3 Méthode de repondération

La méthode de repondération se base sur les travaux de Beaumont et Haziza (2007). Il s'agit d'un développement de la méthode des scores, qui se base donc sur l'estimation de la probabilité de réponse grâce à un modèle logit ou probit. La méthode est un arbitrage entre biais et variance ; en effet, l'estimation de probabilité de réponse individuelle, à condition que le modèle soit correctement spécifié et que l'ensemble des variables explicatives de la non-réponse soient incluses, doit théoriquement permettre une repondération la plus précise possible, et qui par conséquent permet le plus de réduire le biais dû à la non-réponse. Cependant, une telle méthode a le désavantage d'augmenter la dispersion des poids finaux, et par conséquent, d'accroître de manière substantielle l'imprécision liée à l'estimation. Par ailleurs, il n'est pas évident que l'hypothèse de correcte spécification du modèle soit vérifiée, en particulier s'agissant de probabilités. L'idée de la méthode des scores est donc de créer des groupes homogènes en termes de non-réponse, pour lesquels on applique une repondération uniforme. Si cette méthode réduit la dispersion des poids et permet de s'abstenir de faire des hypothèses trop contraignantes sur la forme fonctionnelle des probabilités que l'on souhaite modéliser, elle introduit cependant du biais. Tout l'enjeu de cette méthode est donc de construire les classes de repondération les plus pertinentes possibles afin de réduire la variance sans introduire trop de biais.

La méthode proposée par Beaumont et Haziza (2007) consiste en l'estimation d'une probabilité de réponse π_i^R pour chaque ménage i par un modèle probit, puis de créer des classes de réponse grâce à une méthode d'analyse de données. Pour l'enquête Patrimoine 2010, les classes de réponse ont été constituées grâce à un algorithme de centres mobiles (procédure FASTCLUS en SAS) qui tourne jusqu'à stabilisation des classes. Le nombre de classes est quant à lui déterminé par un arbitrage entre nombre de classes et écart avec la probabilité individuelle de réponse π_i^R :

pour chaque échantillon (standard ou « hauts patrimoines »), on sélectionne le plus petit nombre de classes entre 2 et 20 qui permet d'avoir le moins d'écart possible (R^2 supérieur à 0,99 pour une régression de π_i^R sur les classes de réponse).

La méthode permet effectivement de réduire la variance. Le calcul est possible en estimant l'ajout de variance propre à la phase de non-réponse grâce à la formule (présentée dans la section 5) :

$$\hat{\mathbb{V}}(\hat{Y}) = \sum_{i \in s} (1 - \pi_i^R) \left(\frac{Y_i}{p_i} \right)^2$$

avec p_i la pondération finale du ménage i. On montre grâce à cette formule que la méthode des scores a permis de réduire la variance liée à la non-réponse de 35 % et, par conséquent, la variance totale d'environ 17 % par rapport au résultat obtenu directement par les probabilités estimées.

4 Calage

Les poids de l'enquête Patrimoine 2010 ont également été calés à l'aide de la macro Calmar2. Les poids sont calés sur les champs suivants : France métropolitaine, Réunion et Antilles. Pour la France métropolitaine, les variables de calage sont les suivants :

- nombre de personnes par sexe * âge (12 catégories)
- âge de la personne de référence (6 catégories)
- tranche d'unité urbaine (5 catégories)
- diplôme de la personne de référence (4 catégories)
- ZEAT (3 catégories)
- type de ménage (6 catégories)
- catégorie socioprofessionnelle de la personne de référence (7 catégories)
- masse de revenus d'activité déclarés en 2007
- masse de revenus du patrimoine déclarés en 2007
- masse de patrimoine net pour les « hauts patrimoines »

En ce qui concerne la Réunion, les marges de calage sont les suivantes :

- nombre de personnes par sexe * âge (10 catégories)
- taux de propriétaires de leur résidence principale
- catégorie socioprofessionnelle de la personne de référence (8 catégories)
- lieu de naissance de la personne de référence (3 catégories)
- nombre de pièces du logement (6 catégories)
- micro-région (4 catégories)
- masse des revenus d'activité
- masse des revenus du patrimoine

Enfin, pour la partie de l'échantillon sélectionné dans les Antilles, les variables de calage sont :

- âge de la personne de référence (3 catégories)
- catégorie socioprofessionnelle de la personne de référence (4 catégories)
- type de bâti (3 catégories)
- type de logement, i.e. individuel ou collectif
- nombre total de logements
- nombre de ménages propriétaires de leur résidence principale
- nombre d'individus par tranches d'âge

La fonction de lien utilisée pour le calage est le sinus hyperbolique [4], de la forme :

$$G(x) = \int_1^x \frac{1}{2} \sinh\left(\alpha t - \frac{1}{\alpha t}\right) dt$$

en fixant $\alpha \geq 1$. Ici on a fixé d'emblée $\alpha = 1$, ce qui permet de réduire la limite supérieure des poids après calage obtenus par exemple pour les méthodes logit ou linéaire tronquée.

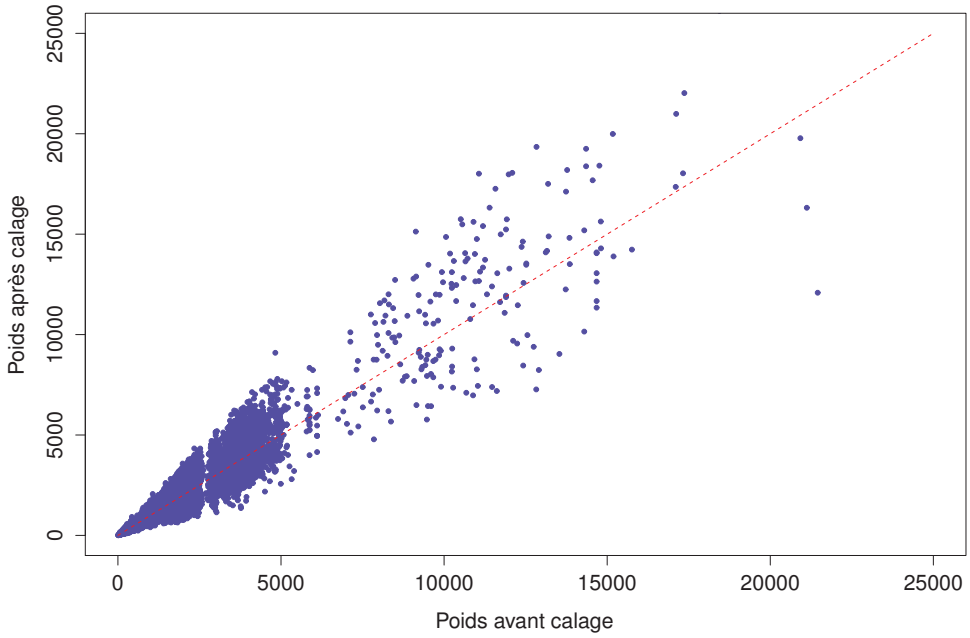


FIGURE 2 – Poids des ménages métropolitains avant et après calage

5 Calcul analytique de la précision de l'enquête Patrimoine 2010

5.1 Principe du calcul

5.1.1 Estimation de l'écart-type associé à l'estimateur

La méthode principale d'estimation de l'aléa de sondage est une méthode analytique, qui repose sur des formules de calcul de la variance. Cette méthode a l'avantage de reposer sur des formules établies sans recours à des hypothèses lourdes. Le seul défaut de cette méthode est qu'elle est difficilement diffusable de manière exhaustive, puisque pour réaliser les calculs de précision, il faut disposer de la routine permettant le calcul et de l'ensemble des informations pertinentes de

4. d'où l'usage de la macro CALMAR2, qui rend possible l'usage de cette fonction.

la base de sondage ayant servi à la sélection des ménages. D'autres méthodes existent, comme par exemple le bootstrap, mis en œuvre dans la partie 6. Cette méthode présente l'avantage d'être facilement diffusable, puisqu'il est possible de fournir aux utilisateurs un jeu de poids *bootstrap*, ou poids répliqués.

Les résultats présentés dans cette partie sont issus des calculs analytiques mis en œuvre par une macro SAS *ad hoc* reprenant la démarche du logiciel POULPE (*cf.* sur ce sujet par exemple Caron, Deville et Sautory (1998)). Comme décrit dans la partie 2, le plan de sondage de l'enquête Patrimoine 2009 peut être vu comme un plan en deux phases. La première phase concerne la sélection des ménages interrogés ; la seconde phase décrit le mécanisme de sélection propre au phénomène de non-réponse. Il faut également tenir compte du processus de calage sur marges. Selon la méthode récursive développée par Durbin (1953), Raj (1966), Rao et Lanke (1984), la variance due au sondage de l'estimation se décompose alors de la manière suivante [5] :

$$\mathbb{V}\hat{Y}_3) = \mathbb{V}(\hat{Y}_1) + \mathbb{E}\left[\mathbb{V}(\hat{Y}_2|s_1)\right] + \mathbb{E}\left[\mathbb{V}(\hat{Y}_3|s_2)\right]$$

où $\hat{Y}_n$ est l'estimateur du total de la variable d'intérêt y, et fonction du degré de tirage n. Ainsi, $\hat{Y}_1$ représente l'estimateur du total de la variable y après sélection des ZAE, calculé sur l'ensemble des logements appartenant aux ZAE sélectionnées. Par ailleurs, s_1 et s_2 symbolisent respectivement l'échantillon de ZAE sélectionnées et l'échantillon de logements sélectionnés. De la même manière, s_3 symbolise l'échantillon de ménages répondants.

Comme vu précédemment, le tirage des ZAE est un tirage proportionnel à la taille, stratifié par région. Pour une région donnée, l'estimation du premier terme se fait grâce à la formule suivante (Caron, Deville et Sautory (1998)) :

$$\hat{\mathbb{V}}(\hat{Y}_1) = \frac{n_1}{n_1 - 1} \sum_{i \in s_1} (1 - \pi_1^{(i)}) \left(\frac{\hat{Y}_{1,i}}{\pi_1^{(i)}} - \frac{\sum_{k \in s_1} (1 - \pi_1^{(k)}) \frac{\hat{Y}_{1,k}}{\pi_1^{(k)}}}{\sum_{k \in s_1} (1 - \pi_1^{(k)})} \right)^2$$

pour laquelle $\pi_1^{(i)}$ représente la probabilité d'inclusion de la ZAE i, n_1 le nombre de ZAE sélectionnées et $\hat{Y}_i$ le total pour la variable d'intérêt y estimé pour la ZAE i.

Pour la suite, le tirage de deuxième degré pouvant être assimilé à un sondage aléatoire simple stratifié, il est simple d'exprimer l'estimation du terme de variance associée à ce tirage de deuxième degré :

$$\hat{\mathbb{E}}\left[\mathbb{V}(\hat{Y}_2|s_1)\right] = \sum_{i \in s_1} \frac{\hat{\mathbb{V}}(\hat{Y}_2^i)}{\pi_1^{(i)}}$$

avec

$$\hat{\mathbb{V}}(\hat{Y}_2^i) = \sum_{h=1}^{H} \left(\frac{n_h^i}{N_h^i} \right)^2 \hat{\mathbb{V}}(Y_{2,h}^i)$$

et

$$\hat{\mathbb{V}}(Y_{2,h}^i) = {N_{i,h}}^2 (1 - \frac{n_{i,h}}{N_{i,h}}) \frac{1}{n_{i,h}} \frac{1}{n_{i,h} - 1} \sum_{j \in s_3(h) \cap \{i\}} \left(\frac{y_{i,j}}{\pi_j^R} - \hat{Y}_{2,h}^i \right)^2$$

De manière générale, les deux premiers n'estiment pas sans biais les variances liées à l'échantillon de 1er et de 2nd degrés. En effet, $\hat{\mathbb{V}}(\hat{Y}_1)$ sur-estime la variance liée au premier degré de

5. La formule présentée n'est pas complètement juste, dans le sens où elle omet un terme liée au processus de non-réponse. Pour des raisons de simplification, nous ne présentons pas ici le calcul de ce dernier terme, bien qu'il soit pris en compte dans le calcul analytique réalisé.

sondage, quand le terme $\hat{\mathbb{E}}\left[\mathbb{V}(\hat{Y}_2|s_1)\right]$ sous-estime la variance liée au second degré de sondage. Au final, la somme des deux termes évalue sans biais la variance liée aux deux premiers degrés de sondage.

Enfin, il reste à estimer le dernier terme, $\mathbb{E}\left[\mathbb{V}(\hat{Y}_3|s_2)\right]$. Comme expliqué dans la section 3, le processus de sélection lié au mécanisme de non-réponse est traditionnellement modélisé comme un processus poissonnien, pour lequel la probabilité de réponse varie en fonction des caractéristiques des ménages. La non-réponse a été traitée par groupes homogènes. On a par conséquent fait ici l'hypothèse que les ménages appartenant à un groupe possédaient tous la même probabilité de répondre à l'enquête, sachant qu'ils appartenaient à s_2, que l'on note π_k^R. Cette hypothèse est par conséquent conservée dans le calcul de la variance. Le tirage poissonnien a ceci de particulier qu'il postule l'indépendance du tirage entre deux éléments distincts. Par conséquent, il est possible d'écrire de manière simple la variance due à la phase de non réponse de la manière suivante :

$$\hat{\mathbb{V}}(\hat{Y}_3|s_2) = \sum_{j \in s_3} (1 - \pi_j^R) \left(\frac{y_j}{\pi_1 \pi_{2|1}^{(j)} \pi_j^R} \right)^2$$

avec $\pi_{2|1}^{(j)}$ la probabilité de sélection des ménages au deuxième degré du tirage, sachant qu'ils appartiennent à une ZAE sélectionnée (égale par ailleurs au terme $\frac{n_h^i}{N_h^i}$). Il est d'ailleurs possible d'exprimer très simplement cette probabilité puisqu'il s'agit ici d'un tirage aléatoire sans remise (en considérant chaque strate de manière isolée, la variance étant additive par strate).

5.1.2 Calcul des intervalles de confiance

Pour les totaux et les moyennes :

On utilise pour ces estimateurs les résultats classiques du théorème central-limite, pour faire l'hypothèse qu'ils suivent une loi normale centrée autour de la valeur estimée $\hat{\theta}$ et d'écart-type $\hat{\sigma}$ l'écart-type estimé en utilisant la méthode décrite dans dans la partie 5.1.1. En notant $q_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite, il est donc possible d'écrire l'intervalle de confiance de niveau α de la façon suivante :

$$IC_\alpha = \left[\hat{\theta} - q_{1-\frac{\alpha}{2}} \hat{\sigma}; \hat{\theta} + q_{1-\frac{\alpha}{2}} \hat{\sigma} \right]$$

Pour les quantiles :

L'écriture d'un intervalle de confiance pour les quantiles est plus compliquée. En particulier, il est possible de montrer que l'estimateur d'un quantile donné suit asymptotiquement une loi normale, mais il faut alors se baser sur des hypothèses beaucoup plus restrictives. En particulier, il faut alors supposer que la fonction de densité f de la variable observée est continue. Une estimation de cette fonction de densité par kernel montre que, pour des montants très élevés de patrimoine par exemple, la valeur de cette densité est très faible, ce qui a pour conséquence directe de faire exploser la taille de l'intervalle de confiance associé à l'estimation des quantiles très élevés (P90, P95, P99 par exemple).

Le choix d'une méthode non paramétrique s'impose alors. En notant $(x_{(1)}, ..., x_{(n)})$ la statistique d'ordre pour la variable x définie sur les n observations de l'enquête, on peut construire un intervalle de confiance basé sur les observations « voisines » du quantile d'ordre τ et la statistique $S_\tau(x_1, ..., x_n) = \sum_{i=1}^n \mathbb{1}_{\{x_i \leq q_\tau\}}$. En notant $q_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$,

on définit ainsi les entiers $i_n = \left\lceil n\tau - q_{1-\alpha/2}\sqrt{\hat{V}(S_\tau)} \right\rceil$ et $j_n = \left\lceil n\tau + q_{1-\alpha/2}\sqrt{\hat{V}(S_\tau)} \right\rceil$, avec $\hat{V}(S_\tau)$ l'estimateur de la variance de la statistique S_τ. Pour n assez grand, ces deux entiers sont compris entre 1 et n, et il est possible de montrer que l'intervalle $\left[x_{(i_n)}; x_{(j_n)}\right]$ est un intervalle de confiance pour q_τ de niveau asymptotique α. Pour une démonstration de ce résultat, on se reportera à l'annexe B.

5.2 Résultats

5.2.1 Précision des principaux indicateurs de l'enquête

Le patrimoine brut moyen est de 259 000 euros, et le calcul de précision montre que l'erreur de mesure associée à cet estimateur est de l'ordre de 2 % de ce montant (tableau 6). Le véritable montant moyen de patrimoine brut se situe avec une probabilité de 95 % entre 249 500 et 268 500 euros.

Indicateurs	Estimation	Ecart-type	CV	Borne inférieure	Borne supérieure
Moyenne	259 000	5 756	2,2 %	249 500	268 500
P99	1 885 200	62 526	3,3 %	1 795 800	1 987 000
P95	841 200	15 214	1,8 %	810 400	867 300
P90	552 300	9 674	1,8 %	534 500	570 900
Q3	307 500	3 623	1,2 %	301 800	314 700
Médiane	150 200	3 042	2,0 %	144 700	154 300
Q1	12 100	937	7,7 %	10 800	14 100
P10	2 700	732	27,1 %	2 500	2 900
P5	1 300	520	40,0 %	1 200	1 400
D9/D1	205	55	26,8 %	114	295
Q3/Q1	25	2	8,0 %	22	29
Part détenue par...					
Les 1 % les plus riches	17,2	1,8	10,4 %	14,3	20,2
Les 10 % les plus riches	48,0	1,3	2,7 %	45,9	50,2
Les 20 % les plus riches	65,1	0,9	1,4 %	63,6	66,6
Les 20 % les plus pauvres	0,2	0,0	3,9 %	0,2	0,2
Les 10 % les plus pauvres	0,1	0,0	4,8 %	0,1	0,1

TABLE 6 – Précision des principaux estimateurs sur le patrimoine brut à partir de l'enquête Patrimoine 2010

Les masses détenues par les déciles et les centiles de ménages sont mesurées avec un peu d'incertitude (tableau 6). La part détenue par le haut de la distribution est estimée avec une précision très satisfaisante (pour le dernier quintile et le dernier décile), beaucoup plus satisfaisante que pour le bas de la distribution. En revanche, la part détenue par le très haut de la distribution (le dernier centile) est estimée de manière beaucoup moins précise. Les rapports de masses (M9/M1 et M8/M2) sont assez instables, ce qui vient essentiellement du fait que la petitesse des dénominateurs rend très sensibles les ratios à une très faible variation de ceux-ci. De fait, ce constat justifie largement le choix d'afficher surtout des rapports de masses avec pour dénominateur la masse détenue par les 50 % des ménages les moins bien dotés. Du côté de la distribution du patrimoine, la précision des estimateurs semble suivre une courbe en U. La surreprésentation des hauts patrimoines a permis d'estimer très précisément le haut de la distribution, et l'erreur commise sur le dernier centile n'est ainsi que de 3 % (tableau 6), ce qui

est un excellent résultat au vu de la forte variabilité propre au haut de la distribution de patrimoine. Les estimations de quantiles sont également très précises dans le milieu de la distribution : ainsi l'erreur commise sur le 3ème quartile est de l'ordre de 1 %. Enfin, les résultats sont moins bons sur le bas de la distribution, pour laquelle le faible nombre de ménages interrogés l'emporte visiblement sur l'homogénéité de cette partie de la population : en effet, l'erreur commise est de l'ordre d'un tiers pour le premier décile (tableau 6), ce qui n'est pas négligeable. De ce point de vue, le choix d'afficher des montants arrondis à la centaine d'euros près est judicieux, puisqu'il a plus d'effet sur les petits montants (cet arrondi « forfaitaire »à la centaine d'euros représente une approximation plus grande pour les montants faibles). Par ailleurs, le calcul des intervalles de confiance montre que l'imprécision en bas de distribution est moins élevée que ne le suggère l'écart-type. En effet, ici, le fait d'utiliser une méthode non paramétrique (dans le sens où l'on ne fait l'hypothèse de normalité de l'estimateur du quantile) d'estimation de l'intervalle de confiance donne des intervalles de confiance beaucoup plus petits qu'ils ne l'auraient été en faisant l'hypothèse de normalité et en linéarisant l'estimateur, deux approximations dont on se dispense dans la méthode adoptée. Les résultats obtenus suggèrent ainsi plutôt un coefficient de variation de l'ordre de 4 % pour le premier décile, tendant à montrer que l'homogénéité dans la bas de la distribution l'emporte sur le petit nombre de ménages sélectionnés.

Indicateurs	Estimation	Ecart-type	CV	Borne inférieure	Borne supérieure
Moyenne	229 300	5 698	2,5 %	218 100	240 400
P99	1 761 800	66 123	3,8 %	1 626 700	1 870 600
P95	755 800	15 798	2,1 %	819 900	878 400
P90	501 600	7 427	1,5 %	484 900	521 200
Q3	275 900	3 276	1,2 %	267 100	283 600
Médiane	113 500	3 377	3,0 %	108 300	118 000
Q1	9 500	924	9,7 %	8 600	10 400
P10	1 600	574	35,9 %	1 400	1 800
P5	300	219	84,2 %	100	400

TABLE 7 – Précision des principaux estimateurs sur le patrimoine net à partir de l'enquête Patrimoine 2010

Les calculs de précision réalisés sur les estimations de patrimoine net apportent des résultats sensiblement identiques (tableau 7). La variabilité de la mesure du patrimoine net moyen est ainsi de l'ordre de 2,5 %, de même ampleur que pour le patrimoine brut. Les constats sur la précision de la mesure de la distribution de patrimoine au sein des ménages français restent également valables une fois que l'on corrige de l'endettement : ainsi le haut de la distribution est mesuré de manière précise, quand le bas est nettement dégradé.

5.2.2 Évaluation de l'efficacité du plan de sondage

On s'intéresse maintenant à l'efficacité relative du plan de sondage. L'objectif est de mesurer à quel point le plan de sondage permet, dans un contexte de contrainte sur la taille globale de l'échantillon, de mesurer le plus précisément possible la répartition du patrimoine au sein de la population française. Pour cela, on compare le résultat obtenu en termes de précision des estimateurs avec ceux qui auraient été obtenus pour un plan de sondage très simple, à savoir un sondage aléatoire simple sans remise. Un tel plan de sondage sert ainsi de *benchmark* pour juger de l'efficacité des choix méthodologiques en matière de précision de la mesure.

La formule du *design effect* est donc la suivante :

$$DEFF = \frac{\hat{V}_s}{\hat{V}_{SAS}}$$

Les résultats obtenus, présentés sur la table 8, montrent ainsi que la surreprésentation des hauts patrimoines a bien permis l'amélioration de la précision des estimations sur le patrimoine des ménages. Ainsi, signe que l'effet du plan de sondage est globalement positif, le *design effect* sur le montant moyen de patrimoine brut ou de patrimoine net est inférieur à 1. Assez corrélé avec le patrimoine, puisqu'il est un moyen d'en acquérir, l'endettement est lui aussi mesuré plus précisément avec le plan de sondage de l'enquête Patrimoine qu'il ne l'aurait été avec un sondage aléatoire simple : le *design effect* associé à l'endettement moyen est de 0,74, légèrement plus élevé que l'estimation des patrimoines net et brut moyens. Le plan de sondage s'avère particulièrement efficace pour capter les montants de patrimoine financier ; le *design effect* associé au montant de patrimoine financier moyen est de 0,31, ce qui révèle que la stratégie de surreprésentation des hauts patrimoines est particulièrement payant pour les éléments patrimoniaux très concentrés, comme le patrimoine financier par exemple. En revanche, sur des éléments beaucoup moins concentrés, comme l'immobilier par exemple, le plan de sondage est beaucoup moins efficace.

De la même manière, le plan de sondage retenu permet de mesurer plus précisément les inégalités : ainsi, le coefficient de Gini a un *design effect* de 0,73. On constate également que la surreprésentation des hauts patrimoines a permis de mieux estimer la part que ceux-ci détiennent dans la masse totale de patrimoine : le *design effect* associé à la part du dernier centile est de 0,65, celui associé à la part du dernier décile de 0,75. En revanche, la situation est dégradée pour le bas de la distribution, pour laquelle la perte de précision consécutive à la surreprésentation des hauts patrimoine est assez élevée.

Indicateurs	Patrimoine brut	Patrimoine net	Endettement	Patrimoine immobilier	Patrimoine financier
Moyenne	0,65	0,69	0,74	0,74	0,31
P99	0,72	0,73	1,10	0,83	0,75
P95	1,23	1,21	1,38	1,56	1,28
P90	1,32	1,36	1,49	1,70	1,47
Q3	1,50	1,41	1,43	1,72	1,56
Médiane	1,41	1,35	1,47	1,51	1,58
Q1	1,68	1,72	1,47	1,37	1,79
P10	2,00	1,91	1,47	1,37	1,88
P5	1,86	1,97	1,47	1,37	1,88

TABLE 8 – *Design effects* pour les principaux estimateurs sur le patrimoine brut à partir de l'enquête Patrimoine 2010

Le calcul des *design effects* permet de mesurer ce qu'aurait été la précision de l'enquête Patrimoine sans surreprésentation des hauts patrimoines. Ainsi, en reprenant les strates ayant servi à construire le plan de sondage et en calculant strate par strate l'effet de sondage, il est possible de reconstruire, pour un estimateur donné, la variance obtenue si la surreprésentation d'une strate était moins forte. Au lieu des 1 766 ménages dits « hauts patrimoines », l'échantillon sans surreprésentation n'aurait contenu que 265 ménages appartenant à cette strate. Ainsi, en utilisant les *design effects* calculés par strate et en recomposant l'échantillon tel qu'il aurait été sans sur-représentation, on trouve que la variance totale relative à l'estimation du patrimoine

brut moyen aurait été 28 % plus élevée, et l'écart-type aurait été accru de 800 euros [6].

5.2.3 Faire la part entre accroissement des inégalités et amélioration de la mesure

L'un des enjeux du calcul de précision des enquêtes est de pouvoir se faire une idée de la significativité des évolutions constatées d'une enquête à l'autre. L'idée, qui peut être formalisée par l'utilisation de tests statistiques, est de vérifier la disjonction des intervalles de confiance entre les différentes estimations. Pour cela, il faut non seulement disposer des estimations de précision pour les différentes enquêtes, mais également s'assurer que l'on procède à méthodologie constante.

Or, parmi les nombreuses innovations qui ont marqué l'enquête Patrimoine 2010, certaines ont pu modifier la nature de l'estimation réalisée grâce à l'enquête. Ainsi, la couverture des DOM en 2010, contrairement à 2004, a pu avoir un impact sur la mesure. Il convient donc de procéder à champ constant et de comparer les chiffres de 2004 et 2010 obtenus sur la France métropolitaine. De la même manière, un certain nombre de modifications sur la façon dont ont été captés les montants des actifs recensés dans l'enquête ont pu eux aussi altérer la mesure : introduction de tranches supplémentaires pour les ménages déclarant la tranche la plus élevée, possibilité offerte aux ménages de déclarer un montant directement en clair, mesure d'un patrimoine dit « résiduel ». Le patrimoine « résiduel » désigne les objets de valeur (bijoux,...), les biens durables (meubles, biens d'équipement, voitures...) qui ne sont pas comptabilisés dans les actifs financiers, les actifs immobiliers ou les actifs professionnels, mais qui cependant ont une valeur patrimoniale, en ce sens qu'ils constituent un élément de richesse, parfois non négligeable, pour les ménages. Ces nouveautés ont modifié la nature de l'imputation réalisée, en modifiant le calcul des coefficients des modèles d'imputation. Là encore, on procède à méthodologie constante, en établissant un jeu de simulations de montants en clair établis comme si le ménage avait répondu avec la méthodologie de 2004 ; en substance, les montants en clair ou les hauts de tranche ont été reversés dans les jeux de tranches proposés en 2004, et les modèles d'imputation ont été ainsi recalculés. On exclut également du calcul du patrimoine brut tout ce qui concerne le patrimoine « résiduel », ce qui a pour conséquence notable de diminuer très nettement le patrimoine des ménages les moins bien dotés.

Pour juger si les inégalités se sont accrues entre 2004 et 2010, il faut retenir les indicateurs susceptibles de renseigner sur le sujet. Naturellement, l'indice de Gini est le premier candidat naturel. Cependant, le recours à cet indice n'est pas concluante, et ce pour plusieurs raisons :

- Tout d'abord, la construction d'un intervalle de confiance pour l'indice de Gini repose sur l'hypothèse que l'estimateur du Gini suit une loi normale, ce qui n'est *a priori* pas évident. Nous verrons dans la section 6 qu'il est possible de se passer de cette hypothèse.
- L'analyse des coefficients de variation, sans même calculer d'intervalle de confiance, laisse à penser que l'évolution entre 2004 et 2010 du Gini n'est pas statistiquement significative. En effet, le coefficient de Gini en 2004 est estimé avec une imprécision de l'ordre de 0,01 pour un niveau à 0,64. En 2010, l'indice de Gini se situe à 0,65, avec une imprécision là aussi de l'ordre de 0,01.
- Cependant, les inégalités peuvent tout à fait avoir crû de manière importante sans que l'indice de Gini en soit affecté. Ainsi, dans un cas extrême où la totalité du patrimoine serait répartie de manière complètement uniforme pour une moitié de la population, un doublement du niveau de celui-ci traduirait une hausse de l'écart entre les deux moitiés de la population, sans que cela ne se reflète sur l'indice de Gini. Ce cas extrême permet ainsi

6. en calculant la variance du patrimoine moyen brut sur un échantillon fictif en s'aidant des *design effects* calculés par strate, on trouve une variance 28 % plus élevée, et un écart-type de 6 500, soit 800 euros de plus que celui réellement observé.

d'illustrer les phénomènes entre 2004 et 2010 : 58 % des ménages, propriétaires de leur logement, ont bénéficié de la hausse importante des prix de l'immobilier entre 2004 et 2010, quand les ménages les moins bien lotis ont vu leur patrimoine, constitués essentiellement de biens durables et d'actifs financiers peu risqués, évoluer beaucoup plus lentement. Pour plus de détail, on peut se reporter à l'Insee Références consacré aux revenus et au patrimoine des ménages, édition 2012 (Lamarche et Salembier (2012)).

Un autre indicateur couramment utilisé pour étudier les inégalités est le rapport inter-quantiles. Nous choisissons ici de regarder le rapport inter-quartile Q3/Q1. Ici non plus, il faut supposer que l'estimateur de ce ratio suit une loi normale si l'on veut pouvoir construire un intervalle de confiance. Cependant, en 2004, le rapport entre le 1er quartile et le 3ème quartile se situait à 27, pour un écart-type de l'ordre de 1. En 2010, le rapport interquartile se situait, **pour la France métropolitaine et à méthodologie constante**, à 40, pour un écart-type de l'ordre de 4. Ici, en revanche, il fait peu de doute que l'accroissement de l'écart entre peu et bien dotés en patrimoine est significatif.

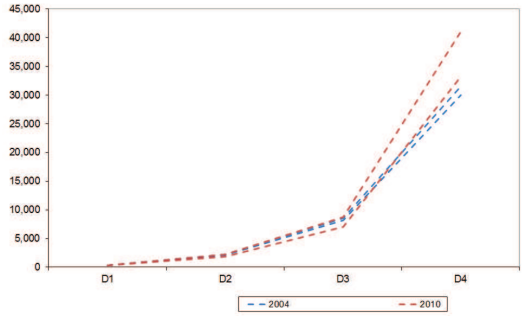


FIGURE 3 – Évolution des patrimoines bruts moyens pour les 4 premiers déciles

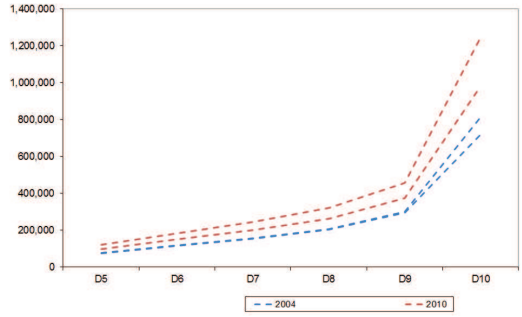


FIGURE 4 – Évolution des patrimoines bruts moyens pour les 6 derniers déciles

Les figures 3 et 4 sont sans ambiguïté sur l'accroissement des inégalités entre 2004 et 2010. Ainsi, comme on peut le voir sur la figure 3, le patrimoine moyen détenu par les 30 % des ménages les moins bien dotés n'a pas évolué de manière significative entre 2004 et 2010, puisque les intervalles de confiance se recouvrent totalement. Mais à partir du 4ème décile de patrimoine brut, on assiste bien à une disjonction complète des intervalles de confiance. Ainsi, les patrimoines moyens se sont considérablement accrus entre 2004 et 2010 pour cette partie de la population. Ceci traduit bien un accroissement des inégalités de patrimoine sur la période considérée, puisqu'une partie de la population voit son patrimoine croître nettement quant l'autre partie constate une stagnation du montant des actifs qu'elle possède.

6 Méthodes de bootstrap appliquée à l'enquête Patrimoine 2010

Les calculs de précision réalisés de manière analytique ont donc permis d'établir un certain nombre de résultats statistiques intéressants. Cependant, ils nécessitent de connaître en détail le plan de sondage de l'enquête, et de disposer de toutes les informations pertinentes dans les données. Précisément, pour des raisons de confidentialité, ces données ne sont en général pas diffusées, et la plupart des utilisateurs n'ont par conséquent pas la possibilité concrète de calculer la précision des estimations réalisées à partir de l'enquête.

Pour autant, il est possible de donner la possibilité aux utilisateurs de calculer eux-mêmes leurs intervalles de confiance, sans avoir recours aux données de plan de sondage. Les méthodes de bootstrap et de poids répliqués permettent ainsi de diffuser l'information relative à la précision sans passer par les données de sondage. La méthode que nous présentons ici a permis le calcul de tels poids répliqués pour l'enquête Patrimoine 2010. Cette méthode s'inspire des travaux menés par exemple par Chauvet (2007), mais elle s'appuie également sur une validation empirique comparant les résultats donnés par les calculs analytiques avec les chiffres fournis par le bootstrap.

6.1 France métropolitaine et Réunion

Pour réaliser le bootstrap, nous distinguons les logements appartenant aux Zones d'Actions Enquêteur dites exhaustives et les autres. Pour les premières, elles ont une probabilité de sélection égale à 1 et par conséquent, il n'y a pas de premier degré de tirage. Pour les autre, la probabilité de tirage est strictement inférieure à 1 et on considère donc qu'un tirage à deux degrés a été concrètement mis en œuvre. Nous appliquons à l'échantillon de ménages répondants vivant dans une Zone d'Action Enquêteur non exhaustive, l'algorithme de Gross. Pour les autres, on applique un simple tirage avec remise. On effectue les opérations suivantes 999 fois, afin d'obtenir un jeu de 1 000 poids permettant d'estimer la précision de l'enquête.

6.1.1 Pour les ZAE non exhaustives

On considère donc ici toutes les Zones d'Action Enquêteurs (ZAE) i de France métropolitaine ayant une probabilité d'inclusion π_i strictement inférieure à 1 et effectivement tirée dans l'enquête Patrimoine. On procède comme suit :

Étape 1 : Duplication des ZAE $\lfloor 1/\pi_i - 1/2 \rfloor$ fois, où $\lfloor . \rfloor$ désigne l'entier le plus proche.

Étape 2 : Tirage à probabilités inégales d'un échantillon supplémentaire de ZAE de taille $N - N^*$, selon les probabilités d'inclusion $\alpha_i = 1/\pi_i - \lfloor 1/\pi_i - 1/2 \rfloor$. Ici, $N = \sum 1/\pi_i = 3\,610$, $N^* = \sum 1/\pi_i - 1/2 = 3\,372$ et $N - N^* = 238$.

Étape 3 : On regroupe les ZAE dupliquées et le tirage supplémentaire de ZAE.

Étape 4 : Dans cette base, on tire selon un sondage proportionnel à la taille, stratifié par région, un échantillon de 488 ZAE, 488 étant le nombre de ZAE non exhaustives effectivement tirées dans l'enquête Patrimoine. Certaines ZAE sont donc tirées plusieurs fois.

Étape 5 : Pour chaque ZAE i tirée à l'étape précédente, on récupère l'échantillon originel de ménages sélectionnés et répondants.

Étape 6 : On sélectionne un échantillon de ménages tirés avec remise, en respectant l'allocation par strate définie au niveau de chaque ZAE. Certains ménages sont donc tirés plusieurs fois.

Étape 7 : On simule un mécanisme de non-réponse : on effectue un tirage poissonnien sur les ménages auxquels on attribue une probabilité π_k^R de répondre à l'enquête, où π_k^R est la probabilité de réponse du ménage k, calculée selon la méthode des scores au moment du traitement de la non-réponse de l'enquête Patrimoine.

Étape 8 : À cette étape, on calcule un premier jeu de poids, qui correspondent aux poids avant calage. Le poids avant calage se calcule de la manière suivante :

$$pd_k = \frac{1}{\pi_i} \times \frac{1}{\pi_{k|i}} \times \frac{1}{\pi_k^R}$$

Étape 9 : Ces poids sont sommés pour les ménages sélectionnés à plusieurs reprises ; les autres ménages répondants mais non sélectionnés lors de la procédure de rééchantillonnage ont un poids nul.

6.1.2 Pour les ZAE exhaustives et la Réunion

Pour les ZAE dont la probabilité de sélection est égale à 1 et pour la Réunion, l'algorithme est simplifié. En effet, les étapes 1 à 4 sont inutiles, et on passe directement à l'étape 5. On procède ensuite comme indiqué ci-dessus.

6.1.3 Calage

Une fois ces différentes opérations réalisées, on regroupe les ménages tirés en un unique échantillon pour la France métropolitaine, puis on cale les poids pd_k sur les marges de calage utilisées lors de la production de l'enquête. De la même manière, on réalise un calage pour l'échantillon réunionnais.

6.2 Guadeloupe et Martinique

Il s'agit d'un sondage à un degré, les quatre premières étapes de l'algorithme métropolitain ne sont donc pas nécessaires. Pour chaque Dom, on effectue les opérations suivantes 999 fois, afin d'obtenir un jeu de 1 000 poids permettant d'estimer la précision de l'enquête.

Étape 1 : On duplique chaque ménage ultra-marin k ayant répondu à l'enquête Patrimoine 2010 $\lfloor 1/\pi_k - 1/2 \rfloor$ fois, avec π_k la probabilité d'inclusion du ménage k dans l'échantillon.

Étape 2 : On effectue un tirage à probabilités inégales d'un échantillon supplémentaire de ménages de taille $N - N^*$, respectant les probabilités d'inclusion $\alpha_k = 1/\pi_k - \lfloor 1/\pi_k - 1/2 \rfloor$.
Pour la Guadeloupe : $N = 109\,871$, $N^* = 109\,352$, $N - N^* = 519$.
Pour la Martinique : $N = 120\,171$, $N^* = 120\,130$, $N - N^* = 41$.

Étape 3 : On tire un échantillon de ménages de façon aléatoire, selon le plan de sondage initial. Les ménages peuvent être tirés plusieurs fois, du fait de la duplication.

Étape 4 : On simule un mécanisme de non-réponse : on duplique $\lfloor 1/\pi_k^R \rfloor$ fois les ménages tirés à l'étape précédente, où π_k^R désigne la probabilité de réponse du ménage k, calculée selon la méthode des scores au moment du traitement de la non-réponse de l'enquête Patrimoine. Puis on effectue un tirage poissonnien sur la base ainsi constituée.

Étape 5 : Les poids finaux de chaque ménage k de notre échantillon valent :

$$poidsfinal_k = \frac{1}{\pi_k} \times \frac{1}{\pi_k^R}$$

Étape 6 : Ces poids sont sommés pour les ménages sélectionnés à plusieurs reprises ; les autres ménages répondants, mais non sélectionnés lors de la procédure de rééchantillonnage ont un poids nul.

Étape 7 : On applique ensuite la procédure de calcul des poids de sorte à obtenir une représentativité sur l'ensemble Antilles + Guyane.

On regroupe les échantillons finaux de Guadeloupe et de Martinique. On effectue enfin le calage sur marges et la prise en compte de la Guyane à partir de cet échantillon final, selon les marges définies lors du calage de l'enquête Patrimoine. On aboutit à un jeu de poids représentatif des Antilles.

6.3 Résultats

On s'intéresse maintenant aux résultats obtenus en termes de variance grâce aux poids bootstrap. Pour avoir une idée de la qualité de l'estimation de la variance « bootstrap », on compare les chiffrages obtenus avec ceux donnés par le calcul analytique décrit dans la section 5.

6.3.1 Estimation de moyennes

Tout d'abord, on examine les résultats de précision obtenus sur les estimateurs de moyennes. Les estimateurs présentés dans cette section peuvent s'apparenter à des estimateurs linéaires, dans la mesure où le dénominateur a été calé (et par conséquent l'incertitude liée au sondage est nulle). La question de la linéarisation du ratio ne se pose pas dans le cadre du bootstrap puisque la variance bootstrap de l'estimateur est estimée de la manière suivante :

$$\hat{V}_B(\hat{\mu}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\mu}^b - \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}^b)^2$$

où $\hat{\mu}^b$ est l'estimateur de la moyenne pour la $b^{\text{ème}}$ itération bootstrap.

Les estimations obtenues en utilisant les poids bootstrap sont globalement proches de celles données par le calcul analytique (*cf.* tableau 9). En effet, la précision de l'estimation du patrimoine brut moyen sur l'ensemble du territoire français est de 2,3 % pour le calcul analytique, et de 2,4 % pour le calcul par bootstrap. La légère surestimation de la variance par la méthode bootstrap que l'on observe pourrait provenir du fait que l'on effectue une hypothèse forte dans le calcul des poids bootstrap en assimilant le tirage systématique réalisé sur les logements à un sondage aléatoire simple (SAS).

En effet, le calcul de variance analytique est en réalité le fruit d'une moyenne de deux estimations, l'une par la méthodes des différences systématiques, l'autre se basant sur les formules classiques de variance pour un SAS. La méthode de différences systématiques utilisée dans la calcul analytique permet de mieux tenir compte de la réduction de l'imprécision que permet le tirage systématique, et a pour effet de sous-estimer la variance réelle de l'enquête. L'hypothèse d'un sondage aléatoire simple a au contraire pour conséquence de surestimer la variance de l'enquête. Par ailleurs, les simulations réalisées par l'Unité des Méthodes Statistiques de l'Insee montrent une variabilité de l'estimation des coefficients de variation de l'ordre de 0,07 point.

Le constat par zone est relativement identique. La France métropolitaine, qui compte pour beaucoup dans l'estimation totale, obtient des résultats similaires. La Réunion est un cas particulier, puisqu'on constate une très nette sous-estimation de la variance par les poids bootstrap. Cette très nette sous-estimation semble provenir du calage, sans qu'il soit possible d'expliquer le phénomène. En effet, les calculs théoriques, qui ne tiennent pas compte de la perte de précision dû à la non-réponse, donnent un écart-type associé à l'estimateur de patrimoine brut moyen de l'ordre de 12 000. Pour réaliser cette estimation, on applique sur les répondants réunionnais la formule suivante :

$$\hat{V}_{SAS} = \sum_{h=1}^{H}\frac{N_h}{N}(1-\frac{n_h}{N_h})\frac{S_h^2}{n_h}$$

Ce résultat est confirmé par le calcul analytique, qui affiche avant calage un écart-type compris entre 13 770 et 13 860. Par ailleurs, les poids bootstrap non calés donnent un écart-type de 13 419. En revanche, les résultats après calage divergent fortement. Pour le calcul par linéarisation, on trouve un écart-type compris entre 12 470 et 12 486, quand le résultat bootstrap indique un gain substantiel de précision dû au calage, puisque l'écart-type est estimé à 7 802. Dans le premier

Champ		Écart-type		CV	
		linéaire	bootstrap	linéaire	bootstrap
France métropolitaine	Patrimoine brut	5 903	6 181	2,3 %	2,4 %
	Patrimoine net	5 843	6 057	2,5 %	2,6 %
	Endettement	765	1 103	2,5 %	3,7 %
	Patrimoine immobilier	2 096	2 246	1,3 %	1,4 %
	Patrimoine financier	1 150	1 348	2,2 %	2,6 %
Réunion	Patrimoine brut	12 348	7 802	6,1 %	3,9 %
	Patrimoine net	10 580	7 203	5,8 %	4,0 %
	Endettement	2 207	1 449	11,2 %	7,4 %
	Patrimoine immobilier	3 993	4 377	2,7 %	3,0 %
	Patrimoine financier	1 722	1 622	11,0 %	10,3 %
Antilles	Patrimoine brut	6 965	8 898	4,6 %	5,9 %
	Patrimoine net	6 838	8 688	5,0 %	6,3 %
	Endettement	813	822	6,4 %	6,4 %
	Patrimoine immobilier	3 176	2 564	3,0 %	2,4 %
	Patrimoine financier	5 109	7 137	30,6 %	42,7 %
France entière	Patrimoine brut	5 764	6 039	2,2 %	2,3 %
	Patrimoine net	5 705	5 918	2,5 %	2,6 %
	Endettement	747	1 078	2,5 %	3,6 %
	Patrimoine immobilier	2 047	2 194	1,3 %	1,4 %
	Patrimoine financier	1 125	1 322	2,2 %	2,6 %

TABLE 9 – Principaux résultats du calcul de précision par zone selon les deux méthodes de calcul

cas, on obtient un gain de précision de l'ordre de 0,01 point sur le coefficient de variation. En revanche, dans le second cas, le gain est de l'ordre de 2,8 points.

Sur la figure 5, on représente la distribution des estimations obtenus à partir des poids répliqués pour le patrimoine moyen. En effet, le fait d'avoir à sa disposition un jeu de 1 000 poids permet de calculer 1 000 estimations différentes, pour laquelle on peut ensuite calculer une distribution. Dans l'hypothèse où l'estimation suit une loi normale, la distribution n'apporte pas plus d'information que ceux déjà fournis par le calcul analytique.

6.3.2 Estimation de quantiles

Il est maintenant intéressant de s'attarder sur les estimateurs de quantiles qui sont fréquemment utilisés dans les études liées au patrimoine. Les résultats sur la précision de ces estimateurs sont essentiels pour juger de la qualité du plan de sondage, en particulier pour vérifier que la sur-représentation des hauts patrimoines a permis d'estimer plus précisément la distribution du patrimoine en France. L'estimation par bootstrap de cette variance se fait pour les quantiles de manière analogue à celle pour les moyennes :

$$\hat{V}_B(\hat{q}_\tau) = \frac{1}{B}\sum_{b=1}^{B}(\hat{q}_\tau^b - \frac{1}{B}\sum_{b=1}^{B}\hat{q}_\tau^b)^2$$

où $\hat{q}_\tau^b$ est l'estimateur du quantile d'ordre τ de la variable x pour la $b^{\text{ème}}$ itération bootstrap.

Pour l'estimation analytique de variance, il est impératif de linéariser les estimateurs des

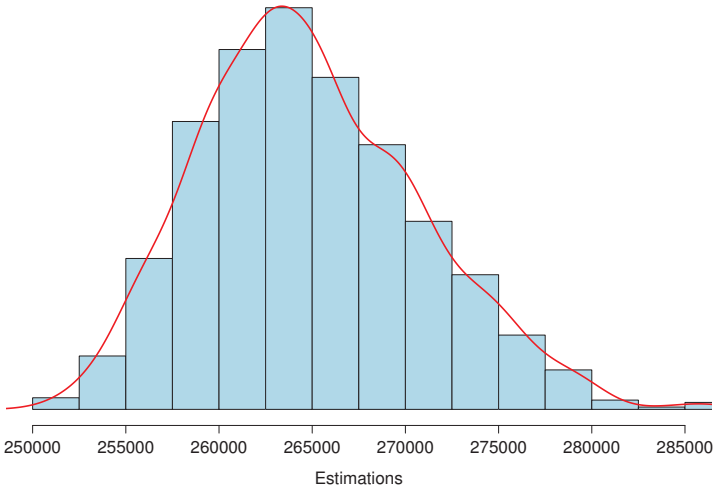


FIGURE 5 – Distribution de l'estimateur de patrimoine brut moyen en France

quantiles, en appliquant la formule de linéarisation suivante :

$$lin_k(\hat{q}_\tau) = -\frac{1}{\sum_{i=1}^{n} \omega_i \hat{f}_{\hat{q}_\tau}(x_i)} (\mathbb{1}_{\{x_k \leq \hat{q}_\tau\}} - \tau)$$

Les résultats obtenus sont globalement assez proches, en particulier pour le haut de la distribution (*cf.* tableau 10). En effet, l'écart d'estimation de la précision que l'on observe sur les quantiles situés entre la médiane et le 99ème centile est très faible. En revanche, sur le bas de la distribution, cet écart peut s'avérer très important : ainsi, pour le premier décile, l'estimation par linéarisation donne un coefficient de variation de 27 %, quand le bootstrap affiche une imprécision de l'ordre de 4 %. De façon similaire à précédemment, la figure 6 donne une idée de la distribution de l'estimateur du patrimoine médian en France.

Quantiles	Estimation	Écart-type		CV	
		linéaire	bootstrap	linéaire	bootstrap
P99	1 885 200	62 526	57 146	3,3 %	3,0 %
P95	841 200	15 214	16 125	1,8 %	1,9 %
D9	552 300	9 674	11 356	1,8 %	2,1 %
Q3	307 500	3 623	3 660	1,2 %	1,2 %
P50	150 200	3 042	3 109	2,0 %	2,1 %
Q1	12 100	937	874	7,7 %	7,2 %
D1	2 700	732	115	27,1 %	4,2 %
P5	1 300	520	70	40,0 %	5,2 %

TABLE 10 – Précision de l'estimation de la distribution du patrimoine brut en France

Pour autant, ce constat n'est pas confirmé lorsqu'on s'intéresse aux intervalles de confiance, pour lesquels les résultats obtenus pour les deux méthodes sont assez proches (*cf.* tableau 11).

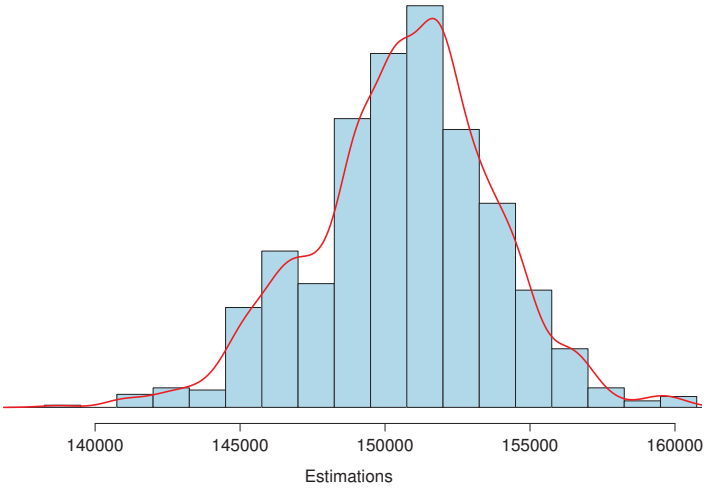


FIGURE 6 – Distribution de l'estimateur de patrimoine brut médian en France

En effet, le résultat de normalité asymptotique sur lequel se base en général la construction des intervalles de confiance pour les estimateurs de moyenne est valable pour les estimateurs de quantiles à la condition que la variable étudiée possède une densité continue. Cette hypothèse n'étant pas forcément vérifiée, le choix est fait de construire des intervalles de confiance de manière non paramétrique, comme décrit précédemment dans la partie 5.1.2.

Indicateur	Estimation	IC analytique		IC bootstrap	
		Borne inférieure	Borne supérieure	Borne inférieure	Borne supérieure
Moyenne	259 000	249 500	268 500	247 100	270 800
P99	1 885 200	1 795 800	1 987 000	1 819 000	2 008 800
P95	841 200	810 400	867 300	822 700	881 400
D9	552 300	534 500	570 900	537 200	579 900
Q3	307 500	301 800	314 700	299 800	315 100
P50	150 200	144 700	154 300	144 500	156 600
Q1	12 100	10 800	14 100	10 500	13 600
D1	2 700	2 500	2 900	2 400	2 900
P5	1 300	1 200	1 400	1 200	1 500

TABLE 11 – Intervalles de confiance de l'estimation de la distribution du patrimoine brut en France

Les intervalles de confiance bootstrap sont calculés avec la méthode du percentile, ce qui permet également d'éviter de faire une hypothèse de normalité asymptotique. De manière générale, les intervalles obtenus, même s'ils ne se recouvrent pas, sont d'ampleur équivalente. Cela souligne une certaine cohérence des résultats sur la précision entre calcul analytique et bootstrap.

6.3.3 Estimation d'indicateurs d'inégalités

On s'intéresse maintenant à la précision obtenue pour des indicateurs d'inégalités, qui sont des estimations centrales pour l'enquête Patrimoine. Ainsi, le rapport interdécile D9/D1 ou encore le coefficient de Gini sont des indicateurs-clés de l'enquête.

Ces indicateurs doivent être linéarisés pour pouvoir calculer analytiquement la variance associée à leur estimation. Pour autant, une fois connu leur écart-type, il n'est pas forcément évident d'obtenir un intervalle de confiance, puisque l'hypothèse de normalité est là encore peu évidente.

Indicateurs	Estimation	Écart-type		CV	
		linéaire	bootstrap	linéaire	bootstrap
D9/D1	205	55	10	26,8 %	40,8 %
Q3/Q1	25	2	2	8,0 %	8,0 %
Part détenue par...					
... les 1 % les plus riches	17,24	0,02	0,02	10,4 %	9,3 %
... les 10 % les plus riches	48,03	0,01	0,01	2,7 %	2,6 %
... les 20 % les plus riches	65,08	0,01	0,01	1,4 %	1,5 %
... les 20 % les plus pauvres	0,23	0,00	0,00	3,9 %	4,3 %
... les 10 % les plus pauvres	0,05	0,00	0,00	4,8 %	5,3 %
Indice de Gini					
Sur l'ensemble des ménages	0,65	0,01	0,01	1,2 %	1,2 %
Sur les 10 % les plus riches	0,38	0,03	0,02	8,5 %	5,9 %

TABLE 12 – Précision de l'estimation de la distribution du patrimoine brut en France

Le calcul de linéarisation fait état, pour l'indice de Gini, d'un écart-type de 0,0077 pour une estimation à 0,65391, soit un coefficient de variation de 1,17 % (*cf.* tableau 12). Le bootstrap donne quant à lui un écart-type de 0,0079, pour un coefficient de variation de 1,19 %. Les intervalles de confiance obtenus par bootstrap donnent pour l'indice de Gini des résultats très proches des intervalles de confiance construits sur l'hypothèse de normalité asymptotique. Il donne en revanche un chiffrage plus important pour l'incertitude liée à l'estimation du rapport D9/D1. Dans tous les cas, ce coefficient est estimé avec une précision assez faible, ce qui milite pour l'usage d'indicateurs alternatifs.

7 Estimation de la variance due à l'imputation des montants

7.1 Rappel sur la méthode des résidus simulés

L'enquête Patrimoine s'appuie sur une méthodologie d'imputation des montants propre à la problématique de la mesure du patrimoine et développée depuis le début des enquêtes portant sur ce thème, en 1986. L'idée est qu'il est difficile de capter un montant précis correspond à la valeur des actifs déclarés par les ménages ; pour aider les ménages, ceux-ci sont invités à fournir des montants en tranches quand ils ne peuvent pas fournir de montant précis. Sauf à sommer les bornes supérieures et inférieures des montants déclarés (ce qui ne résout pas le problème des tranches supérieures) au prix d'une très forte imprécision, la solution retenue est de générer un montant dans la tranche déclarée par les ménages à l'aide de la méthode dite des résidus simulés. Cette méthode a été décrite par Gourieroux, Monfort, Renault et Trognon (1987), et reprise par Lollivier et Verger (1988).

Elle s'appuie sur un modèle de durée ; en effet, la déclaration d'un montant en tranche peut se voir comme une censure à gauche et une censure à droite. Ainsi, pour un actif i dont la valeur

$ln(y_i)$ est positionnée dans la tranche $[m; M]$ et peut s'écrire comme suit :

$$ln(y_i) = X_i\beta + \sigma u_i$$

avec X_i les descripteurs de l'actif i, u_i un résidu de loi normale centrée réduite, on peut écrire la vraisemblance comme suit :

$$L_i = \Phi(\frac{M - X_i\beta}{\sigma}) - \Phi(\frac{m - X_i\beta}{\sigma})$$

Il est alors possible de calculer la vraisemblance globale sur les données et d'estimer, par un algorithme de maximisation, le cofficient β du modèle. On dispose alors d'une estimation $X_i\hat{\beta}$ du montant $ln(y_i)$ de l'actif i, et on génère un résidu $\tilde{u_i}$ selon une loi normale pour obtenir une prédiction $\widetilde{ln(y_i)}$ du montant. Ici on initialise un algorithme d'acceptation-rejet, pour lequel on réitère la génération du résidu $\tilde{u_i}$ jusqu'à ce qu'un certain nombre de contraintes prédéfinies soient respectées (par exemple que le montant $ln(y_i)$ soit inclus dans la tranche $[m; M]$, ce que le modèle ne garantit pas).

7.2 Méthode d'évaluation de la variance d'imputation

Cette méthode a l'avantage de respecter l'information délivrée par le ménage, en tenant compte du résultat de la modélisation. Comme il s'agit d'une imputation de nature stochastique, elle génère une incertitude supplémentaire sur l'évaluation du patrimoine dans l'enquête. Pour autant, il est possible d'évaluer cette imprécision, grâce à la méthode décrite par exemple par Rubin (1996). Cette méthode permet de prendre en compte à la fois la variance due à l'échantillonnage (à la description de laquelle nous nous sommes attachés jusqu'à présent), mais également la variance due à l'imputation. Pour cela, il faut produire non pas une imputation, mais M imputations. Pour chaque imputation θ_m de l'estimateur θ, il est possible de calculer la variance due au sondage que nous noterons U_m, par l'un ou l'autre des méthodes présentées ci-avant. On obtient alors la formule de la variance totale T suivante :

$$T = W + \left(1 + \frac{1}{M}\right) Q$$

où $W = \frac{1}{M}\sum_{m=1}^{M} U_m$ est la variance intra-imputation, et $Q = \frac{1}{M-1}\sum_{m=1}^{M}(\theta_m - \bar{\theta})^2$ est la variance inter-imputation.

7.3 Résultats sur l'enquête Patrimoine 2010

Les résultats obtenus montrent que la prise en compte de la variance due à l'imputation n'accroît pas substantiellement l'imprécision de la mesure du patrimoine dans l'enquête (*cf.* tableau 13). On constate que la prise en compte de la variance due à l'imputation ne modifie globalement pas l'incertitude liée à l'estimation des montants de patrimoine brut ou net. En effet, l'erreur commise sur l'estimation du patrimoine brut moyen est de 2,8 %, soit 0,4 points de plus. La prise en compte de l'imputation n'a par ailleurs que très peu d'effet sur la distribution : l'incertitude associé à l'estimation du patrimoine médian n'a pas varié. Les estimations les plus sensibles aux imputations sont liées au patrimoine financier : ainsi, l'erreur commise sur le patrimoine financier est de l'ordre de 10 %, une fois prise en compte l'incertitude due à l'imputation. En revanche, la distribution n'est là non plus affectée par l'erreur due à l'imputation.

Variables	CV	
	Moyenne	Médiane
Patrimoine brut	2,8 %	2,1 %
Patrimoine net	3,1 %	3,4 %
Endettement	3,6 %	0,0 %
Patrimoine immobilier	1,4 %	3,1 %
Patrimoine financier	9,9 %	4,2 %

TABLE 13 – Coefficient de variation pour les principales mesures du patrimoine

8 Conclusion

Dans une enquête telle que l'enquête Patrimoine 2010, les sources d'imprécision statistique sont de deux ordres : la première est liée au fait que l'on interroge qu'une très petite proportion de la population (et qu'en particulier, une partie des ménages sélectionnés refusent de répondre à l'ensemble du questionnaire), la seconde est liée au fait que les ménages qui acceptent de répondre fournissent cependant une information imprécise ou pas d'information du tout à une partie des questions qui leur sont posées. Les résultats présentés dans la dernière partie montrent clairement que la seconde source joue un rôle relativement mineur dans l'incertitude générale liée à l'enquête. Cela valide *a posteriori* le choix initial de chercher à améliorer substantiellement le plan de sondage de l'enquête au travers de la sur-représentation du haut de la distribution. Les résultats exposés confirment que les choix méthodologiques qui ont été adoptés ont considérablement amélioré la précision de l'enquête. Enfin, l'objet de ce document de travail est également de présenter une caractéristique additionnelle de l'enquête, pour laquelle des poids répliqués ont été calculés. Ces poids répliqués donnent ainsi la possibilité à l'utilisateur d'avoir une idée pour toute estimation effectuée sur les données de l'enquête de la précision de cette estimation. En ce sens, la mise à disposition de ce type de variables dans les fichiers de l'enquête, bien que très expérimental, constitue un progrès pour l'utilisateur.

Références

Beaumont, J.-F. et D. Haziza (2007) : "On the Construction of Imputation Classes in Surveys," *International Statistical Review*, 75(1), 25–43.

Beaumont, J.-F. et Z. Patak (2012) : "On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling," *International Statistical Review*, 80(1), 127–148.

Bertail, P. et P. Combris (1997) : "Bootstrap généralisé d'un sondage," *Annales d'Économie et de Statistique*.

Canty, A. J. et A. C. Davison (1999) : "Resampling-Based Variance Estimation for Labour Force Surveys," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 48(3), pp. 379–391.

Caron, N., J.-C. Deville et O. Sautory (1998) : *Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE*, Document de travail - INSEE. INSEE.

Chauvet, G. (2007) : *Méthodes de bootstrap en population finie*Université de Rennes 2.

Dell, F., X. d'Haultfoeuille, P. Février et E. Massé (2002) : "Mise en œuvre du calcul de variance par linéarisation," *Actes des Journées de Méthodologie Statistique*, INSEE.

Durbin, J. (1953) : "Some Results in Sampling Theory when the Units are Selected with Unequal Probabilities," *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2), pp. 262–269.

Gourieroux, C., A. Monfort, E. Renault et A. Trognon (1987) : "Generalised residuals," *Journal of Econometrics*, 34(1), 5–32.

Lamarche, P. et L. Salembier (2012) : "Les déterminants du patrimoine : facteurs personnels et conjoncturels," dans *Les revenus et le patrimoine des ménages*, Insee Références. Insee.

Lollivier, S. et D. Verger (1988) : "D'une variable discrète à une variable continue : une application de la méthode des résidus simulés," *Mélanges économiques, essais en l'honneur de Edmond Malinvaud*, pp. 811–831.

Preston, J. (2009) : "Rescaled bootstrap for stratified multistage sampling," Document de Travail, Statistics Canada.

Raj, D. (1966) : "Some Remarks on a Simple Procedure of Sampling Without Replacement," *Journal of the American Statistical Association*, 61(314), pp. 391–396.

Rao, J. N. K. et J. Lanke (1984) : "Simplified Unbiased Variance Estimation for Multistage Designs," *Biometrika*, 71(2), pp. 387–395.

Rao, J. N. K. et C. F. J. Wu (1988) : "Resampling Inference With Complex Survey Data," *Journal of the American Statistical Association*, 83(401), pp. 231–241.

Rubin, D. B. (1996) : "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, 91(434), 473–489.

Rubin, D. B. (2008) : *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.

Saigo, H. (2010) : "Comparing Four Bootstrap Methods For Stratified Three-Stage Sampling," *Journal of Official Statistics*, 26(1), pp. 193–207.

Annexes

Annexe A

Constante	0.01
	(0.14)
Age	0.02^{***}
	(0.00)
Age au carré	0.00^{***}
	(0.00)
Commune rurale	-0.01
	(0.06)
UU de moins de 5 000 hab.	0.11
	(0.08)
UU de 5 000 à 9 999 hab.	0.04
	(0.08)
UU de 10 000 à 19 999 hab.	0.13
	(0.07)
UU de 20 000 à 49 999 hab.	0.00
	(0.07)
UU de 50 000 à 99 999 hab.	-0.13
	(0.07)
UU de 100 000 à 199 999 hab.	-0.03
	(0.08)
UU de 200 000 à 1 999 999 hab.	0.06
	(0.07)
Ménage occupant une maison	0.02
	(0.03)
Logement occupé par le propriétaire	-0.01
	(0.03)
Logement vacant	-0.05
	(0.25)
Logement HLM	0.10^{*}
	(0.04)
Protocole spécifique	-0.24^{**}
	(0.09)
Echantillon de réserve	-0.12^{***}
	(0.03)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Strate agriculteur	0.13^{*}
	(0.05)
Strate âgés	0.04
	(0.05)
Strate hauts indépendants	-0.18^{***}
	(0.04)
Strate Revenus du patrimoine	-0.02
	(0.05)
Strate cadres	-0.15^{**}
	(0.06)
Logement de moins de 40 m^2	-0.31^{**}
	(0.10)
Logement de 40 à moins de 60 m^2	-0.16^{*}
	(0.07)
Logement de 60 à moins de 80 m^2	-0.11^{*}
	(0.05)
Logement de 80 à moins de 100 m^2	-0.03
	(0.05)
Logement de 100 à moins de 120 m^2	0.02
	(0.04)
Ménage complexe	0.02
	(0.04)
Personne seule	-0.11^{**}
	(0.03)
Couple avec enfant	0.03
	(0.03)
Famille monoparentale	-0.11^{*}
	(0.05)
2eme quintile de revenus d'activité	0.07^{*}
	(0.03)
3eme quintile de revenus d'activité	0.09^{*}
	(0.04)
4eme quintile de revenus d'activité	0.06
	(0.04)
5eme quintile de revenus d'activité	0.05
	(0.05)
1er quintile de revenus du patrimoine	-0.04
	(0.07)
4eme quintile de revenus du patrimoine	-0.04
	(0.03)
5eme quintile de revenus du patrimoine	-0.10^{*}
	(0.04)
Revenus d'activité $>$ 100 000 euros annuels	-0.16^{*}
	(0.07)
Revenus du patrimoine $>$ 100 000 d'euros annuels	-0.37
Les régions de gestion sont également des variables de contrôle.	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 14 – Résultats du modèle probit sur l'échantillon standard, France métropolitaine

Constante	8.09
	(12.15)
Age	0.05^{**}
	(0.02)
Age au carré	0.00^{**}
	(0.00)
Ménage occupant une maison	0.15^{*}
	(0.06)
Strate riches urbains	-0.04
	(0.14)
Strate dominante mobilière	-0.13
	(0.10)
Strate dominante immo.	-0.09
	(0.07)
Log du patrimoine brut	-1.25
	(1.62)
Log du patrimoine brut au carré	0.04
	(0.05)
Ménage complexe	-0.10
	(0.08)
Personne seule	-0.25^{***}
	(0.07)
Couple avec enfant	-0.14
	(0.08)
Famille monoparentale	-0.12
	(0.18)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

2eme quintile de revenus d'activité	0.07
	(0.11)
3eme quintile de revenus d'activité	0.05
	(0.10)
4eme quintile de revenus d'activité	0.26^{**}
	(0.09)
5eme quintile de revenus d'activité	0.33^{***}
	(0.09)
1er quintile de revenus du patrimoine	0.27
	(0.17)
4eme quintile de revenus du patrimoine	0.23
	(0.15)
5eme quintile de revenus du patrimoine	0.19
	(0.13)
Revenus d'activité $>$ 100 000 euros annuels	-0.11
	(0.07)
Revenus du patrimoine $>$ 100 000 euros annuels	-0.16
	(0.09)
Revenus du patrimoine $>$ 1 million d'euros annuels	-0.17
	(0.14)
Revenus du patrimoine $>$ 10 million d'euros annuels	-0.32
	(0.25)
Ménage propriétaire du logement	0.10
	(0.18)
Les régions de gestion sont également des variables de contrôle.	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 15 – Résultats du modèle probit sur l'échantillon hauts patrimoines, France métropolitaine

Annexe B

Soit $\tau \in]0;1[$, et x_τ le quantile d'ordre τ de la variable d'intérêt x, et $(X_{(1)}, ..., X_{(n)})$ sa statistique d'ordre. On note également $q_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. Plutôt que de s'intéresser à la statistique x_τ, on s'intéresse plutôt à la statistique $S_n^\tau(x_1, ..., x_n) = \sum_{i=1}^n \mathbb{1}_{\{x_i \leq x_\tau\}}$. On introduit également $\hat{V}(S_n^\tau)$ l'estimateur de la variance de la statistique S_n^τ.

On pose alors $Z_n = \sqrt{n}\,\frac{S_n^\tau - n\tau}{\sqrt{\hat{V}(S_n^\tau)}}$. D'après le théorème central-limite, la suite $(Z_n, n \geq 1)$ converge en loi vers Z de loi normale centrée réduite.
Par conséquent, si l'on considère les entiers $i_n = \left\lceil n\tau - q_{1-\alpha/2}\sqrt{\hat{V}(S_n^\tau)} \right\rceil$ et $j_n = \left\lceil n\tau + q_{1-\alpha/2}\sqrt{\hat{V}(S_n^\tau)} \right\rceil$, on a :

$$\begin{aligned}
\mathbb{P}(X_{(i_n)} \leq x_\tau \leq X_{(j_n)}) &= \mathbb{P}(i_n \leq S_n^\tau \leq j_n) \\
&= \mathbb{P}\left(\sqrt{n}\frac{i_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}} \leq Z_n \leq \sqrt{n}\frac{j_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}}\right)
\end{aligned}$$

Or pour $n \leq n_0 \leq 1$, et en utilisant le fait que $i_n \leq n\tau - q_{1-\alpha/2}\sqrt{\hat{V}(S_n^\tau)}$ et $j_n \leq n\tau + q_{1-\alpha/2}\sqrt{\hat{V}(S_n^\tau)}$:

$$\begin{aligned}
\mathbb{P}\left(\sqrt{n}\frac{i_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}} \leq Z_n \leq \sqrt{n}\frac{j_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}}\right) &\geq \mathbb{P}(-q_{1-\alpha/2} \leq Z_n \leq q_{1-\alpha/2} - \frac{1}{\sqrt{n_0}\sqrt{\hat{V}(S_{n_0}^\tau)}}) \\
\mathbb{P}\left(\sqrt{n}\frac{i_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}} \leq Z_n \leq \sqrt{n}\frac{j_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}}\right) &\leq \mathbb{P}(-q_{1-\alpha/2} - \frac{1}{\sqrt{n_0}\sqrt{\hat{V}(S_{n_0}^\tau)}} \leq Z_n \leq q_{1-\alpha/2})
\end{aligned}$$

En faisant tendre n et n_0 vers $+\infty$, on a :

$$\mathbb{P}\left(\sqrt{n}\frac{i_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}} \leq Z_n \leq \sqrt{n}\frac{j_n - n\tau}{\sqrt{\hat{V}(S_n^\tau)}}\right) \overset{+\infty}{\to} \mathbb{P}(-q_{1-\alpha/2} \leq Z \leq q_{1-\alpha/2}) = 1 - \alpha$$

Par conséquent, l'intervalle de confiance $[X_{(i_n)}; X_{(j_n)}]$ est bien défini et de niveau α pour x_τ.