Estimation des inégalités dans l'enquête Patrimoine 2004

Eric Gautier, Cédric Houdré

Estimation des inégalités dans l'enquête *Patrimoine* 2004

Éric Gautier* et Cédric Houdré**

Dans les enquêtes sur le patrimoine, les questions sur les montants, celles où le ménage doit par exemple donner un montant détenu sur tel ou tel produit financier ou immobilier, proposent souvent de fournir un intervalle plutôt qu'un montant précis. Cette stratégie permet de réduire le taux de non-réponse. En contrepartie, les montants déclarés ne sont plus des valeurs ponctuelles permettant de calculer directement des indicateurs d'inégalités.

Cet article décrit une procédure générale permettant l'estimation ponctuelle d'indices d'inégalité et l'obtention d'intervalles de confiance. Cette méthode est adaptée à une collecte par sondage et à des données en intervalles. Elle repose sur une modélisation des indices d'inégalité à deux ou trois « étages », qui constituent, par emboîtement, un modèle « hiérarchique ». Le premier étage décrit le sondage, les deux autres le processus de génération des données de patrimoine total. La modélisation de ce processus utilise, outre les observations de variables socio-démographiques disponibles dans l'enquête, différents ensembles d'information : les intervalles déclarés par les ménages pour les encours de patrimoine et des informations auxiliaires telles que l'imposition à l'Impôt de Solidarité sur la Fortune (ISF). La procédure permet d'obtenir des intervalles de confiance tenant compte de l'aléa de sondage et de l'incertitude sur les montants qui sont observés de manière imprécise. L'article discute plus particulièrement la modélisation de la variable de patrimoine brut total et deux modèles sont considérés : un modèle décrivant directement la variable de patrimoine brut total qui est recueillie dans l'enquête, ainsi qu'un modèle à équations simultanées décrivant simultanément plusieurs composantes agrégées du patrimoine brut total.

En utilisant l'ensemble d'information le plus complet, le patrimoine brut moyen se situerait début 2004 autour de 205 000 euros et l'indice d'inégalité de Gini vaudrait environ 0,65, ce qui constitue un niveau d'inégalité comparable à celui estimé par le passé sur les enquêtes Patrimoine. Toutefois, l'ajout d'information sur l'imposition à l'ISF permet de réduire significativement la largeur des intervalles de confiance.

** ENSAE - CREST, Timbre J120, 3 avenue Pierre Larousse, 92240 Malakoff, gautier@ensae.fr. Éric Gautier travaillait à l'Unité Méthodes Statistiques de l'INSEE lorsque ce travail a été initié.*

*** Cédric Houdré appartenait à la Division Revenus et Patrimoine des Ménages de l'INSEE au moment de la rédaction de cet article, cedric.houdre@dgtpe.fr*

Les auteurs remercient leurs collègues à la Direction des Statistiques Démographiques et Sociales de l'INSEE et au CREST et les membres du groupe de travail Patrimoine pour des discussions enrichissantes, parmi lesquels : Luc Arrondel, Céline Bessière, Pascal Chevalier, Marie Cordier, Sibylle Gollac, Christian Robert, Muriel Roger, Catherine Rougerie, Alain Trognon et Daniel Verger. Nous remercions également les participants des séminaires de la Direction des Statistiques Démographiques et Sociales, de recherche en économétrie de Yale et de la European Conference on Quality in Survey Statistics *de 2006 où une partie de ces résultats a été présentée. Les remarques de deux rapporteurs anonymes ont permis d'améliorer substantiellement la présentation de l'article.*

L'analyse microéconomique des inégalités de patrimoine s'appuie généralement sur des indicateurs synthétiques nécessitant l'utilisation de valeurs exactes pour les montants de patrimoine ou les encours d'actifs patrimoniaux. En France, les données sur le patrimoine ne sont pas très nombreuses. Il existe principalement des données d'origine fiscale, issues des déclarations à l'Impôt de Solidarité sur la Fortune (ISF) ou des enregistrements de successions, et des données d'enquêtes auprès des ménages, réalisées par l'Insee.

Les données sur les successions ont l'avantage d'être quasi exhaustives, mais les petites successions ne font pas l'objet d'une obligation de déclaration (1), ce qui peut conduire à sous-estimer les inégalités, et ces données concernent une population très spécifique : les défunts. Il est donc nécessaire, pour obtenir des indices d'inégalité pour la population totale, d'extrapoler le patrimoine au décès au patrimoine de la population en vie. Ceci repose sur l'utilisation de tables de mortalité et d'hypothèses sur la relation entre mortalité et niveau de richesse (2).

Les données sur le patrimoine assujetti à l'ISF ont également leurs limites. D'une part, entre 2 et 3 % des foyers fiscaux seulement sont redevables de cet impôt, le seuil d'imposition s'élevant à 770 000 euros en 2008. Par ailleurs, l'assiette d'imposition exclut pour une large part les actifs professionnels et les objets d'art. Ces données ne peuvent donc pas être utilisées en tant que telles pour mesurer les inégalités de patrimoine sur l'ensemble de la population. Elles ne peuvent servir que de complément à des sources plus complètes à la fois en termes de composantes de patrimoine recensées et de population observée.

Les enquêtes *Patrimoine* de l'Insee constituent une source naturelle pour la mesure des inégalités. Réalisées auprès d'environ 10 000 ménages tous les six ans, elles portent sur l'ensemble de la population métropolitaine et collectent une information très détaillée sur l'ensemble des éléments de patrimoine des ménages. Pour le seul patrimoine financier, plus de 30 types de produits différents sont recensés : des livrets d'épargne règlementée aux valeurs mobilières en passant par les livrets soumis à l'impôt (livrets B, livrets Orange), les produits d'assurance-vie et d'épargne-retraite, ceux d'épargne-logement, ou encore d'épargne salariale. Le questionnaire recense également le patrimoine immobilier de jouissance (résidence principale et secondaire) et de rapport (logements mis en location), ainsi que le patrimoine professionnel, exploité ou non par le ménage. Deux questions servent enfin à collecter une information récapitulative sur la somme des composantes du patrimoine financier et sur le patrimoine total. Cependant, ces enquêtes font face, en France comme à l'étranger, à une difficulté majeure dans l'observation des encours : la non-réponse. Juster et Smith (1997) rapportent que dans les enquêtes américaines *Health and Retirement Study* (*HRS*) et *Aging and HEAlth Dynamics* (*AHEAD*), les taux de non-réponse aux questions de montants peuvent atteindre 20 à 40 %. D'autre part, même lorsque le ménage fournit pour un montant une valeur ponctuelle, celui-ci est souvent déclaré avec une marge d'erreur non négligeable. Pour contourner cet obstacle, il est possible de proposer au ménage de donner une réponse en intervalle. Cette stratégie a l'avantage de conserver une part de l'information mais, en contrepartie, il n'est plus possible d'utiliser les procédures statistiques standards qui requièrent l'observation ponctuelle des montants pour tout l'échantillon.

La mesure des encours dans l'enquête : valeurs ponctuelles, intervalles et non-réponse

Pour favoriser la restitution d'information sur les encours, même partielle, l'enquête *Patrimoine* retient deux stratégies suivant les actifs patrimoniaux considérés. Pour la résidence principale, le ménage est d'abord interrogé sur le montant exact de sa résidence. S'il répond une valeur, nous appelons par la suite ce type de réponse une valeur ponctuelle, sinon, il est invité à donner des bornes inférieures et supérieures, qu'il choisit lui-même (des « fourchettes »), encadrant la valeur de son bien. Même pour ce type de bien, tangible et dont les ménages ont une connaissance pratique de la valeur d'usage, la fréquence des réponses autres qu'une valeur ponctuelle est très élevée (tableau 1). Malgré ce procédé de collecte, environ 8 % des ménages répondant à l'enquête et possédant une résidence principale, ne déclarent aucune valeur ponctuelle ou en intervalle. Il s'agit du cas usuel de non-réponse partielle.
Pour les actifs financiers en revanche, le ménage choisit des intervalles prédéfinis parmi une grille proposée par l'enquêteur. C'est cette fois-

1. Le seuil d'obligation de déclaration était par exemple de 50 000 euros en 2008 pour les successions en ligne directe ou entre conjoint et de 3 000 euros pour les autres.
2. Voir par exemple Piketty, Postel-Vinay et Rosenthal (2006).

ci dans un système d'intervalles prédéfinis (des « tranches ») que les ménages doivent situer le total de leurs actifs financiers, et le total de leur patrimoine, y compris les biens durables, objets d'art et de valeurs à travers deux questions récapitulatives. Pour les actifs professionnels, la question recueille d'emblée un intervalle dont les bornes sont choisies par le répondant.

Le procédé permet de substituer une partie des données qui seraient manquantes par des données en intervalles. Les intervalles peuvent être de nature légèrement différente : majorés, minorés par une valeur strictement positive et non majorés. On pourrait aussi considérer que les valeurs ponctuelles et la non-réponse partielle correspondent à des intervalles particuliers. En pratique, les différents actifs sont plus ou moins affectés par l'incertitude provenant de l'observation d'intervalles (tableau 1). Pour les contrats d'assurance-vie par exemple, la dernière tranche proposée au répondant avait une borne inférieure à 230 000 euros, mais si plus de 95 % des réponses données sont des intervalles usuels (tableau 1), seulement 0,1 % des contrats recensés sont situés dans cette dernière tranche. De même, le seuil de la dernière tranche à 450 000 euros est moins préoccupant pour le patrimoine financier (puisque moins de 1 % des ménages déclarent disposer d'un patrimoine financier supérieur à ce montant) que pour le patrimoine total (puisque c'est alors plus de 7 % des ménages qui se placent dans cette tranche supérieure). Par conséquent, si l'estimation des indices d'inégalité ne devait reposer que sur le montant récapitulatif de patrimoine total, l'incertitude sur les montants liée à l'observation d'intervalles serait vraisemblablement plus forte qu'avec une estimation qui utilise aussi les montants déclarés pour des composantes plus détaillées de patrimoine. Notons enfin que la détention des différentes composantes de patrimoine est ici supposée parfaitement observée.

Pour illustrer l'apport d'information lié à la description dans le détail du patrimoine des ménages, le plus simple est de s'intéresser aux ménages dont le patrimoine brut total est supérieur à 450 000 euros, soit 1 059 ménages sur les 9 692 ménages de l'enquête *Patrimoine* 2004. Compte tenu des grandes disparités de patrimoine, ce seuil est relativement bas. En utilisant les intervalles déclarés pour les composantes de patrimoines, il est possible de réduire le niveau d'incertitude sur le montant de patrimoine total. Par exemple, la somme des bornes inférieures des différentes composantes peut dépasser 450 000 euros dans certains cas. Par ailleurs, il est possible, en utilisant une décomposition adéquate du patrimoine total, de calculer des majorants et minorants du patrimoine assujetti à l'ISF (voir les formules ci-dessous), et, par appariement avec des sources fiscales, d'utiliser l'imposition du ménage pour préciser l'intervalle dans lequel se trouvent les composantes de son patrimoine total (en manipulant les formules (1) et (2) et les bornes des différents intervalles) puis son patrimoine total. L'assiette d'imposition de l'ISF est moins large que le patrimoine recensé dans l'enquête : le patrimoine professionnel n'est pas entièrement pris en compte (par exemple si l'on ne possède qu'une part trop faible d'une entreprise, il n'est pas possible de déduire ce montant du calcul du patrimoine imposable), la résidence principale bénéficie d'un abattement de 20 %, les objets d'art ne sont pas non plus imposés. Pour pouvoir utiliser l'imposition à l'ISF, ce que nous proposerons dans une des estimations, la décomposition du patrimoine total doit au moins distinguer les composantes suivantes : le patrimoine

Tableau 1
Formulation des questions et comportements de réponse

En %

	Résidence principale	Assurance-vie	Patrimoine financier	Patrimoine total
Part de ménages détenteurs	55,7	29,7	100,0	100,0
Valeur des actifs patrimoniaux				
Ponctuelle	12,3	0,4	-	-
Non-réponse	8,3	6,6	4,9	4,8
Intervalle	79,4	95,1	95,1	94,2
Dont				
Borne inférieure nulle	0,5	24,9	24,5	7,1
Borne supérieure manquante	2,8	0,1	0,7	7,5

Lecture : 55,7 % des ménages sont propriétaires de leur résidence principale, et 12,3 % des valeurs de résidences principales déclarées dans l'enquête par les propriétaires sont des valeurs ponctuelles.
Champ : ensemble des ménages de France métropolitaine.
Source : enquête Patrimoine *2004, Insee.*

financier (FIN), la résidence principale (RP), les autres logements (ALG), le patrimoine professionnel exploité à titre professionnel ou non (PROF), le patrimoine professionnel non imposable (NDED), les autres éléments de patrimoine comme les biens meubles ou les objets de valeur (RESTE) et les éléments de passif (PASSIF). Avec une telle décomposition, deux situations sont possibles :

• Lorsqu'un ménage est imposé à l'ISF, son patrimoine imposable est supérieur, en 2003, à 720 000 euros. Le majorant suivant du patrimoine imposable

$$FIN_k + 0.8 * RP_k + ALG_k + MIN(PROF_k, NDED_{max,k}) + RESTE_k - PASSIF_k \quad (1)$$

doit donc être supérieur à 720 000 euros. La variable $NDED_{max,k}$ correspond à une borne supérieure de la valeur maximale de l'ensemble du patrimoine professionnel non déductible construite à partir des informations détaillées. Nous supposons que le passif est constamment déduit.

• Lorsqu'un ménage n'est pas imposé à l'ISF, son patrimoine imposable est inférieur à 720 000 euros. Le minorant suivant du patrimoine imposable

$$FIN_k + 0.8 * RP_k + ALG_k + NDED_{min,k} - PASSIF_k \quad (2)$$

doit donc être inférieur à 720 000 euros. $NDED_{min,k}$ est une borne inférieure du patrimoine professionnel non déductible construite à partir des informations détaillées.

Les intervalles renseignés pour des composantes permettent bien de définir une borne inférieure au-dessus de 450 000 euros pour près de la moitié des ménages initialement situés dans cette dernière tranche (tableau 2). L'imposition à l'ISF apporte effectivement de l'information puisque la somme des minorants des composantes dépasse alors 450 000 euros pour plus de 52 % des ménages de l'échantillon contre 40 % sinon. Plus généralement, utiliser l'information sur l'imposition à l'ISF et les minorants et majorants des composantes permet de déplacer vers le haut les minorants du patrimoine total.
Il n'est pas possible de déduire de la non-imposition à l'ISF, par analogie, une borne supérieure du patrimoine total, puisqu'une partie du patrimoine professionnel et les biens durables et objets d'art sont exclus du patrimoine imposable. Toutefois, cette information est utilisée dans une des estimations et permet de déterminer des bornes supérieures pour les composantes qui constituent ce patrimoine imposable et qui apparaissent dans la formule (2) ci-dessus.

Les patrimoines élevés étant assez rares puisque la distribution du patrimoine est très concentrée, le plan de sondage de l'enquête surreprésente certaines catégories plus aisées que d'autres (ce qui améliorerait la précision des estimations si les montants étaient effectivement des valeurs ponctuelles). Néanmoins, il est relativement

Tableau 2
Distribution de la borne inférieure du patrimoine brut total pour les ménages se situant dans la tranche supérieure (égale ou supérieure à 450 000 euros) de la question récapitulative

En %

Borne inférieure du patrimoine brut total	Information mobilisée		
	Variable récapitulative seule	Somme des bornes inférieures des composantes sans prise en compte de l'ISF	Somme des bornes inférieures des composantes avec prise en compte de l'ISF
Moins de 450 000 euros	0	59,9	47,4
Plus de 450 000 euros	100	40,1	52,6
Dont			
Entre 450 et 500 000 euros	-	9,3	10,6
Entre 500 et 750 000 euros	-	21,4	25,7
Entre 750 et 1 000 000 euros	-	5,8	8,8
Entre 1 000 et 3 000 000 euros	-	3,4	6,7
Entre 3 000 et 10 000 000 euros	-	0,2	0,7
Plus de 10 000 000 euros	-	0,0	0,1

Lecture : 25,7 % des ménages situant leur patrimoine brut total au-delà de 450 000 euros à la question récapitulative disposent en fait d'un patrimoine dont on peut situer une borne inférieure entre 500 et 750 000 euros en s'appuyant sur les intervalles qu'ils ont déclarés pour les différents actifs de patrimoine et sur leur imposition à l'ISF.
Champ : ménages interrogés dans l'enquête Patrimoine *2004.*
Source : enquête Patrimoine *2004, Insee, calculs des auteurs.*

paradoxal d'utiliser un tel plan de sondage à probabilités inégales et de conjointement procéder à une collecte par intervalles avec une question récapitulative dont le plancher de 450 000 euros de la dernière tranche est relativement faible. Au vu de la seule variable récapitulative du patrimoine brut total, un ménage milliardaire est parfaitement substituable à un ménage au patrimoine de 451 000 euros. Nous verrons dans la partie suivante comment nous parvenons à estimer, grâce à une approche basée sur la simulation, les résumés de la distribution de patrimoine malgré des données en intervalles. Mais, pour autant, il semble très difficile de pouvoir dire si un tel plan réduit la couverture des intervalles de confiance. On conçoit également que pour vraiment exploiter les propriétés du plan de sondage il faille mobiliser le plus d'information possible sur le patrimoine total et ne pas se contenter de la variable synthétique de patrimoine total.

Estimation d'indices d'inégalité de patrimoine des ménages à partir de données d'enquête et en présence de non-réponse et de réponses en intervalles

Résumons à ce stade les données du problème. L'objectif du travail est de produire des estimations, ponctuelles ou en intervalles, de certains « résumés » de la distribution du patrimoine des ménages, par exemple la moyenne, la médiane, certains quantiles, des rapports inter-quantiles, ou encore des indices d'inégalité plus complexes comme l'indice de Gini ou l'indice de Theil. La démarche générale d'estimation est illustrée sur l'indice de Gini (3). Si tous les patrimoines dans l'échantillon étaient des valeurs ponctuelles, on saurait donner un estimateur $\hat{G}$ de l'indice de Gini et un intervalle de confiance à 95 % tenant compte de l'aléa de sondage (voir encadré 1).

Cependant, le patrimoine est observé sous forme d'intervalles : il est impossible de calculer directement l'estimateur $\hat{G}$ et une approximation de sa variance asymptotique, puisqu'ils nécessitent tous les deux l'observation de valeurs ponctuelles du patrimoine des ménages répondants. La démarche d'estimation proposée dans cet article repose sur une modélisation du niveau de patrimoine des ménages à travers la description du *processus de génération des données* (PGD). La valeur exacte du patrimoine de chaque ménage de l'ensemble de répondants r est considérée comme un nombre au hasard (qui sera en fait compris dans l'intervalle qui est observé pour la valeur du patrimoine). De cette façon, l'estimation peut intégrer dans la largeur des intervalles de confiance de l'indice de Gini l'incertitude sur la valeur des patrimoines. Chaque valeur aléatoire dans l'intervalle déclaré par le ménage est, en quelque sorte, un scenario possible pour la vraie valeur mesurée de façon imprécise. Choisir tel ou tel processus de génération des données revient à choisir un modèle pour les scenarios de valeurs ponctuelles de patrimoines. Ce choix est critique et délicat. De manière générale, si un modèle possible appartient à une famille de modèles indexée par un paramètre θ, l'observation des intervalles déclarés peut permettre de choisir le « meilleur » paramètre (ou le « meilleur » modèle) au sens d'un critère statistique à définir. Dans la suite de l'article, les résultats d'estimation obtenus à partir de deux familles de modèles seront comparés et discutés.

Un modèle (PGD) est constitué d'une forme fonctionnelle reliant la valeur du patrimoine *pt* à celles d'autres variables X observées dans l'enquête, à un vecteur de paramètre θ et à un terme d'erreur aléatoire u :

$$pt = f(X,\theta,u)$$

Ainsi, du fait de l'observation de patrimoines en intervalles, les grandeurs $\hat{G}$ et $\hat{V}(\hat{G})$ sont doublement aléatoires : d'une part parce que l'échantillon de répondants est un ensemble aléatoire de la population totale (encadré 1) ; d'autre part parce que les valeurs des patrimoines $(pt_k)_{k\in r}$ sont désormais supposées aléatoires. Nous utilisons donc par la suite le modèle hiérarchique, c'est à dire l'emboîtement de modèles, suivant :

$$G = \hat{G}(pt_1,...,pt_r) + \sqrt{\hat{V}(\hat{G}(pt_1,...,pt_r))}\cdot\varepsilon \quad (3)$$

$$pt = f(X,\theta,u) \quad (4)$$

En statistique « fréquentiste », les valeurs de patrimoine (ici partiellement observées) sont issues d'un unique modèle, le modèle (4) pour une unique valeur de paramètre $\theta = \theta_0$. Les observations, ici des intervalles et des covariables, permettent en général, lorsque la taille de l'échantillon tend vers l'infini, de trouver θ_0. Par

3. L'indice de Gini est un nombre compris entre 0 et 1 qui correspond à 2 fois l'aire entre la première bissectrice et la courbe de Lorentz. Cette dernière représente la proportion du total du patrimoine des Français possédée par chaque pourcentage des ménages, ordonnés du plus pauvre au plus riche. La première bissectrice correspond au cas d'égalité complète.

contre, comme en pratique l'échantillon est de taille finie, le paramètre est connu avec erreur et on ne dispose que d'un estimateur de θ_0. En statistique bayésienne, on considère le paramètre θ comme une variable inobservable. De même que nous avons ajouté le modèle (4) au modèle (3), il convient de spécifier une distribution au hasard, dite « *a priori* » (cf. annexe 1), pour la variable inobservable θ. Nous introduisons dans ce cas un « troisième étage » à notre modèle hiérarchique (3)+(4) :

$$\theta \text{ est une variable aléatoire de loi } \pi(\theta) \qquad (5)$$

Le modèle (5) sera utilisé plus loin lorsque le modèle (4) est multivarié et ce pour des raisons pratiques (4).

Une des étapes de l'estimation consiste à simuler des valeurs ponctuelles du patrimoine (encadré 2). Les valeurs fictives $(pt_1^i,...,pt_I^i)$ sont « fabriquées » selon les hypothèses de modélisation et satisfont les contraintes propres à chaque ménage (les intervalles, les valeurs observées

4. L'estimation des résumés de la distribution des patrimoines des ménages repose sur des simulations (cf. plus loin) obtenues par échantillonnage de Gibbs (cf., par exemple, Arnold (1993), Robert (1995)). Adopter le point de vue bayésien permet de recourir exclusivement à de la simulation et ce suivant une suite d'étapes élémentaires. L'estimation de modèles à variables cachées (le patrimoine ou ses composantes) par maximum de vraisemblance (cf. par exemple Schafer (2001) et Train (2003)) est d'autant plus complexe que la dimension du modèle (nombre de composantes de patrimoine modélisées conjointement) augmente mais aussi que le domaine d'intégration est complexe.

Encadré 1

L'ALÉA DE SONDAGE

Étant donnée la distribution $(pt_k)_{k=1}^N$ des patrimoines totaux de l'ensemble des ménages français numérotés de 1 à N (taille de la population des ménages français), l'indice de Gini peut se calculer à l'aide de la formule suivante :

$$G = \frac{\sum_{k=1}^{N}(2r(k)-1)pt_k}{N\sum_{k=1}^{N}pt_k} - 1$$

où *r(k)* est le rang du patrimoine possédé par le k[ième] ménage. Après tirage de l'enquête nous ne disposons pas de tous les ménages mais d'un sous-ensemble $s \subset \{1,..,N\}$. L'ensemble *s* est tiré au hasard en respectant un plan de sondage préalablement spécifié. Chaque ménage de l'ensemble {1,…,*N*} est affecté d'une probabilité d'être sélectionné. À chaque ménage de l'échantillon *s* est alors associé un poids de sondage w_k égal à l'inverse de la probabilité de sélection. Un estimateur de *G* est donné par :

$$\hat{G} = \frac{\sum_{k\in s}(2\hat{r}(k)-1)w_k pt_k}{\sum_{k\in s}w_k \sum_{k\in s}w_k pt_k} - 1$$

où $\hat{r}(k) = \sum_{j\in s} w_j 1_{(pt_j \le pt_k)}$. Cet estimateur $\hat{G}$ est une grandeur aléatoire du fait que *s* est tiré au hasard dans {1,…,*N*} : si l'enquête était réalisée une seconde fois, d'autres ménages appartenant à un sous-échantillon s′ de {1,…,*N*} seraient interrogés et une autre valeur de $\hat{G}$ serait obtenue. On fournit alors un intervalle de confiance approché pour *G* en faisant une approximation normale $\hat{G} \approx G + \sqrt{\hat{V}(\hat{G})}\varepsilon$ où $\hat{V}(\hat{G})$ est une approximation de la variance asymptotique de $\hat{G}$ et ε est une variable aléatoire normale centrée réduite. Shao (1994) propose une justification théorique à cette approximation normale. Ce qui par inversion donne :

$$G \approx \hat{G} + \sqrt{\hat{V}(\hat{G})} \cdot \varepsilon .$$

À partir de cette approximation, un intervalle de confiance à 95 % pour *G* est donné par

$$\hat{G} - 1{,}96 * \sqrt{\hat{V}(\hat{G})} \le G \le \hat{G} + 1{,}96 * \sqrt{\hat{V}(\hat{G})}$$

L'estimateur $\hat{V}(\hat{G})$ de la variance asymptotique de $\hat{G}$ peut être obtenu de différentes façons. La méthode généralement utilisée à l'Insee repose sur la linéarisation puis la décomposition de la variance (Deville, 1999). Le calage sur marge permet d'utiliser des informations auxiliaires sur des totaux connus de certaines variables pour améliorer la précision des estimateurs (Deville et Särndal, 1992).

In fine, la procédure d'estimation d'intervalle de confiance de l'indice de Gini est la suivante :

- Linéariser les estimateurs ;

- Récupérer les résidus d'une régression des linéarisés sur les variables dont les totaux sont connus et qui sont utilisés pour le calage sur marge ;

- Calculer la variance d'un estimateur de total où les variables sont les résidus, par décomposition en éléments plus simples, permettant de tenir compte du plan de sondage.

Des formules pour linéariser les estimateurs des grandeurs étudiées ici sont données dans Dell *et al.* (2002). Dans le cadre de l'enquête *Patrimoine*, le plan de sondage est un plan de sondage en trois phases, stratifié et à probabilités inégales. Le fait que certains ménages n'aient pas répondu à l'enquête (non-réponse totale) conduit en fait l'échantillon de répondants *r* à être un ensemble aléatoire inclus dans le sous-échantillon initial *s*. Pour tenir compte de la non-réponse totale, les poids de sondage sont modifiés en supposant la non-réponse uniforme par groupes. Nous remplaçons dans la définition de $\hat{G}$, *s* par *r* et les poids w_k par les nouveaux poids corrigés $\ddot{w}_k$. La non-réponse totale est interprétée comme une phase supplémentaire. Plus de détails sur le plan de sondage et le traitement de la non-réponse totale sont disponibles sur le site internet de l'Insee *http://www.insee.fr/fr/themes/detail.asp?ref_id=fd-patri04.*

pour d'autres variables de l'enquête, etc.). Cette phase correspond à une imputation (5). Dans le problème d'estimation d'indices d'inégalité abordé dans l'article, il ne s'agit que d'une phase intermédiaire. Les imputations ne sont pas intéressantes en elles-mêmes et sont tirées d'un PGD adapté au problème et aux données. *In fine*, ce sont bien les estimations des « résumés » de la distribution des patrimoines des ménages, et plus particulièrement les intervalles de confiance, que nous cherchons à obtenir.

5. On remplace les données manquantes ou en intervalles par des valeurs artificielles.

Encadré 2

ESTIMATION, UNE APPROCHE PAR MÉTHODE DE MONTE-CARLO

Choix optimal des estimateurs

Une fois un modèle de type (3)+(4) ou (3)+(4)+(5) défini, l'estimateur ponctuel de l'indice de Gini est obtenu en minimisant le risque *a posteriori* suivant :

$$\hat{G} = \underset{G^*}{\operatorname{argmin}} E\left[(G^* - G)^2 \middle| observables \right]$$

où $\rho(G^*,G) = (G^* - G)^2$ est la fonction de perte (plus l'estimateur s'éloigne de la vraie valeur, plus la perte est importante). L'espérance $E[(G^* - G)^2|observables]$ est alors la perte moyenne sachant les observables, qu'on appelle le risque *a posteriori*. Ici, les observables sont les observations des variables X_k qui apparaissent dans le processus de génération des données, ainsi que des informations comme des intervalles pour le patrimoine total, pour des composantes, l'imposition ou non à l'ISF. Naturellement, il est possible de choisir des fonctions de perte très différentes, mais le choix fait ici est très courant et revêt un aspect très pratique puisque l'estimateur qui minimise le risque *a posteriori* est alors :

$$\hat{G} = E[G|observables] \quad (6)$$

Pour une estimation de l'indice de Gini sous forme d'intervalle de confiance de niveau α (en pratique α=95 % ou α=90 %), l'intervalle $\hat{I}_\alpha = [\hat{b}, \hat{h}]$ est tel que la probabilité que l'intervalle contienne l'indice soit égale à α :

$$P(G \in \hat{I}_\alpha | observables) = \alpha$$

Par simplicité, les bornes de cet intervalle ont été calculées de manière à ce que :

$$P(G \leq \hat{b} \middle| observables) = \frac{1-\alpha}{2} \text{ et}$$

$$P(G \geq \hat{h} \middle| observables) = \frac{1-\alpha}{2} \quad (7)$$

Approche par méthode de Monte-Carlo

Les deux problèmes (6) et (7) requièrent le calcul de moyennes théoriques (l'espérance $E[\ ]$ et les probabilités $P(\ )$). Une méthode de Monte-Carlo consiste à approcher la moyenne théorique par sa contrepartie empirique sur des scenarios tirés au hasard. On remplace donc les moyennes théoriques par des moyennes d'un nombre fini de valeurs simulées de la grandeur d'intérêt *G*. Les simulations de *G* sont obtenues en utilisant successivement les différents étages du modèle hiérarchique :

- modèle (4) sachant les observables, puis modèle (3) si le paramètre est connu, c'est-à-dire si on fait l'approximation que l'estimateur du paramètre est le vrai paramètre ;
- ou modèle (5) sachant les observables, puis modèle (4) sachant les observables, puis étage (1).

Les simulations peuvent être obtenues dans des lois indépendantes ou non. Si les tirages sont indépendants, la moyenne empirique approche bien la moyenne théorique par la loi des grands nombres. Néanmoins, produire des simulations exactement dans les lois souhaitées et indépendantes entre elles est parfois difficile, par exemple si l'on souhaite simuler conjointement plusieurs composantes de patrimoine, ce qui sera fait dans la suite. Une alternative consiste alors à simuler une trajectoire de chaine de Markov bien choisie, il s'agit des méthodes de Monte-Carlo par Chaînes de Markov (MCMC) (cf., par exemple, Robert et Casella, 2004). Dans le cas d'une méthode de Monte-Carlo avec tirages indépendants, par exemple, les valeurs simulées G^i s'obtiennent à partir de tirages aléatoires de patrimoines des répondants $(pt_1^i,...,pt_r^i)$ dans la loi du processus de génération des données (2ème étage du modèle), sachant les observables $(X_1,...,X_r)$, les intervalles sur le patrimoine total et éventuellement ses composantes et l'imposition à l'ISF, et sachant θ conditionnellement aux observables (modèle (5)) et des tirages aléatoires ε^i du terme d'erreur (modèle (3)).

Les moyennes théoriques dans (6) et (7) étant approchées par leur contrepartie empirique, l'estimateur de *G* est donné par :

$$\hat{G}_{B,T} = \frac{1}{T-B} \sum_{i=B}^{T} G^i ,$$

les bornes $\hat{b}_{B,T}$ et $\hat{h}_{B,T}$ de l'intervalle de confiance correspondent quant à elles aux quantiles empiriques à α/2 et (1-α)/2 des valeurs simulées, pour les simulations *B* à *T*. *T* peut être choisi arbitrairement grand, plus *T* est grand meilleur est l'approximation des moyennes théoriques. La valeur *B* est appelée *burn-in* dans la littérature MCMC. *B* est nul dans le cas de tirages indépendants. Lorsque l'on utilise une méthode MCMC, on prend souvent B suffisamment grand afin de commencer le calcul de la moyenne empirique lorsque la chaîne de Markov s'est stabilisée proche de l'équilibre.

Contrairement à l'approche par imputation aléatoire simple mise en œuvre à l'Insee depuis de nombreuses années (cf., par exemple, Lollivier et Verger, 1987), nous procédons à des imputations multiples (6). Ces imputations multiples permettent de fournir une estimation optimale (au sens où elle minimise le risque *a posteriori*) et d' obtenir des intervalles de confiance tenant compte de l'aléa de sondage, de la non-réponse totale par la troisième phase, de la réduction de variance par calage, et de l'incertitude sur les valeurs des patrimoines (intervalles et connaissance imprécise du paramètre).

La suite de l'article présente deux modélisations possibles du processus de génération des données PGD 1 et PGD 2, en distinguant pour le modèle PGD 2 un ensemble d'observables qui inclut l'information sur l'imposition ou non à l'ISF et un autre qui l'exclut. La loi « a priori » et l'algorithme de simulation dans le cadre du processus de génération des données PGD 2 ne sont que brièvement présentés. Gautier (2008) développe de manière plus détaillée ces aspects.

Le choix d'un processus de génération des données : d'un modèle univarié à un modèle multivarié

Le patrimoine des ménages peut être modélisé de plusieurs façons. Un point de départ naturel pour le PGD est de modéliser une unique composante de patrimoine - en fait, directement le patrimoine total - sous la forme d'une relation linéaire entre le logarithme du patrimoine, des variables observables (comme la position dans le cycle de vie, les niveaux d'étude et de revenus, le fait d'avoir reçu une donation ou un héritage ou d'avoir transmis des biens en donation) et un résidu de loi normale. De tels modèles sont depuis longtemps utilisés à l'Insee (Lollivier et Verger, 1987) pour imputer des valeurs ponctuelles de patrimoine ou de revenu en présence de non-réponse ou de réponse en intervalles. Le choix d'une forme log-normale est souvent bien adapté pour décrire la loi du revenu ou du salaire conditionnelle à des observables. Cependant, le patrimoine en général et certaines de ses composantes en particulier, comme les valeurs mobilières ou les actifs immobiliers de rapport, ont des distributions nettement plus concentrées que celles des revenus ou des salaires. Des lois de Pareto pourraient être mieux adaptées pour décrire la distribution du patrimoine (conditionnelle aux observables), au moins pour les ménages aisés. D'autres travaux (Avery *et al.*, 1988) retiennent par exemple une forme log-normale en deçà du patrimoine médian et une forme parétienne au-delà. Même si les sources fiscales sur l'impôt sur la fortune ou sur les actifs successoraux peuvent fournir des indications intéressantes sur la forme de la queue de distribution, elles ne sont pas disponibles et ne correspondent pas non plus au concept de patrimoine brut total que nous étudions ici. Il est clair que l'hypothèse de log-normalité a une incidence sur les résultats et il est envisageable que l'utilisation de lois de Pareto pourrait conduire à des indices d'inégalité plus élevés. Ceci pourrait faire l'objet de travaux ultérieurs (7). Le modèle retenu, noté PGD 1, distingue les propriétaires de leur résidence principale et les non-propriétaires afin de tenir compte de l'hétérogénéité de ces deux sous-populations, et introduit des variables explicatives standards dans l'analyse de l'accumulation patrimoniale (tableau 3).

Travailler sur une unique variable a l'avantage de la simplicité. Cependant, en travaillant directement sur le patrimoine total, il n'est pas possible d'utiliser au mieux l'information de l'enquête et l'information auxiliaire sur l'imposition ou non des ménages à l'ISF. D'une part, l'enquête contient des informations très détaillées sur chaque actif patrimonial, par exemple pour les contrats d'assurance-vie : l'année de souscription et les versements annuels, pour la résidence principale : la surface, etc. Le PGD sur la variable récapitulative seule n'intègre pas ces informations dans les variables observables. Seules des observables relatives au ménage sont utilisées. D'autre part, en l'absence de décomposition adéquate du patrimoine, le modèle ne permet pas d'intégrer les intervalles construits en utilisant l'information auxiliaire sur l'ISF. Ceci conduit, comme on le verra, à des intervalles de confiance plus larges pour les indices d'inégalité. Quitte à travailler avec des composantes plus fines, il parait intéressant d'envisager de modéliser directement chaque composante, conjointement aux autres composantes, afin d'obtenir un modèle pour le patrimoine total comme somme de ses composantes.

---

*6. Néanmoins, notre méthode n'exige pas.de produire des imputations bayésiennes propres et ne repose pas sur des formules pour combiner les variances (*cf. *Little et Rubin, 2002).*

7. L'hypothèse de log-normalité pose par ailleurs un autre problème lié à la sélection de notre échantillon. L'échantillon surreprésente les catégories aisées et la sélection est donc volontairement liée à la variable d'intérêt. Néanmoins, en dehors de la non-réponse totale que l'on connait mal, le tirage à probabilités inégales correspond à une sélection exogène car nous disposons des variables ayant servi à cette surreprésentation. Nous avons donc veillé à inclure parmi les régresseurs ces variables quand elles étaient significatives. Cela est essentiel car si l'hypothèse de log-normalité de la loi conditionnelle du patrimoine des français est crédible, celle des ménages sélectionnés de façon endogène pourrait ne pas l'être.

Le processus de génération des données PDG 2 comprend cinq composantes de patrimoine : les actifs financiers (FIN), la résidence principale (RP), les autres actifs immobiliers (ALG), les actifs professionnels (PROF) et les autres éléments de patrimoine comme les biens durables, bijoux, objets d'art ou de valeur (RESTE). La prise en compte des comportements de détention d'actifs est obtenue en spécifiant autant de modèles que de combinaisons de détention possibles de ces cinq composantes. En réalité, seules trois composantes sont susceptibles d'être ou non détenues, tous les ménages étant supposés détenir des actifs financiers (les compte-chèques en font partie) et des éléments de patrimoine tels que des biens durables ou d'autres objets de valeur. Il existe alors huit portefeuilles différents (tableau 4) et, pour chaque portefeuille, un modèle à équations simultanées est spécifié, avec autant d'équations que de composantes détenues dans le portefeuille considéré. Les systèmes portent sur le logarithme des montants, la moyenne est linéaire en les variables explicatives, le vecteur des résidus suit une loi normale de dimension correspondant au nombre de composantes détenues et a une matrice de variance-covariance quelconque. Par souci de parcimonie (8),

8. Augmenter le nombre de paramètres, à taille d'échantillon fixée, augmente l'incertitude sur la valeur des paramètres. La parcimonie correspond à arbitrer entre la flexibilité du modèle et la taille des intervalles de confiance. Ceci est relié au célèbre arbitrage en statistique entre le biais et la variance.

Tableau 3
Variables explicatives retenues pour le modèle univarié

Variables explicatives \ Groupe	Propriétaires	Non propriétaires
Position dans le cycle de vie	X	X
Niveau socio-professionnel	X	X
Niveau de diplôme de la personne de référence	X	X
Ressources économiques	X	X
Niveau de revenus	X	X
Perception d'aides sociales	X	X
Perception d'une rente	X	X
Perception de revenus autres que revenus d'activité ou de remplacement	X	X
Zone géographique	X	X
Histoire du patrimoine	X	X
Existence d'une donation reçue	X	X
Existence d'une donation versée	X	X
Existence d'une aide reçue	X	X
Existence d'une aide versée	X	X
Décès des deux parents	X	X

Source : auteurs.

Tableau 4
Fréquence des portefeuilles dans l'échantillon

Portefeuille	Actifs financiers (FIN)	Résidence principale (RP)	Autres actifs immobiliers (ALG)	Actifs professionnels (PROF)	Autres éléments de patrimoine (1) (RESTE)	Pourcentage de ménages disposant du portefeuille dans l'échantillon (en %)
1	x				X	32,8
2	X	X			X	33,8
3	X		X		X	3,5
4	x			X	X	2,8
5	X	X	X		X	10,2
6	X	X		X	X	8,6
7	X		X	X	X	1,5
8	X	X	X	X	X	6,8

1. Autres éléments de patrimoine comme les biens durables, bijoux, objets d'art ou de valeur.

Lecture : parmi les ménages constituant l'échantillon de l'enquête, 32,8 % possèdent un patrimoine composé uniquement d'épargne financière et de biens durables ou objets de valeurs.
Champ : ménages interrogés dans l'enquête Patrimoine 2004.
Source : enquête Patrimoine *2004, Insee.*

nous faisons l'hypothèse que les coefficients apparaissant dans les moyennes sont constants quel que soit le portefeuille. Par contre, nous introduisons des indicatrices du type de portefeuille. Ainsi, pour chaque portefeuille, les constantes des logarithmes des composantes sont différentes, tout comme les matrices de variance-covariance. Néanmoins, les autres coefficients restent égaux d'un portefeuille à l'autre (cf. encadré 3 pour une version simplifiée avec uniquement deux composantes).

De nombreuses variables observées dans l'enquête sont introduites pour expliquer les différentes composantes de patrimoine. Pour la résidence principale, il a été possible d'introduire des caractéristiques propres du bien immobilier telles que la surface. En revanche, pour les autres composantes, ceci n'était pas possible puisque ce sont des composantes agrégées. Les variables explicatives retenues (tableau 5) comprennent des critères de position dans le cycle de vie (âge, composition familiale et interactions entre ces variables), des variables sur les ressources culturelles et économiques du ménage (niveau d'études, niveau de revenu d'activité, perception de revenu complémentaire comme des aides sociales, ou des revenus de remplacement), une variable de localisation et enfin quelques informations sur la trajectoire du patrimoine du ménage (existence de donation reçue ou versée, évolution récente du patrimoine, composition du patrimoine des parents). Toujours par souci de parcimonie, seul les régresseurs qui étaient significatifs ont été inclus dans le modèle.

La structure de variance-covariance des résidus des composantes détenues (entre 3 et 5) est la

Encadré 3

LE PROCESSUS DE GÉNÉRATIONS DES DONNÉES POUR DEUX COMPOSANTES

Pour simplifier, l'encadré décrit le modèle retenu pour le processus de générations des données (PGD) de patrimoine dans le cas où le patrimoine total est la somme de deux composantes uniquement. Le PGD est proche de celui spécifié dans par Heeringa *et al.* (2002).

Soit $y_k^* = \left(y_{k1}^*, y_{k2}^*\right)$ le vecteur des deux composantes patrimoniales en question pour le ménage k et $D_k = (D_{k1}, D_{k2})$ le vecteur des indicatrices de détention de ces composantes, qui définit les composantes présentes ($D_{ki} = 1$ si $y_{ki}^* > 0$) ou absentes ($D_{ki} = 0$ si $y_{kj}^* = 0$) du portefeuille du ménage. La détention est supposée être parfaitement observée, c'est-à-dire que tous les ménages déclarent détenir ou non chacune des composantes. Il y a donc quatre types de portefeuilles, qu'on indexe par $\{p\}$.

Soit $INF_k = (INF_{1k}, INF_{2k})$ et $SUP_k = (SUP_{1k}, SUP_{2k})$ les bornes inférieures et supérieures observées des logarithmes des valeurs des deux composantes. Plusieurs cas se présentent pour la composante j dont :

$INF_{jk} = SUP_{jk}$:	l'observation est une valeur exacte,
$INF_{jk} = \infty$ et $SUP_{jk} = +\infty$:	l'observation est une valeur complètement manquante,
$INF_{jk} = \infty$ et $SUP_{jk} < +\infty$:	l'observation est un intervalle non minoré,
$INF_{jk} > \infty$ et $SUP_{jk} = +\infty$:	l'observation est un intervalle non majoré,
$INF_{jk} > \infty$ et $SUP_{jk} < +\infty$:	l'observation est un intervalle borné.

Les lois des encours y_k^* conditionnelles à des régresseurs sont modélisées sous forme log-normale. On définit des variables $z_k = (z_{k1}, z_{k2})$ observées partiellement telles que :

$$\left(y_{1k}^*, y_{2k}^*\right) = \begin{cases} (\exp(z_{1k}), \exp(z_{2k})) & si \quad D_k = (1,1) \\ (\exp(z_{1k}), 0) & si \quad D_k = (1,0) \\ (0, \exp(z_{2k})) & si \quad D_k = (0,1) \\ (0,0) & si \quad D_k = (0,0) \end{cases}$$

Le portefeuille de type $p = p_1$ comprend deux composantes modélisées de la manière suivante :

$$\begin{cases} z_{1k} = \alpha_{1p_1} + x_{1k}' \cdot \beta_1 + \varepsilon_{1p_1 k} \\ z_{2k} = \alpha_{2p_1} + x_{2k}' \cdot \beta_2 + \varepsilon_{2p_1 k} \end{cases}$$

où $\left(\varepsilon_{1p_1}, \varepsilon_{2p_1}\right)'$ suit une loi normale de dimension 2, centrée, et de matrice de variance-covariance Σ_{p_1}. Pour le portefeuille de type $p = p_2$ nous avons l'équation $z_{1k} = \alpha_{1p_2} + x_{1k}' \cdot \beta_1 + \varepsilon_{1p_2 k}$ et pour $p = p_3$ nous avons $z_{2k} = \alpha_{2p_3} + x_{2k}' \cdot \beta_2 + \varepsilon_{2p_3 k}$ où ε_{1p_2} et ε_{2p_3} sont de loi normale centrée, indépendantes entre elles et indépendantes de résidus pour $p = p_1$, de variances $\sigma_{p_2}^2$ et $\sigma_{p_3}^2$. Le modèle PGD 2 se limite à cinq composantes et permet l'utilisation de l'information sur l'imposition à l'ISF. Il serait possible de retenir une décomposition plus fine en réduisant le nombre de variables explicatives ou, de manière voisine à Heeringa *et al.* (2002), en faisant des hypothèses plus fortes sur la structure de variance-covariance.

plus générale possible : nous autorisons des corrélations entre les résidus des différentes composantes de patrimoine détenues et les corrélations dépendent du type de portefeuille. Ceci permet d'introduire un PGD relativement peu contraignant, évitant ainsi de faire trop d'hypothèses de structure qui pourraient ne pas être justifiées et entraîner des biais dans l'estimation des inégalités. En contrepartie de cette généralité sur la structure de variance-covariance, le nombre de paramètres devient élevé et conduit, par souci de parcimonie, à n'introduire que des effets fixes propres à chaque portefeuille dans les moyennes. Une structure de covariance générale semble pertinente au vu de résultats classiques dans l'analyse des comportements de patrimoine. En effet, les caractéristiques observables dans les enquêtes n'expliquent au mieux que la moitié des variations de niveau de patrimoine entre ménages. Les enquêtes plus récentes, comme l'enquête *Patrimoine 1998*, permettent de tenir compte de mesures directes de préférences individuelles comme l'horizon temporel ou l'aversion pour le risque. Cependant, ces enquêtes ne permettent pas de réduire très significativement la part de variance non expliquée (9). Ainsi, même conditionnellement aux autres observables de l'enquête, l'existence de cette forte hétérogénéité inobservée (capacités cognitives, éducation financière, esprit d'entreprise, etc.) est une source de corrélation potentielle entre les composantes patrimoniales, qui ne peut être prise en compte qu'en choisissant un PGD multivarié. D'autres données, comme celles de la Direction Générale des Finances Publiques sur les actifs successoraux, mettent en évidence une relation assez forte entre le niveau de patrimoine et sa structure (tableau 6). Une analyse sur des tranches plus élevées d'actifs successoraux montrerait que la part de l'immobilier de rapport et des valeurs mobilières augmente encore sensiblement avec le niveau de patrimoine au-delà de 600 000 euros. La structure du patrimoine peut s'expliquer à l'aide

9. Arrondel, Masson et Verger (2005b) montrent que l'inclusion des ces paramètres de préférences dans les régressions augmente le R^2 de 0,12, mais près de 45 % de la variance reste malgré tout inexpliquée.

Tableau 5
Variables explicatives retenues pour chacune des composantes de patrimoine

Variables explicatives \ Composantes	Actifs financiers (FIN)	Résidence principale (RP)	Autres actifs immobiliers (ALG)	Actifs professionnels (PROF)	Autres éléments de patrimoine (1) (RESTE)
Cycle de vie					
Seul et sans enfant		X	X	X	X
Âge et âge au carré		X	X	X	X
Position dans le cycle de vie	X				
Ressources culturelles et sociales					
Niveau social	X	X	X	X	X
Niveau d'études	X	X	X	X	X
Ressources économiques					
Niveau de revenus	X	X	X	X	X
Perception d'aides sociales	X				
Perception d'une rente	X	X		X	
Perception de revenus autres que revenus d'activité ou de remplacement	X		X	X	
Zone géographique	X	X	X		X
Histoire du patrimoine					
Existence d'une donation reçue	X	X	X		X
Existence d'une donation versée	X				
Augmentation ou baisse récente du patrimoine	X	X		X	X
Composition patrimoine des parents	X		X	X	
Caractéristiques du produit					
Surface et surface au carré		X			
Patrimoine professionnel					
Existence de patrimoine professionnel				X	
Possession d'une entreprise				X	

1. Autres éléments de patrimoine comme les biens durables, bijoux, objets d'art ou de valeur.

Source : auteurs.

de caractéristiques observables comme l'âge, les revenus, etc. Cependant, de manière analogue à ce qui se passe pour l'explication du niveau de patrimoine, il est également vraisemblable que ces caractéristiques n'expliquent qu'une partie des corrélations entre les composantes. Il est donc justifié de tenir compte de corrélations entre les résidus des composantes (10).

Comparaison des résultats en fonction du type de PGD considéré et des observables utilisés dans le conditionnement

L'estimation ponctuelle du patrimoine moyen des français fin 2003 varie entre 200 000 et 230 000 euros suivant le modèle retenu (tableaux 7 et 8). L'estimation à partir du modèle PGD 1 repose sur un modèle hiérarchique du type (3)+(4), donc sous l'hypothèse que le paramètre estimé est le vrai paramètre, ce qui sous-estime la couverture réelle des intervalles de confiance (11) (tableau 7). Au contraire, les estimations reposant sur le PGD 2

10. Il est possible de tester plus rigoureusement l'hypothèse d'indépendance entre les lois de différents actifs patrimoniaux, conditionnelles aux observables à partir des tests utilisant les résidus généralisés (cf. Gouriéroux et al.*, 1987). Ces tests sont plus simples à mettre en œuvre que ceux basés sur l'estimation d'un modèle multivarié à variables latentes. En effet, ils ne requièrent que l'estimation de modèles univariés à variables latentes. Gautier et Houdré (2008) montrent que sur la détention seule, l'hypothèse d'indépendance est rejetée.*

11. Par ailleurs, dans les résultats présentés ci-dessous, les calculs des approximations des variances asymptotiques des différents estimateurs de sondage sont obtenus sans tenir compte du calage sur marge, et en approximant le plan complexe par un plan à probabilités inégales et entropie maximale comme dans Dell et al. *(2002) en utilisant une formule analytique due à Deville. Cette simplification a tendance à surestimer la largeur des intervalles de confiance cette fois. Des intervalles tenant compte intégralement de la procédure présentée dans l'encadré 1 sont obtenus dans Gautier (2008)*

Tableau 6
Composition des actifs successoraux en fonction de leur valeur en 2006

En %

Montant de l'actif brut de succession :	Résidence principale	Autres immobiliers (1)	Liquidités	Autres mobiliers (2)	Passif
Moins de 60 000 €	16	9	64	11	- 11
De 60 000 à 120 000 €	35	13	42	10	- 5
De 120 000 à 180 000 €	51	11	28	9	- 3
De 180 000 à 300 000 €	55	12	22	12	- 3
De 300 000 à 600 000 €	50	17	20	13	- 3
Plus de 600 000 €	30	25	14	31	- 5
Ensemble	**45**	**17**	**22**	**17**	**- 4**

1. Foncier bâti et non bâti, immobilier de rapport, immobilier professionnel.
2. Valeurs mobilières y compris bons, créances, fonds de commerce, meubles meublant, bijoux, etc.

Lecture : en 2006, la valeur des résidences principales constitue en moyenne 16 % de la valeur des actifs transmis dans les successions de montant brut inférieur à 60 000 €.
Champ : successions déclarées à l'administration fiscale en 2006.
Source : Azoulay (2008), enquête Mutations à titre gratuit 2006*.*

Tableau 7
Estimation d'indices d'inégalités de patrimoine PGD 1

Caractéristique de la distribution de patrimoine	Estimation ponctuelle	Intervalles de confiance à 95 %	
		Borne inférieure	Borne supérieure
Quantiles (1)			
P99	1 624 800	1 402 900	1 889 800
P95	619 970	587100	654 580
P90	412 820	399 930	427 970
Moyenne	204 550	194 080	217 610
Médiane	112 950	108 210	117 850
P10	4 370	3 160	5 660
D9/D5	3,66	3,49	3,84
Gini	0,652	0,632	0,676
Theil	0,968	0,852	1,000

1. En euros.

Lecture : sous l'hypothèse du modèle PGD 1, un estimateur de sondage de l'indice de Gini du patrimoine des français vaut 0,652. La vraie valeur de cet indice se trouve à 95 % entre 0,632 et 0,676. Les calculs sont approchés et utilisent T=1 000 valeurs simulées.
Champ : ensemble des ménages de France métropolitaine.
Source : enquête Patrimoine *2004, Insee, calculs des auteurs.*

ont été réalisées avec un modèle hiérarchique à trois étages et les intervalles de confiance présentés (tableau 8) tiennent compte également de l'incertitude sur la vraie valeur du paramètre. Dans le cadre du PGD 2, nous distinguons en PGD 2a les résultats obtenus sans tenir compte dans les observables de l'imposition ou non à l'ISF et en PGD 2b où nous tenons compte de cette dernière information. En PGD 2a et en PGD 2b, des informations externes sur la distribution du patrimoine ont également servi à majorer artificiellement le niveau de patrimoine total. Heeringa *et al.* (2002) constatent que cela a relativement peu d'effet si les bornes sont lâches. On s'attend néanmoins à une légère sous-estimation des intervalles de confiance. Les poids de sondage varient entre 450 et 12 000. Nous avons donc limité la possibilité que le patrimoine simulé dépasse une valeur au-delà de laquelle le principe de représentativité soit mis en défaut. Sur la base de Cordier *et al.* (2006b) et d'informations publiques sur les plus grosses fortunes professionnelles françaises (12), nous avons introduit des plafonds relativement généreux sur les composantes de patrimoine. Au final, le patrimoine du ménage apparemment le plus fortuné a été plafonné à 50 millions d'euros et celui des autres ménages à 10 millions d'euros. L'avantage principal de l'introduction de ces bornes est de permettre une convergence exponentielle de l'échantillonnage de Gibbs utilisé pour la simulation (ceci peut se prouver avec les mêmes arguments que dans Roberts et Polson, 1994). Des représentations graphiques (cf. annexe 2) montrent qu'une valeur d'arrêt de la procédure à 10 000 simulations est très satisfaisante (13).

Les estimations des trois modèles sont cohérentes au sens où les intervalles de confiance à 95 % se recoupent pour la plupart des caractéristiques de la distribution, excepté les estimations des quantiles P95 et P90 pour lesquels les modèles univarié PGD 1 et multivarié PGD 2 avec dans l'ensemble d'observables l'information sur l'imposition ou non à l'ISF (PGD 2b dans le tableau 8) conduisent à des intervalles de confiance d'intersection vide. Les intervalles de confiance obtenus en PGD 1 et en PGD 2b sont de largeur tout à fait comparable (et même légèrement inférieure pour le modèle multivarié pour l'estimation de l'indice de Gini) et nous avons vu que les intervalles présentés en PGD 1 sont en réalité sous-estimés. Ceci accrédite le fait que la prise en compte d'information supplémentaire sur les composantes et l'ISF permet de réduire la longueur des intervalles de confiance.

Les indices de Gini varient entre 0,65 et 0,68 (tableaux 7 et 8), ce qui constitue une différence

12. Le magasine Challenges *publie chaque année un classement des 500 plus grosses fortunes professionnelles de France.*
13. En pratique, il semble qu'un nombre de simulations de l'ordre du millier aurait pu suffire.

Tableau 8
Estimation d'indices d'inégalités de patrimoine PGD 2

Ensemble d'information	Résultats obtenus sans tenir compte dans les observables de l'imposition à l'ISF **(PGD 2a)**			Résultats obtenus en tenant compte dans les observables de l'imposition à l'ISF **(PGD 2b)**		
Caractéristique de la distribution de patrimoine	Estimation ponctuelle	Intervalles de confiance à 95 %		Estimation ponctuelle	Intervalles de confiance à 95 %	
		Borne inférieure	Borne supérieure		Borne inférieure	Borne supérieure
Quantiles (1)						
P99	2 560 520	1 820 080	3 315 320	1 584 600	1 359 260	1 825 360
P95	760 320	682 820	838 160	690 790	636 920	746 760
P90	454 110	430 190	482 480	434 460	416 410	452 000
Moyenne	231 030	210 760	250 850	205 000	192 880	217 650
Médiane	113 360	107 310	119 490	111 460	105 670	117 560
P10	3 970	2 760	5 220	3 960	2 870	5 070
D9/D5	4,00	3,78	4,24	3,90	3,71	4,09
Gini	0,681	0,655	0,704	0,652	0,633	0,672
Theil	1,000	0,91	1,000	0,904	0,814	1,000

1. En euros.

Lecture : sous l'hypothèse du modèle PGD 2b, un estimateur de sondage de l'indice de Gini du patrimoine des français vaut 0,652. La vraie valeur de cet indice se trouve à 95 % entre 0,632 et 0,676. Les calculs sont approchés et utilisent T = 10 000 valeurs simulées et un burn-in des B = 1 000 premières valeurs simulées.
Champ : ensemble des ménages de France métropolitaine.
Source : enquête Patrimoine *2004, Insee, calculs des auteurs.*

significative en termes d'inégalités, l'indice étant relativement peu élastique. Toutefois, la prise en compte de plusieurs composantes de patrimoine et de leur corrélation résiduelle ne bouleverse pas la mesure des inégalités à partir des données de l'enquête *Patrimoine*. Si, par contre, on compare ces résultats avec ceux présentés dans Cordier *et al* (2006a), on se rend compte que si l'on n'autorise pas la corrélation des résidus dans le PGD, nous pouvons obtenir des résultats assez différents.

* *
*

L'estimation présentée dans cet article repose sur la simulation. Les simulations, notamment multivariées, pourraient avoir une utilisation plus large que celle présentée dans cet article. En effet, la présence de données en intervalles complique très fortement les analyses statistiques qu'il est possible de mener sur l'enquête Patrimoine. Les utilisateurs finaux des enquêtes réalisées par l'Insee ne disposent parfois pas des informations nécessaires pour traiter correctement de telles données. Par conséquent, certains considèrent qu'il est du ressort du producteur de données de fournir aux utilisateurs des données complètes ou complétées par imputation sur lesquelles ces derniers peuvent mener des analyses statistiques comme ils le feraient sur des données complètes. Cela n'est pourtant pas très aisé. En effet, l'inférence doit absolument tenir compte de l'incertitude initiale liée aux observations en intervalles (pour des tests d'hypothèses ou le calcul d'intervalles de confiance). Une façon de faire est de fournir plusieurs jeux de données complétées par simulations, des imputations multiples (Little et Rubin, 2002), et de faire autant d'analyses que de jeux de données, ces analyses devant in fine être recombinées. En effet, il n'est pas possible, en utilisant directement des procédures statistiques sur données complétées par imputation aléatoire simple comme si c'était des vraies mesures, de produire des intervalles de confiance honnêtes ou de se fier aux tests qui sont erronés, par exemple la significativité de variables dans un modèle de régression. Mais aussi, on espère grâce aux imputations, pouvoir étudier des questions diverses et variées. Cela nécessiterait de disposer d'imputations issues d'un PGD bien spécifié, au niveau de détail le plus fin et très souple, c'est-à-dire avec le minimum de structure non justifiable. Cette généralité est tout à fait illusoire tout d'abord pour des raisons d'identification.

Dans cet article, nous avons répondu, du mieux que nous avons pu, à une question spécifique: déterminer des intervalles pour le niveau des inégalités de patrimoine total. Nous avons pu spécifier un modèle souple autorisant la dépendance conditionnelle des modèles pour les différentes composantes intervenant dans le calcul du patrimoine total. Par là même, cela nous a permis de considérer conjointement les composantes de patrimoine d'un même ménage et de les additionner pour reconstituer le patrimoine total. Par ailleurs, nous avons pu nous satisfaire d'un niveau de détail relativement fruste mais suffisant pour exploiter de l'information auxiliaire comme l'imposition à l'ISF. Le niveau de détail du modèle PGD 2 peut être insuffisant pour une autre utilisation que celle de l'étude des inégalités de patrimoine total. Considérer un plus grand nombre de composantes tout en autorisant une structure souple de covariance des résidus requiert, *par souci de parcimonie*, que l'on se restreigne à un petit nombre de variables explicatives. Mais omettre certaines variables entraîne des biais de sélection. D'autre part, le PGD devrait inclure toutes les variables explicatives qui sont susceptibles d'être utilisées dans des analyses statistiques ultérieures. Ainsi, imputer à un niveau de détail fin augmente le risque de biais dans les résultats des analyses sur données complétées. L'imputation des données d'enquête manquantes et en intervalles a porté sur le niveau de détail le plus fin des composantes. En contrepartie, une hypothèse forte d'indépendance conditionnelle a dû être faite. Autoriser la corrélation des résidus était impossible car nous aurions eu beaucoup trop de paramètres relativement à la taille de l'échantillon. Mais aussi, les modèles sur les montants n'ont pas pu tenir compte de la composition des portefeuilles. Il faut donc manipuler ces imputations avec beaucoup de précaution et être conscient des limites en fonction de l'étude considérée. □

BIBLIOGRAPHIE

Arnold S. F. (1993), « Gibbs sampling », *Handbook of Statistics*, n° 9, pp. 599-625.

Arrondel L., Masson A. et Verger D. (2005a), « Les comportements de l'épargnant à l'égard du risque et du temps », *Économie et Statistique*, n° 374-375, pp. 9-19.

Arrondel L., Masson A. et Verger D. (2005b), « Préférences individuelles et disparités de patrimoine », *Économie et Statistique*, n° 374-375, pp. 129-157.

Avery R. B., Elliehausen G. E. et Kennickell A. B. (1988), « Measuring Wealth in Survey Data: an Evaluation of the 1983 Survey of Consumer Finances », *Review of Income and Wealth*, vol. 34, n° 4, pp. 339-369.

Azoulay J. (2008), « Les transferts intergénérationnels familiaux », dans *La répartition des prélèvements obligatoires entre générations et la question de l'équité intergénérationnelle*, Rapport du Conseil des Prélèvements Obligatoires.

Cordier M., Gautier E. et Houdré C. (2006a), « Simulation et mesure des inégalités de patrimoine en 2004 », communication au séminaire de la Direction des Statistiques Démographiques et Sociales, Insee, janvier 2006.

Cordier M., Houdré C. et Rougerie C. (2006b), « Les inégalités de patrimoine des ménages entre 1992 et 2004 », dans *Insee Références - Les revenus et le patrimoine des ménages*, pp. 47-58.

Dell F., d'Haultfoeuille X., Février P. et Massé E. (2002), « Mise en œuvre de calcul de variance par linéarisation », *Actes des Journées de Méthodologie Statistique*, http://jms.insee.fr/site/index.php.

Deville J. C. (1999), « Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques », *Survey Methodology*, vol. 25, n° 2, pp. 193-203.

Deville J. C. et Särndal C. E. (1992), « Calibration Estimators in Survey Sampling », *Journal of the American Statistical Association*, vol. 87, n° 418, pp. 376-382.

Gautier E. (2008), « Bayesian Estimation of Inequalities with Non-rectangular Censored Survey Data », téléchargeable sur le site arXiv.org : arXiv:0802.4190.

Gautier E. et Houdré C. (2008), « Approche multivariée de l'estimation des inégalités dans l'enquête Patrimoine 2004 », *Document de Travail de la Direction des Statistiques Démographiques et Sociales*, n° F0801, Insee, http://www.insee.fr/fr/publications-et-services/docs_doc_travail/f0801.pdf.

Gourieroux C., Monfort A., Renault E. et Trognon A. (1987), « Generalized Residuals », *Journal of Econometrics*, vol. 34, n° 1-2, pp. 5-32.

Heeringa S. G., Little R. J. A. et Raghunathan T. E. (2002), « Multivariate Imputation of Coarsened Survey Data on Household Wealth » dans *Survey Nonresponse*, edité par Robert M. Groves et al., pp. 357-372, Wiley, New York.

Juster T. F. et Smith J. P. (1997), « Improving the Quality of Economic Data: Lessons from the HRS and AHEAD », *Journal of the American Statistical Association*, vol. 92, n° 440, pp. 1268-1278.

Little R. J. A. et Rubin D. B. (2002), « Statistical Analysis with Missing Data », 2ème édition, Wiley, New York.

Lollivier S. et Verger D. (1987), « D'une variable discrète à une variable continue : la technique des résidus simulés » dans *Mélanges économiques - Essais en l'honneur de Edmond Malinvaud*, Economica, Paris.

Piketty T., Postel-Vinay G. et Rosenthal J.-L. (2006), « Wealth Concentration in a Developing Economy: Paris and France, 1807-1994 », *American Economic Review*, vol. 96, n° 1, pp. 236-256.

Robert C. P. (1995), « Simulation of Truncated Normal Variables », *Statistics and Computing*, vol. 5, n° 2, pp. 121-125.

Robert C. P. (2007), *The Bayesian Choice*, Springer, New York.

Robert C. P. et Casella G.(2004), *Monte Carlo Statistical Methods*, 2ème édition, Springer, New York.

Roberts G. O. et Polson N. G. (1994), « On the Geometric Convergence of the Gibbs Sampler », *Journal of the Royal Statistical Society – B*, n° 56.

Schafer J. L. (2001), *Analysis of Incomplete Multivariate Data*, 2ème édition, Chapman & Hall, Londres.

Shao J. (1994), « L-statistics in complex survey problems », *Annals of Statistics*, vol. 22, n° 2, pp. 946-967.

Train K. E. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.

ANNEXE 1

LOI A PRIORI

Nous détaillons dans cette partie la loi *a priori* utilisée dans le cadre du modèle hiérarchique avec le modèle PGD 2. Nous prenons, pour la loi $\pi(\theta)$, la mesure produit d'une mesure de Lebesgue pour le vecteur des paramètres « β », *i.e.* ceux apparaissant dans la moyenne des logarithmes des composantes de patrimoine (incluant les constantes spécifiques par portefeuille de l'encadré 3), et

de la mesure de probabilité de densité proportionnelle à

$\prod_{i=1}^{8} \det(\Sigma_i)^{-(d_i+1)/2}$ pour les matrices de variances-covariances des huit portefeuilles. d_i correspond au nombre de composantes détenues dans le portefeuille *i*. La loi *a priori* est telle que nous ne faisons aucune hypothèse sur les paramètres « β ». Les matrices de variance-covariance sont par ailleurs *a priori* indépendantes des coefficients « β » et mutuellement indépendantes. Dans le cas où il n'y aurait eu qu'un seul portefeuille, cette loi *a priori* est très usuelle dans les modèles linéaires Gaussiens multivariés (Little et Rubin, 2002 ; Schafer, 2001). Elle est parfois dite non-informative ou de loi de Jeffreys. Il s'agit d'une limite de lois normale/inverse-Wishart, lorsque $\tau \to 0$, $m \to -1$, $\Lambda_1^{-1} \to 0$, définie par :

- $\beta = (\beta_{1,1},\ldots,\beta_{1,5})$ suit une loi normale $N(\beta_0,\tau^{-1}\sum_1)$
- la loi de $\sum_1$ sachant β est inverse-Wishart $W^{-1}(m,\Lambda_1)$.

La loi conditionnelle sachant les variables explicatives et les composantes de patrimoine est bien une normale/inverse-Wishart. Ainsi, bien que $\pi(\theta)$ ne soit pas associé à une probabilité, la loi conditionnelle sachant les covariables et les composantes de patrimoine est bien une probabilité. Cette seconde loi apparait dans la construction de la chaîne de Markov présentée dans l'Annexe 2. Cette loi *a priori* est dite non-informative ou objective. Elle implique des intervalles de confiance plus larges qu'une loi a priori qui serait une vraie loi de probabilité. Nous préférons arbitrer en faveur de l'objectivité, éventuellement au détriment de la précision. C'est pour la même raison qu'il nous semble gênant d'imposer trop de structure dans la spécification du PGD, par exemple en posant des modèles conditionnellement indépendants (matrice de variance-covariance des résidus diagonale).

ANNEXE 2

MÉTHODE DE MONTE-CARLO DANS LE CAS DU MODÈLE HIÉRARCHIQUE AVEC PGD 2

Dans le cas du modèle PGD 2, nous produisons les valeurs G^i de l'encadré 2 à partir de valeurs d'un vecteur $v_i = (\theta_i, p_{1,i}, \ldots p_{r,i}, \varepsilon_i)$ où $p_{k,i}$ sont des vecteurs de valeurs pour les composantes de patrimoine détenues par le ménage d'indice k dont la somme vaut $pt_{k,i}$. $(v_1,\ldots,v_T)$ correspond à la réalisation d'une trajectoire d'une chaîne de Markov de probabilité invariante : la loi du vecteur $V = (\theta, p_1, \ldots, p_r, \varepsilon)$ conditionnelle aux observables (variables explicatives, différents intervalles). La trajectoire est simulée par échantillonnage de Gibbs. Cet algorithme est particulièrement efficace lorsque la moyenne théorique repose sur la loi tronquée de vecteurs de dimension élevée (cf., par exemple, Robert (1995)) et son extension dans le cadre d'une estimation bayésienne (avec le modèle (5)) est naturelle. D'un point de vue pratique, cela permet de ne reposer que sur des étapes de simulation élémentaires. Nous rappelons brièvement le principe de la méthode sur un exemple où le vecteur est découpé en deux sous-vecteurs, le cas d'un découpage en un nombre plus élevé de composantes est analogue. Supposons que nous souhaitons produire un tirage dans la loi d'un vecteur aléatoire $X = (X^{1\prime}, X^{2\prime})'$ et qu'il soit aisé de simuler dans les lois conditionnelles notées $\mathcal{L}(X^1|X^2 = x^2)$ et $\mathcal{L}(X^2|X^1 = x^1)$. Dans ce cas, partant d'une condition initiale $x_0 = (x_0^{1\prime}, x_0^{2\prime})'$, on construit une suite de vecteurs $x_n = (x_n^{1\prime}, x_n^{2\prime})'$ telle que, connaissant x_{n-1}, on tire successivement :

1. x_n^1 dans $\mathcal{L}(X^1 \mid X^2 = x_{n-1}^2)$,

2. x_n^2 dans $\mathcal{L}(X^2 \mid X^1 = x_n^1)$.

L'algorithme permet d'effectuer des simulations dans des lois de vecteurs de dimension inférieure et la probabilité invariante de la chaîne de Markov est bien la loi jointe : la loi de $X = (X^{1\prime}, X^{2\prime})'$.

Du point de vue de la mise en œuvre pratique nous procédons comme suit. Nous donnons une valeur initiale pour le sous-vecteur de l'ensemble des composantes de patrimoine des r ménages satisfaisant l'ensemble des contraintes. En ce qui concerne une condition initiale θ_0 pour θ, il n'est pas nécessaire d'initier les paramètres apparaissant dans la moyenne car leur loi conditionnelle aux données et covariances ne dépend des données initiales que *via* les composantes de patrimoine des r ménages et les matrices de variance-covariance. Les matrices $\sum_{1,0},\ldots, \sum_{8,0}$ que nous initions comme des matrices diagonales où les termes diagonaux sont les variances des résidus dans des modèles pour les marginales. Décrivons la mise à jour du vecteur V à l'étape 1. Étant données les matrices $\sum_{1,0},\ldots, \sum_{8,0}$, nous procédons successivement aux étapes suivantes :

Mise à jour du vecteur β

Nous calculons

$$\Sigma_\beta = \left(\sum_{k=1}^{r} X_k' \left(\Sigma_{P(k),0} \right)^{-1} X_k \right)^{-1}$$

où X_k est la matrice diagonale par blocs où figurent sur la diagonale les lignes de variables explicatives pour chaque composante de patrimoine détenue pour le ménage k et $P(k)$ le type de portefeuille appartenant à {1,…8}, et

$$\hat{\beta} = \Sigma_\beta \left(\sum_{k=1}^{r} X_k' \left(\Sigma_{P(k),0} \right)^{-1} p_{k,0} \right)$$

où $p_{k,0}$ correspond aux données initiales des composantes de patrimoine détenues pour le ménage k. Enfin nous simulons β_1 dans la loi $N(\hat{\beta}, \Sigma_\beta)$.

Mise à jour des matrices de variance-covariance

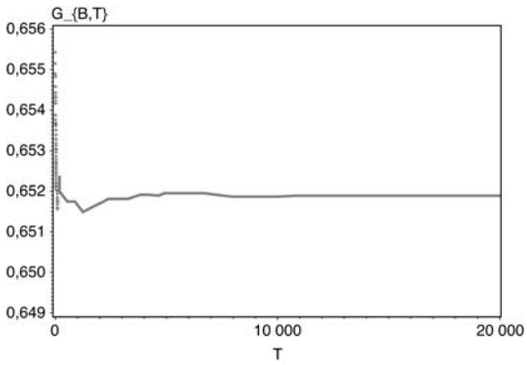
Il convient de construire les matrices $\Lambda_{1,1},\ldots, \Lambda_{8,1}$ définissant les matrices $\sum_{1,1},\ldots, \sum_{8,1}$. Si on note π_i la projection qui permet de passer de β_1 au sous-vecteur pertinent dans le cas d'un portefeuille de type i, $\Lambda_{i,1}$ est donné par

$$\Lambda_{i,1} = \sum_{k:P(k)=i} \left(p_{k,0} - X_k \pi_i \beta_1 \right) \left(p_{k,0} - X_k \pi_i \beta_1 \right)'.$$

Simulation des composantes de patrimoine possédées pour l'ensemble des *r* ménages

Les simulations sont indépendantes pour des ménages différents. Au sein d'un ménage, les composantes de patrimoine sont simulées l'une après l'autre suivant un ordre prédéfini, conditionnellement aux valeurs des paramètres qui ont été actualisés comme expliqué ci-dessus, aux observations des covariables et aux valeurs des autres composantes pour le même ménage, et à l'intervalle sensé la contenir. Comme l'actualisation du vecteur se fait de manière itérative, certaines des autres composantes ont déjà été simulées à nouveau alors que d'autres devant être simulées plus tard ont pour valeur leur valeur à l'étape 0. Nous fabriquons, composante après composante la contrainte de troncature à satisfaire compte tenu des valeurs simulées des autres composantes de patrimoine. Il s'agit donc à chaque

Graphique
Convergence de l'estimateur de l'indice de Gini



Lecture : le graphique présente, pour différentes valeurs de T, la valeur de la moyenne empirique $\hat{G}_{B,T}$ présentée dans l'encadré 2 dans le cadre du modèle hiérarchique avec le modèle PGD 2 et un ensemble d'information comprenant l'imposition à l'ISF. Le nombre d'itérations T vaut 10 000, ici B = 0 .
Source : calculs des auteurs.

fois d'une simulation dans une loi normale tronquée de dimension 1, ce qui est obtenu très simplement, par exemple par inversion de la fonction de répartition (voir aussi Robert, 1995).

Longueur de la trajectoire simulée

Nous avons choisi comme valeur T = 10 000 pour la longueur de la trajectoire $(\upsilon_1,\ldots,\upsilon_T)$ pour obtenir les résultats présentés dans le Tableau 8 et avons procédé comme indiqué dans l'encadré 2. Les T indices de Gini simulés G_t correspondants ont été obtenus à partir des valeurs des patrimoines et de ε_t dans chaque vecteur υ_t. Nous avons pour cela considéré plusieurs critères graphiques (cf. Robert et Casella, 2004).

Le graphique relatif à l'indice de Gini est un exemple des nombreuses représentations graphiques utilisées. Toutes donnent une convergence très rapide.