Les méthodes de pseudo-panel et un exemple d'application aux données de patrimoine

Marine Guillerm*

Les méthodes de pseudo-panel sont une alternative à l'utilisation de données de panel pour l'estimation de modèles à effets fixes, lorsque seules des données en coupes répétées indépendantes sont disponibles. Leur usage est courant pour estimer des élasticités-prix ou revenu et mener des analyses en cycle de vie qui demandent des données sur longue période, alors que les données de panel présentent des limites de disponibilité dans le temps et rencontrent des problèmes d'attrition.

Les pseudo-panels consistent à suivre au cours du temps, plutôt que des individus, des cohortes – c'est-à-dire des groupes stables d'individus. Les variables individuelles sont remplacées par leurs moyennes intra-cohortes. Du fait de la linéarité de cette transformation, au modèle linéaire avec effet fixe individuel correspond son homologue sur les données du pseudo-panel. À l'effet fixe individuel se substitue un effet cohorte et l'estimation du modèle est particulièrement aisée si cet effet cohorte peut être, lui aussi, considéré comme fixe. Le critère de constitution des cohortes doit ainsi prendre en compte un certain nombre de contraintes. Il doit évidemment être observable pour l'ensemble des individus et former une partition de la population (chaque individu est classé dans exactement une cohorte) ; au-delà, il doit correspondre à une caractéristique des individus fixe dans le temps, par exemple l'année de naissance. Enfin, la taille des cohortes répond à un arbitrage biais-variance. Elle doit être suffisante pour limiter l'ampleur des erreurs de mesure des moyennes intra-cohortes des différentes variables qui génèrent biais et imprécision des estimateurs des paramètres du modèle. Cependant, l'augmentation de la taille des cohortes fait diminuer le nombre de cohortes observées, ce qui détériore la précision des estimateurs.

L'extension aux modèles non linéaires n'est pas directe et seulement introduite ici. Enfin une application sur les données des enquêtes *Patrimoine* est donnée.

Codes JEL : C21, C23, C25, D91.

Mots clés : pseudo-panel, données groupées, modèles à effets fixes, données en coupes répétées.

** Au moment de la rédaction de cet article, Insee-DMCSI-DMS (Département des méthodes statistiques) – Division des méthodes appliquées de l'économétrie et de l'évaluation (marine.guillerm@travail.gouv.fr).*

Cet article a bénéficié des commentaires, corrections et remarques de nombreuses personnes. L'auteur les remercie et tout particulièrement Pauline Givord pour son aide et son soutien tout au long de ce travail, ainsi que Simon Beck, Didier Blanchet, Richard Duhautois, Bertrand Garbinti, Stéphane Gregoir, Ronan Le Saout, Simon Quantin et Olivier Sautory. Elle reste seule responsable des erreurs qui pourraient y demeurer. Elle remercie également Pierre Lamarche pour son aide sur l'utilisation des enquêtes Patrimoine*.*

L'analyse économique des comportements se heurte généralement au fait que de nombreuses dimensions importantes pour l'analyse ne sont pas observables dans les données disponibles. Par exemple, les comportements de consommation dépendent de préférences individuelles qui ne sont qu'imparfaitement captées dans les données statistiques. Les estimations d'élasticités-revenu s'en trouvent alors biaisées. Parfois, il est difficile de dissocier les effets de plusieurs variables alors même qu'elles sont observées simultanément. Ainsi, bien que l'âge et la génération soient généralement disponibles, il sera impossible à partir d'une source de données « en coupe » (à une date donnée) de distinguer ce qui relève de l'un ou de l'autre. C'est particulièrement dommageable pour des analyses en cycle de vie. Supposons que l'on s'intéresse aux carrières salariales au cours de la vie, que l'on tenterait de décrire à partir d'une seule enquête. Celle-ci permet bien d'observer des personnes à des âges différents, et donc à des moments successifs de leur vie professionnelle. Mais il ne sera pas possible de distinguer ce qui, dans l'évolution observée du salaire, s'explique effectivement par un effet d'âge (ou d'expérience professionnelle acquise), plutôt que par un effet génération. Ce dernier conditionne en partie le fait d'avoir fait des études plus ou moins longues, d'être entré sur le marché du travail à un moment plus ou moins propice... autant de facteurs qui influent aussi sur le salaire.

Il est classique d'utiliser des données de panel pour répondre à ces questions. À partir des observations répétées dans le temps d'unités identiques, on tente de neutraliser d'éventuelles spécificités individuelles. Cela se fait en général par l'introduction d'un « effet fixe » individuel censé capter ces spécificités. Le fait d'observer les mêmes variables à plusieurs dates est aussi un moyen de traiter en partie les problèmes d'identification décrits ci-dessus. L'âge varie avec le temps, contrairement à la génération, ce qui permet de suivre une même génération à différents âges. Ces données sont cependant rares, souvent limitées à des échantillons de petite taille et couvrent des périodes de temps réduites (ce qui diminue leur intérêt pour une analyse en cycle de vie par exemple). Elles sont en outre sujettes à des problèmes d'attrition ou de non-réponses : il est difficile de suivre les mêmes individus sur une longue période. Au fil du temps, la représentativité des données de panel peut devenir problématique.

Les méthodes de pseudo-panel constituent une manière de pallier l'absence de données de panel. Leur usage remonte à Deaton (1985) qui, le premier, a suggéré d'utiliser des méthodes de panel à partir de données en coupes répétées. L'avantage de ces données est qu'elles sont très souvent disponibles et permettent de couvrir de longues périodes. En effet, de nombreuses enquêtes sont menées à des intervalles réguliers dans le temps. Elles constituent en général des données en coupes répétées indépendantes, au sens où elles portent sur des échantillons différents. Les méthodes de panel ne peuvent pas être directement appliquées, les individus observés changeant à chaque date. Même lorsque l'on dispose de sources exhaustives comme le recensement ou certaines données administratives, il n'est pas possible de suivre des personnes dans le temps par exemple pour des raisons de confidentialité. Cependant, à défaut de suivre des mêmes individus, on peut suivre des types d'individus, qu'on désigne généralement sous le terme de « cohortes » ou encore « cellules ». Ces cohortes sont identifiées par un ensemble de caractéristiques observées dans les données et stables dans le temps (comme la génération ou le sexe). Dans les estimations, on captera les spécificités inobservées qui pourraient biaiser les estimations par un effet fixe « cohorte ». Les pseudo-panels ont été utilisés pour modéliser des sujets aussi différents que l'investissement (Duhautois, 2001), la consommation (Gardes, 1999 ; Gardes et al., 2005 ; Marical & Calvet, 2011), ou encore l'évolution des comportements sur longue période, comme la carrière salariale (Koubi, 2003), l'activité féminine (Afsa & Buffeteau, 2005), le bien-être subjectif (Afsa & Marcus, 2008) ou le niveau de vie (Lelièvre et al., 2010), pour ne citer que les travaux les plus récents. En pratique, la mise en œuvre de ces méthodes repose sur la manière de définir les cohortes. Dans le cas de modèles linéaires, les méthodes d'estimation classiques sur données de panel peuvent alors être adaptées assez simplement.

Cet article propose une introduction à ces techniques, en insistant sur les aspects pratiques. Après un bref rappel des estimations des modèles à effets fixes sur données de panel, il insiste sur les principes qui doivent guider le choix des critères définissant les cohortes. La seconde partie présente les techniques d'estimation. Ces deux premières parties ne traitent que le cas des modèles linéaires. La troisième partie apporte des compléments techniques et évoque notamment l'extension aux modèles dichotomiques. Enfin, la dernière partie propose un exemple pratique tiré de l'exploitation des enquêtes *Patrimoine*.

Il n'aborde pas les questions de mise en œuvre informatique dans les logiciels statistiques. Des exemples de programmes sur les logiciels SAS, R et Stata sont fournis dans Guillerm (2015), dont cet article est issu.

Principe général : de l'effet fixe individuel à l'effet cohorte

Pourquoi utiliser des données de panel, que faire en leur absence ?

Le point de départ des modèles de pseudo-panel sont les modèles linéaires à effets fixes, dont l'usage est classique lorsque l'on dispose de données de panel. Il est donc utile de les présenter (pour une présentation plus détaillée, voir, par exemple, Magnac, 2005). Typiquement, on souhaite modéliser l'influence d'une ou de plusieurs variables explicatives sur une variable d'intérêt. On s'intéresse ici au cas où la variable d'intérêt est continue. Lorsqu'elle est discrète, il faut mobiliser des méthodes spécifiques (voir la partie « Estimation de modèles dichotomiques »). La difficulté de l'estimation de tels modèles vient en général du fait que tous les déterminants de cette variable d'intérêt ne sont pas observés. Si ces déterminants inobservés sont en partie corrélés aux variables explicatives du modèle, le risque existe d'attribuer à tort une partie de leur effet à ces variables explicatives.

L'estimation de l'élasticité-revenu d'un bien de consommation offre une illustration classique d'une telle difficulté. Par exemple, le prix supporté par les ménages lorsqu'ils consomment de la nourriture n'est qu'imparfaitement observé. Au prix des biens de consommation (les denrées) s'ajoute du temps (celui nécessaire à la préparation du repas et à sa consommation) qui n'est pas valorisé de la même manière par tous les ménages. Sa valeur croît avec le revenu (Gardes et al., 2005). Ne pas en tenir compte conduit à sous-estimer l'élasticité-revenu.

Une solution classique est alors d'utiliser des données de panel (c'est-à-dire des observations pour le même individu répétées dans le temps), qui permettent de contrôler certains facteurs dont l'effet est supposé constant dans le temps. On ajoute alors un effet fixe individuel au modèle linéaire classique, censé capter l'effet des caractéristiques individuelles constantes dans le temps sur la variable d'intérêt[1] :

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \quad (1)$$
$$i = 1,\ldots,N \quad t = 1,\ldots,T$$

où y_{it} est la variable d'intérêt (dans notre exemple, le niveau de consommation du bien), x_{it} est un vecteur (ligne) de K variables explicatives observées pour l'individu i à la date t (dans notre exemple, le revenu de l'individu ou du ménage, l'âge...), β est l'effet de ces variables (soit un vecteur de paramètres de dimension K). α_i est l'effet fixe individuel. Il capte tous les déterminants de la variable d'intérêt fixes au cours du temps. En effet, seuls les paramètres associés à des variables non constantes dans le temps sont identifiables dès lors qu'on introduit un effet fixe dans le modèle. Par exemple, on n'obtiendra pas une estimation de l'effet intrinsèque du sexe si le modèle intègre un effet fixe. Enfin, ε_{it} est un terme résiduel, c'est-à-dire tout ce qui n'est pas pris en compte par le modèle. Ignorer l'effet fixe dans l'estimation conduit à des estimateurs biaisés de l'effet des variables explicatives considérées dès lors que ces variables sont corrélées à cet effet fixe.

Quand on dispose d'observations répétées, il est possible dans le cadre de ce modèle linéaire d'estimer l'impact de variables explicatives en neutralisant l'impact des effets fixes individuels. En pratique, cela se fait en utilisant non plus les variables en niveau mais des variables transformées de manière à faire disparaître l'effet fixe individuel. L'estimateur le plus couramment utilisé (car le plus efficace sous certaines hypothèses) est obtenu en procédant à une transformation « *within* » : on utilise à chaque date les observations centrées par rapport à la moyenne individuelle sur la période, c'est-à-dire les variables transformées $z_{it} - \overline{z}_i$, avec $\overline{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}$ la moyenne des valeurs individuelles pour une

1. Les modèles à effets aléatoires sont un autre type de modélisation classiquement mis en œuvre sur des données de panel. Ces modèles incluent aussi un effet individuel et sont une autre façon de tenir compte dans la modélisation du fait que des caractéristiques inobservées de l'individu et fixes dans le temps ont un effet sur la variable d'intérêt. Mais contrairement aux modèles à effets fixes, ils reposent sur l'hypothèse que l'effet individuel n'est pas corrélé aux variables explicatives (l'effet individuel permet de tenir compte de la corrélation des différentes observations associées à un même individu et de ne pas surestimer la précision des estimateurs). Si on est prêt à faire une telle hypothèse, alors les pseudo-panels ne présentent aucun intérêt. En effet, avec des coupes transversales indépendantes, il n'y a pas de corrélation entre les observations, chaque individu n'étant observé qu'une fois. Les modèles peuvent donc être estimés directement sur les données individuelles empilées.

variable z sur l'ensemble de la période d'observation. Une autre solution serait d'estimer directement les effets fixes comme des paramètres du modèle, mais cela suppose l'estimation d'un très grand nombre de paramètres (un effet fixe pour chacun des individus observés en plus des paramètres des variables explicatives), sans grand intérêt en termes d'interprétation[2].

Cet estimateur « *within* » converge vers les vraies valeurs des paramètres d'intérêt dès lors que les variables explicatives ne sont pas corrélées aux termes résiduels restants. Dit autrement, il ne faut pas que les chocs individuels à chaque date, pour un individu donné, soient liés à la réalisation d'une des variables explicatives incluses dans le modèle[3].

Les méthodes de panel reposent cependant sur le fait d'observer les mêmes individus à des dates différentes, ce qui est rare. Dans de nombreux cas, on dispose de données en coupes transversales indépendantes répétées. Le principe des pseudo-panels est alors de suivre dans le temps non plus des individus, mais des cohortes, c'est-à-dire des groupes d'individus partageant un ensemble de caractéristiques fixes dans le temps. Le modèle sera considéré au niveau de ces cohortes d'individus et non plus au niveau des individus qui les composent. En pratique, cela signifie qu'on remplace les variables observées par les moyennes de ces variables au sein de chaque cohorte. Ces données sont assimilées à des données de panel et, quand les conditions le permettent, les techniques d'estimation sur données de panel leur sont appliquées.

Les analyses en cycle de vie sont un autre exemple, avec celui déjà évoqué des estimations d'élasticité-revenu et d'élasticité-prix, où l'usage des méthodes de pseudo-panel est fréquent. Lorsque l'on souhaite étudier l'accumulation du patrimoine au cours du cycle de vie, une analyse naïve consiste, à partir d'observations à une date donnée, à étudier les différences de patrimoine selon l'âge. Cependant, de nombreuses autres caractéristiques individuelles expliquent les différences de patrimoine entre individus : la carrière salariale, le niveau d'étude, les ressources familiales, la plus ou moins grande propension à épargner... Certaines caractéristiques sont corrélées à l'âge. Ce serait le cas, par exemple, si des générations ont connu des conditions d'entrée dans la vie active plus favorables. Ne pas tenir compte de ces déterminants risque de conduire à des estimations biaisées de l'effet de l'âge sur le niveau de patrimoine. Une solution classique est d'introduire ces dimensions supplémentaires (on « contrôle » de l'effet de ces variables) dans un modèle linéaire. Cependant, si certains de ces déterminants sont couramment disponibles dans la plupart des enquêtes, tous ne le sont pas. On aura ainsi facilement une mesure de l'âge, du niveau d'étude ou du salaire actuel, mais il est moins fréquent de disposer d'indications précises sur l'ensemble de la carrière salariale, l'héritage dont les personnes interrogées ont pu bénéficier et encore moins s'ils sont plutôt « fourmi » ou « cigale » au sens d'une plus ou moins grande propension à épargner. Une solution est alors – comme décrit précédemment – d'estimer un modèle à effets fixes similaire à (1).

Analyses en cycle de vie et estimations d'élasticités-revenus ou d'élasticités-prix sont deux exemples de problématiques où l'usage des pseudo-panels est courant. Ce sont spécifiquement des analyses où les données de panel sont rares. Les analyses en cycle de vie demandent de disposer de données sur longues périodes. Des séries de coupes transversales offrent plus souvent cet horizon temporel que des panels. Ceci justifie que, même en présence de données de panel, on ait parfois recours à des estimations par pseudo-panel. Par exemple, Antman et McKenzie (2005) disposent d'un panel rotatif pour analyser la mobilité salariale. Ne retenir à chaque trimestre que le cinquième des nouveaux entrants dans le panel leur permet de disposer de données sur longue période, alors qu'ils seraient limités à cinq trimestres avec le panel. De plus, contrairement aux panels, les pseudo-panels ne sont pas confrontés à de l'attrition liée notamment à la difficulté de suivre des ménages. Dans l'exemple de l'étude de la mobilité salariale, cette attrition pose problème car elle peut être liée à un déménagement qui peut en outre être consécutif à une évolution salariale. Disposant des données de panel, Gardes et al. (2005) mènent des estimations d'élasticités-revenus sur données de panel et sur données de pseudo-panels. Ils montrent que, dans leur exemple, les deux estimations sont proches.

Formellement, on s'intéresse à $y^*_{ct} = E\left(y_{it} | i \in c, t\right)$, espérance de la variable d'intérêt sur la cohorte c à la date t. On obtient à partir du modèle

2. D'autant moins que si on dispose de peu d'observations temporelles par individu, l'estimation des effets fixes est peu précise.
3. Rappelons que ce terme résiduel représente dans le modèle à effets fixes tous les facteurs individuels variables dans le temps que l'on n'observe pas.

précédent (en l'intégrant conditionnellement à la date et à la cohorte) :

$$y^*_{ct} = x^*_{ct}\beta + \alpha^*_{ct} + \varepsilon^*_{ct} \quad (2)$$
$$c = 1,\ldots,C \quad t = 1,\ldots,T$$

où pour chaque variable z, $z^*_{ct} = E(z_{it}|i \in c,t)$.

Comme le modèle initial au niveau individuel, le modèle du pseudo-panel (2) est linéaire en ses paramètres, ce qui permet en principe d'appliquer les techniques d'estimation classiques des modèles de panel. Cependant, en pratique les choses sont un peu plus complexes.

Tout d'abord, les « vraies » valeurs y^*_{ct} et x^*_{ct} ne sont pas connues. On ne dispose que d'une estimation, leur contrepartie empirique au sein de la cohorte observée : $\bar{y}_{ct} = \frac{1}{n_{ct}} \sum_{i \in c,t} y_{it}$ et $\bar{x}_{ct} = \frac{1}{n_{ct}} \sum_{i \in c,t} x_{it}$ (c'est-à-dire, à chaque date, les moyennes des valeurs observées pour les individus de l'échantillon appartenant à cette cohorte). L'estimation à partir de ce sous-échantillon d'individus risque de ne pas correspondre exactement aux « vraies » valeurs. Les fluctuations d'échantillonnage des individus d'une même cohorte d'une date à une autre constituent une deuxième difficulté. À chaque date, les individus observés n'étant pas les mêmes, la moyenne des effets fixes $\bar{\alpha}_{ct}$ est susceptible de varier au cours du temps, alors qu'elle est en théorie constante.

Les erreurs de mesure posent des difficultés différentes pour l'estimation du modèle (2) selon qu'elles portent sur les covariables ou sur la variable d'intérêt. Celles des covariables sont source de biais dans les estimateurs (pour plus de détails, voir « Modèle à erreurs de mesure » et l'annexe B). Le point positif est que plus le nombre d'individus de la cohorte est grand dans l'échantillon d'observations, plus cette estimation sera proche de la vraie valeur et les estimateurs des valeurs moyennes suffisamment précis pour pouvoir négliger les erreurs de mesure dans le modèle économétrique. De leurs côtés, les erreurs de mesure sur la variable d'intérêt et la variabilité temporelle de l'effet cohorte réduisent la précision des estimateurs et posent un problème d'efficacité si l'erreur de mesure est hétéroscédastique. Enfin, le problème de la variabilité dans le temps des effets cohortes peut aussi venir de la définition des cohortes : il faut qu'en amont les effets α^*_{ct} puissent bien être considérés comme constants, au risque sinon de produire des estimateurs biaisés. Ces remarques orientent les critères qu'on retiendra pour définir les cohortes d'individus.

Comment constituer les cohortes ?

En premier lieu, le critère de regroupement doit être observable pour l'ensemble des individus et former une partition de la population (chaque individu est classé dans exactement une cohorte). Au-delà de ces évidences, le critère de constitution des cohortes ne peut pas être choisi au hasard. Il doit viser à rendre plausible l'hypothèse que les termes de cohorte $\bar{\alpha}_{ct}$ sont effectivement fixes au cours du temps. Deux facteurs distincts peuvent remettre en cause cette hypothèse. Avec des données d'enquête, seul un échantillon des vraies cohortes est observé. La première source de variation de $\bar{\alpha}_{ct}$ est liée aux fluctuations d'échantillonnage : $\bar{\alpha}_{ct}$ correspond à la moyenne des effets fixes sur les observations de la cohorte *c* de l'échantillon disponible à la date *t*. Il s'agit d'un estimateur de la vraie valeur α^*_{ct}, inobservée. Même si la vraie cohorte est stable, les individus la représentant changent d'une date à une autre. α^*_{ct} peut aussi varier si la vraie cohorte regroupe une population mouvante au cours du temps, notamment si le critère retenu ne correspond pas à une caractéristique stable au cours du temps des individus. Il s'agit de la deuxième source de variation possible de $\bar{\alpha}_{ct}$.

Un critère stable sur une population stable

Choisir un critère de constitution des cohortes de sorte à rendre α^*_{ct} constant dans le temps permet d'éliminer dans une certaine mesure une des sources de variation de $\bar{\alpha}_{ct}$. α^*_{ct} est fixe lorsque les vraies cohortes regroupent à chaque date les mêmes individus. Deux conditions sont requises : définir les cohortes sur une population stable et sur la base d'un critère stable (cela signifierait sinon que les personnes pourraient changer de profil au cours du temps).

L'année de naissance est évidemment un exemple de critère de regroupement qui correspond à une caractéristique stable des individus. Dans ce cas, on suit des générations d'individus. Ce critère est très fréquemment retenu dans les estimations par pseudo-panel. Le terme cohorte ne doit pas laisser penser que seul ce critère est valide (certains auteurs utilisent le terme « cellule »). D'autres regroupements sont possibles et plusieurs critères peuvent aussi être combinés. Par exemple, Bodier (1999) forme des cohortes par génération et diplôme

de fin d'étude pour étudier les effets d'âge sur le niveau et la structure de consommation des ménages. À l'inverse, un critère de regroupement fondé sur le salaire ou la situation sur le marché du travail ne serait *a priori* pas pertinent : il est susceptible de changer pour une même personne au cours du temps[4].

Mais cette condition de stabilité du critère au niveau individuel n'est pas suffisante. Il faut aussi que la cohorte n'évolue pas elle-même dans le temps. Cette question est particulièrement cruciale lorsque l'on s'intéresse à des données d'enquêtes répétées, sur des échantillons différents. Dans une enquête, les individus d'un certain profil constituent un échantillon de l'ensemble de la cohorte d'intérêt. Mais dans certains cas, leur représentation dans le champ de l'enquête peut varier en fonction des critères retenus pour définir la cohorte. Supposons, par exemple, que l'on constitue des cohortes à partir de l'année de naissance. Selon la date de l'enquête, les différentes générations seront plus ou moins bien représentées : elles rentreront progressivement en fonction de l'âge minimum requis pour être enquêté (ou de la prise d'indépendance des jeunes pour des enquêtes auprès des ménages), tandis qu'à l'inverse les plus âgées en sortiront progressivement (décès, départs en institutions spécialisées si celles-ci sont hors du champ de l'enquête). Il faut faire attention à ces effets de composition pour l'analyse, s'ils sont liés à la variable d'intérêt. Supposons, par exemple, que l'on s'intéresse au profil de revenu de générations successives. L'espérance de vie et le revenu sont en partie corrélés (voir par exemple Blanpain, 2011). Aux âges avancés, les personnes aux revenus les plus élevés sont donc surreprésentées parmi les « survivants » d'une même génération. Une analyse par cohorte qui suivrait une génération laisserait penser que le revenu des individus de cette génération augmente avec l'âge, ce qui n'est probablement pas le cas. En pratique, une analyse au cas par cas est nécessaire pour évaluer si les cohortes représentent au cours du temps une population stable, quitte à restreindre le champ de l'analyse. Par exemple, pour une analyse sur les effets d'âge et de génération sur le niveau et la structure de la consommation, Bodier (1999) restreint la population d'étude aux individus de 25 à 84 ans, considérant que les ménages constitués de personnes au-delà de ces limites risquent de ne plus être représentatifs de l'ensemble des personnes de leur génération.

Il faut souligner que ce problème n'est pas spécifique aux pseudo-panels, mais il est particulièrement apparent lors du suivi sur de longues périodes pour lesquelles ces phénomènes d'entrée-sortie (entrées sur le marché du travail, constitution d'un ménage autonome, créations d'entreprises, décès, migrations, etc.) sont susceptibles d'apparaître. En revanche, contrairement aux données de panels classiques, on n'est pas confronté à des problèmes d'attrition liés à la difficulté de suivre des individus identiques au cours du temps (déménagement, refus de répondre à nouveau).

Former des cohortes de taille suffisante...

Le principe des pseudo-panels est de constituer des cohortes, autrement dit des profils, regroupant des individus dont les comportements sont considérés comme proches. Cette hypothèse sera d'autant plus plausible qu'on définira des profils précis. Néanmoins, en particulier avec des données d'enquête, ceci peut avoir un coût. En effet, plus les cohortes sont petites, plus l'ampleur des erreurs de mesure des moyennes empiriques $\overline{y}_{ct}$ et $\overline{x}_{ct}$ et la variabilité temporelle des moyennes des effets individuels $\overline{\alpha}_{ct}$ sont grandes. Les problèmes de biais et d'imprécision de l'estimateur classique (l'estimateur « *within* ») déjà évoqués *supra* seront d'autant plus importants (pour plus de détails, voir « Modèle à erreurs de mesure » et l'annexe B).

Il est possible de limiter biais et imprécision des estimateurs en formant des cohortes plus grandes. En pratique, dans les études empiriques, il est généralement considéré que le seuil de 100 individus par cohorte est suffisant pour négliger les erreurs d'échantillonnage (et donc simplifier l'estimation). Ce choix s'appuie en particulier sur les études de Verbeek et Nijman (1992, 1993). À partir de données simulées, ces derniers concluent que l'hypothèse est raisonnable (au sens où le biais qui en résulte n'est pas trop élevé) pour des catégories regroupant 100 individus minimum. Cependant, ils préconisent des tailles deux fois supérieures pour réduire significativement les risques de biais.

4. On trouve en pratique des cas de construction de pseudo-panels sur la base de critères non stables dans le temps. L'opportunité de construire de tels pseudo-panels doit être discutée au cas par cas. Marical et Calvet (2011) construisent un pseudo-panel par âge du ménage pour estimer des élasticités-prix de carburant. L'âge n'étant pas une caractéristique stable des individus, même en présence de données de panels les cohortes constituées ne regrouperaient pas les mêmes individus. Mais un pseudo-panel par âge leur permet de suivre des ménages qui ne vieillissent pas et dont la composition familiale qui est liée à la consommation de carburant, évolue peu d'une date à une autre.

... tout en conservant de la variabilité

L'amplitude des erreurs de mesure et le biais et l'imprécision des estimateurs qu'elles génèrent se réduisent à mesure que la taille des cohortes augmente. Cependant, la taille des cohortes n'est pas le seul paramètre à prendre en compte. De fait, il est assez simple de se rendre compte qu'à taille d'échantillon total fixée, constituer de larges cohortes signifie qu'on réduit le nombre d'observations utilisées pour le modèle de pseudo-panel. Supposons par exemple que le critère de constitution des cohortes soit l'année de naissance, mais que les données en coupes répétées contiennent à chaque date peu de personnes d'une génération. Pour réduire les fluctuations d'échantillonnage qui risquent d'en découler, une solution classique est d'augmenter la taille des cohortes en formant des générations définies plus largement (par exemple, par tranche de cinq ans). Mais dans ce cas, la variabilité des observations à une date donnée se réduit, le nombre final d'observations utiles diminuant. En outre, regrouper des générations proches mais différentes signifie aussi qu'on réduit la variabilité au cours du temps de ces moyennes. Ces deux éléments (nombre d'observations utilisées pour l'estimation, faible variabilité) sont deux facteurs qui classiquement réduisent la précision de l'estimateur final. Intuitivement, moins on a d'observations et moins l'estimation est précise. Mais il est aussi nécessaire d'observer des valeurs différentes des grandeurs d'intérêt, autrement dit que ces valeurs varient dans le temps, pour mesurer la force de leur corrélation. On est ainsi confronté à un classique arbitrage biais-variance : former des cohortes de grande taille permet de limiter le biais de l'estimateur, mais fait perdre de la variabilité, de nature à réduire la précision des estimateurs. Verbeek et Nijman (1992) montrent que le biais de l'estimateur *within* classiquement utilisé (cf. *infra*) peut être élevé même si les cohortes sont de grande taille si la variabilité inter-temporelle est faible par rapport aux erreurs de mesure.

En résumé, un bon critère de regroupement doit : (1) être une caractéristique qui ne change pas au cours du temps au niveau individuel, définir une (sous-)population stable, et résulter d'un arbitrage permettant de (2) former des cohortes suffisamment grandes tout en (3) ne faisant pas perdre trop de variabilité. Ces différentes contraintes limitent fortement le choix des critères de constitution des cohortes. En pratique, de nombreuses études utilisent l'année de naissance car ce critère répond à beaucoup de ces contraintes. Il est très souvent disponible dans les données, et il est stable. En outre, selon la taille des échantillons des enquêtes en coupe, il est possible de jouer sur le regroupement de générations proches pour constituer des cohortes plus ou moins grandes. Enfin, il ne faut pas négliger que cette dimension a un intérêt en tant que tel dans de nombreuses études. L'effet cohorte s'interprète ainsi directement comme un effet génération, qu'il peut être intéressant d'étudier. Dans les analyses en cycle de vie en particulier, regrouper les individus par génération permet de garder de la variabilité sur la variable « âge ».

L'estimation des modèles en pseudo-panels

Lorsque le critère de constitution des cohortes a les qualités requises pour considérer le modèle (2) comme un modèle de panel à effets fixes, l'estimation des paramètres repose généralement sur les techniques classiques d'estimation sur données de panel. En pratique, le modèle estimé est donc :

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\alpha}_c + \bar{\varepsilon}_{ct} \qquad (3)$$
$$c = 1,\ldots,C \quad t = 1,\ldots,T$$

On applique une transformation « *within* » évoquée plus haut, dans laquelle, pour chaque cohorte, on centre les différentes variables sur la moyenne des valeurs observées pour cette cohorte sur l'ensemble des dates d'observation. On régresse donc $\bar{y}_{ct} - \bar{y}_c$ sur $\bar{x}_{ct} - \bar{x}_c$, où pour chaque variable z, $\bar{z}_c = \frac{1}{T}\sum_{t=1}^{T}\bar{z}_{ct}$. On obtient l'estimateur *within* :

$$\hat{\beta}_W = \left[\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c)\right]^{-1}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c) \qquad (4)$$

On en déduit un estimateur de l'effet cohorte :

$$\hat{\alpha}_c = \bar{y}_c - \bar{x}_c\hat{\beta}_W \qquad (5)$$

En pratique, on obtient l'estimateur *within* en procédant d'abord à une transformation *within* et en calculant l'estimateur des moindres carrés sur ces variables centrées. Mais il faut faire attention car, du fait que l'on travaille sur des variables transformées, l'estimateur

standard de la variance fourni par la procédure des moindres carrés ordinaires ne correspond pas directement à l'estimateur sans biais de la variance du modèle *within*. Il la sous-estime. Il faut tenir compte d'un facteur multiplicatif $(CT - K) / (CT - C - K)$ où C est le nombre de cohortes, T le nombre de dates d'observation et K le nombres de variables explicatives. Sous SAS, la macro Bwithin de Duguet (1999) tient compte de cette difficulté (voir Guillerm (2015) pour d'autres procédures sous Stata et R).

L'estimateur *within* est obtenu de manière équivalente, soit en incluant des indicatrices de cohortes, soit par instrumentation. Inclure les indicatrices de cohortes dans le modèle (3) permet d'obtenir directement des estimateurs des effets fixes[5] qui ont parfois un intérêt en tant que tel. Dans une analyse en cycle de vie dans laquelle les cohortes seraient constituées des générations, on estime ainsi directement l'effet génération. Attention cependant, l'estimation de ces effets fixes ne sera précise que si le nombre de périodes d'observation est suffisant.

Une méthode d'estimation alternative par instrumentation est proposée par Moffitt (1993). Il montre que l'estimateur *within* (4) du modèle en pseudo-panel correspond techniquement à l'estimateur des doubles moindres carrés sur les données individuelles (variables explicatives ainsi que des indicatrices de cohortes), dans lequel on utiliserait comme instrument l'ensemble des indicatrices de cohortes croisées avec les indicatrices de temps. La preuve formelle est fournie en annexe A. Pour en saisir l'intuition, rappelons que dans la première étape des doubles moindres carrés, on projette les variables explicatives sur les instruments. La projection de x_{it} sur les indicatrices cohorte x date d'observation correspond exactement à la moyenne empirique $\overline{x}_{ct}$, c étant la cohorte à laquelle appartient l'individu i. La deuxième étape consiste à remplacer dans le modèle initial les variables instrumentées par leur projection, soit ici à régresser y_{it} sur $\overline{x}_{ct}$ et les indicatrices de cohortes. On obtient le même estimateur que l'estimateur *within* (4).

Ceci peut simplifier l'estimation : on travaille directement sur les données individuelles. Cette analogie sert également de base à l'extension des pseudo-panels aux modèles dichotomiques (voir « Estimation de modèles dichotomiques »). Un autre intérêt de cette approche est que d'autres types d'instruments plus parcimonieux peuvent être utilisés. Par exemple, si l'année de naissance est retenue, on peut utiliser une fonction de l'année de naissance (un polynôme par exemple) pour construire l'instrument plutôt que des indicatrices associées à une partition des années de naissance.

On remarque d'ailleurs que cette approche permet de retrouver là encore les critères de regroupement des individus en cohortes[6]. Rappelons que deux conditions sont requises pour définir un bon instrument. Il doit d'abord être corrélé aux variables explicatives. Ici, cela renvoie au fait que la constitution des cohortes doit conserver suffisamment de variabilité pour permettre l'estimation du modèle agrégé au niveau des cohortes. Pour en comprendre l'intuition, on peut penser au cas extrême où ces indicatrices croisées « cohorte x date » seraient totalement indépendantes des variables explicatives du modèle : dit autrement, que la distribution de ces variables explicatives est identique à chaque date et d'une cohorte à l'autre. Dans ce cas, les moyennes empiriques de ces variables au niveau d'une date et d'une cohorte sont très proches, ce qui signifie qu'on ne pourra estimer le modèle. L'autre propriété d'un instrument valide est qu'il ne doit pas être corrélé avec les déterminants inobservés de la variable d'intérêt. Moffitt montre que cette propriété est vérifiée si les cohortes sont définies sur un critère stable et quand la taille des cohortes tend vers l'infini.

Au-delà de l'estimation proprement dite, plusieurs remarques s'imposent. Tout d'abord, sur le choix des variables explicatives. Rappelons que dans le modèle linéaire à effets fixes classique, seuls les paramètres associés à des variables non constantes dans le temps sont identifiables : l'effet fixe « absorbe » l'effet des variables constantes. Dans un modèle en pseudo-panel, l'agrégation en cohortes crée artificiellement de la variabilité et donne l'impression que les paramètres associés aux caractéristiques fixes sont identifiables. Par exemple, une variable constante au niveau individuel telle que l'indicatrice « être une femme » devient dans les données du pseudo-panel « la proportion de femmes dans la cohorte c à la date t ». Les variations temporelles observées (normalement faibles) ne sont dues qu'à l'erreur d'échantillonnage. L'introduction de ce type de variables dans l'analyse n'est donc pas recommandée.

---

5. L'estimation directe des effets fixes est déconseillée avec des données individuelles, car elle demande d'estimer un très grand nombre de paramètres. Dans le cadre des pseudo-panels, le nombre de cohortes est en général limité. Si chaque cohorte regroupe environ 100 individus, le nombre d'effets fixes à estimer dans le modèle de pseudo-panel est divisé d'autant par rapport au modèle de panel.

6. Pour plus de détails, voir Moffitt (1993) et Verbeek (2008).

Compléments techniques

Cette partie propose deux extensions au cadre classique pour traiter des difficultés techniques que peuvent soulever les estimations en pseudo-panel : la prise en compte 1) de l'hétéroscédasticité des termes résiduels et 2) des erreurs de mesure dans l'estimation. Enfin, les modèles présentés jusqu'à présent ne sont adaptés qu'au cas où la variable d'intérêt est continue. Lorsqu'elle est discrète, il faut mettre en œuvre des techniques d'estimation spécifiques. Une introduction est proposée dans un troisième temps.

Hétéroscédasticité dans les pseudo-panels

En pratique, la taille des cohortes varie d'une cohorte à une autre et pour une même cohorte, d'une date à une autre. Ces variations de taille sont susceptibles de créer de l'hétéroscédasticité dans le modèle (2). En effet, la précision de l'estimateur dépendant directement de ce nombre, on introduit des termes d'erreur plus ou moins importants selon les cohortes. En présence d'hétéroscédasticité, l'estimateur *within* (4) est sans biais mais l'estimateur de sa précision est biaisé et par conséquent les statistiques de test sont invalides.

L'estimateur intra-efficace est obtenu en pondérant les observations par la taille de la cohorte, ce qui revient à estimer par les moindres carrés le modèle suivant :

$$\sqrt{n_{ct}}\,\bar{y}_{ct} = \sqrt{n_{ct}}\,\bar{x}_{ct}\beta + \sqrt{n_{ct}}\,\alpha_c + \sqrt{n_{ct}}\,\bar{\varepsilon}_{ct} \qquad (6)$$

De même que pour le modèle homoscédastique, $K + C$ paramètres sont à estimer. Cette estimation est facile à mettre en œuvre sauf si le nombre de cohortes est trop important, auquel cas, on souhaite en général procéder à une transformation *within* pour éliminer les effets fixes avant estimation. Mais dans ce modèle, une transformation *within* classique n'élimine pas les indicatrices de cohorte car le poids qui est affecté à chaque cohorte (n_{ct}) varie dans le temps. Gurgand et al. (1997) montrent que dans ce cas l'estimateur intra efficace est :

$$\hat{\beta}_{WP} = \left(X'(WDW)^{-}X\right)^{-1}\left(X'(WDW)^{-}y\right) \qquad (7)$$

où X matrice de dimension CT × K empile les vecteurs lignes $\bar{x}_{ct}$, y vecteur de dimension CT empile les valeurs $\bar{y}_{ct}$, $(WDW)^{-}$ est l'inverse généralisée de la matrice WDW, où W est la matrice *within* classique de dimension CT et D est la matrice diagonale dont les éléments diagonaux sont $1/n_{ct}$.

Modèle à erreurs de mesure

Les méthodes d'estimations présentées dans la section précédente ne tiennent pas compte du fait que les vraies moyennes intra-cohortes notées y^{*}_{ct} et x^{*}_{ct} sont mesurées avec erreurs par les moyennes calculées sur l'échantillon (notées $\bar{y}_{ct}$ et $\bar{x}_{ct}$). Comme présenté *supra*, ces erreurs de mesure sont à l'origine de deux problèmes : celles sur les variables explicatives créent un biais ; celles sur la variable d'intérêt de même que la variabilité temporelle de l'effet cohorte réduisent la précision des estimateurs. Les techniques d'estimation présentées précédemment reposent implicitement sur l'hypothèse que l'on puisse négliger les erreurs de mesure. Quand ce n'est pas le cas, on sera amené à utiliser des techniques appropriées. Les estimateurs du modèle (2) proposés par Deaton (1985) reposent ainsi sur des modèles à erreurs de mesure qui prennent en compte ce problème. Il reprend la théorie développée par Fuller (1986) et l'adapte à l'estimation par pseudo-panel.

On note u_{ct} et v_{ct} les erreurs de mesure :

$$\bar{y}_{ct} = y^{*}_{ct} + u_{ct}$$

$$\bar{x}_{ct} = x^{*}_{ct} + v_{ct}$$

En les intégrant au modèle (2), on obtient :

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \alpha_c + \tilde{\varepsilon}_{ct} \qquad (8)$$

$$c = 1,\ldots,C \quad t = 1,\ldots,T$$

avec $\tilde{\varepsilon}_{ct} = \varepsilon^{*}_{ct} + u_{ct} - v_{ct}\beta$. On montre que ce résidu est corrélé à $\bar{x}_{ct}$.

L'estimateur du paramètre β proposé par Verbeek et Nijman (1993) repose sur une spécification paramétrique de l'erreur de mesure et de sa corrélation avec la variable d'intérêt (pour plus de détails, voir annexe B). Il vaut :

$$\tilde{\beta} = \left(\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}\left(\bar{x}_{ct}-\bar{x}_{c}\right)'\left(\bar{x}_{ct}-\bar{x}_{c}\right) - \frac{T-1}{T}\times\frac{1}{n}\hat{\Sigma}\right)^{-1} \left(\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}\left(\bar{x}_{ct}-\bar{x}_{c}\right)'\left(\bar{y}_{ct}-\bar{y}_{c}\right) - \frac{T-1}{T}\times\frac{1}{n}\hat{\sigma}\right) \qquad (9)$$

$\sum$ et σ correspondent respectivement à la matrice de variance-covariance des erreurs de mesure de x^*_{ct} et à la covariance entre les erreurs de mesure de x^*_{ct} et y^*_{ct}. Elles ne sont en général pas connues. Deaton propose de les estimer à partir des données individuelles :

$$\widehat{\Sigma} = \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}\widehat{\Sigma}_{ct} \tag{10}$$

avec $\widehat{\Sigma}_{ct} = \frac{1}{n-1}\sum_{i\in c,t}\left(x_{it}-\bar{x}_{ct}\right)'\left(x_{it}-\bar{x}_{ct}\right)$

$$\hat{\sigma} = \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}\hat{\sigma}_{ct} \tag{11}$$

avec $\hat{\sigma}_{ct} = \frac{1}{n-1}\sum_{i\in c,t}\left(x_{it}-\bar{x}_{ct}\right)'\left(y_{it}-\bar{y}_{ct}\right)$

Plusieurs types de convergence peuvent être envisagées dans le cas des estimations en pseudo-panels, car plusieurs paramètres entrent en jeu : N le nombre d'individus observés à chaque date, C le nombre de cohortes constituées, n_{ct} la taille des cohortes formées et T le nombre de dates d'observation.

Intuitivement, quand la taille des cohortes augmente, les moyennes intra-cohortes sont des estimateurs des vraies moyennes intra-cohorte d'autant plus précis que la taille des cohortes est grande. Les erreurs de mesure deviennent négligeables et on retrouve l'estimateur *within* classique.

L'estimateur *within* a un biais asymptotique quand la taille des cohortes est fixée mais une variance plus faible que l'estimateur de Verbeek et Nijman (pour plus de détails, voir Verbeek & Nijman, 1993). On est donc confronté à un classique arbitrage biais-variance.

Estimation de modèles dichotomiques

Les estimateurs précédents ne sont adaptés qu'aux modèles linéaires et pas au cas où la variable d'intérêt est binaire. Pour cela, il faut faire appel à des techniques d'estimation spécifiques. En présence de données de panel, le passage du linéaire au non linéaire pour estimer un modèle à effets fixes n'est déjà pas aisé. L'usage de pseudo-panels complexifie encore l'estimation. Jusqu'à présent, peu de travaux ont mis en œuvre les méthodes d'estimation développées pour de tels modèles. Nous n'en donnons ici que les grands principes.

Le modèle à estimer se présente sous la forme :

$$\tilde{y}_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \tag{12}$$
$$i = 1,\ldots,N \quad t = 1,\ldots,T$$

où $\tilde{y}_{it}$ est une variable latente (non observée). La variable binaire observée y_{it} vaut 1 si $\tilde{y}_{it}$ est positive et 0 sinon. x_{it} est un vecteur de variables explicatives, α_i est un effet fixe individuel et ε_{it} un terme d'erreur dont on suppose en général qu'il suit une loi *logit* ou une loi normale.

De même que dans le cas linéaire, on souhaite estimer un modèle à effets fixes. Avec des données de panel, deux techniques d'estimation classiques existent : le *logit* conditionnel qui consiste à transformer les données de sorte à éliminer l'effet fixe (voir, par exemple, Davezies, 2011), ou l'approche de Chamberlain (1984).

L'approche de Chamberlain est le point de départ de la méthode d'estimation sur données de pseudo-panel proposée par Collado (1998). Elle consiste à expliciter la relation entre l'effet fixe individuel et les covariables :

$$\alpha_i = x_{i1}\lambda_1 + \ldots + x_{iT}\lambda_T + \theta_i \tag{13}$$
avec $E\left(\theta_i \mid x_{i1},\ldots,x_{iT}\right) = 0.$

En substituant (13) dans (12), on obtient la forme réduite :

$$\tilde{y}_{it} = x_{i1}\pi_{t1} + \ldots + x_{iT}\pi_{tT} + \theta_i + \varepsilon_{it} \tag{14}$$
$$i = 1,\ldots,N \qquad t = 1,\ldots,T$$

avec $\pi_{ts} = \beta + \lambda_s$ si $s = t$ et $\pi_{ts} = \lambda_s$ sinon. Le terme d'erreur $\theta_i + \varepsilon_{it}$ n'est pas corrélé aux covariables.

En l'absence de données de panel, on ne dispose pas de la série complète des covariables pour un même individu. On ne peut donc pas estimer directement le modèle (14). Collado (1998) propose d'estimer ce modèle en remplaçant dans (14) chaque valeur individuelle des covariables x_{it} par la moyenne de la cohorte à laquelle appartient l'individu, soit $\bar{x}_{ct}$. La construction des cohortes obéit ici aux mêmes règles que celles présentées dans le cadre linéaire (voir *supra*). On note que la variable d'intérêt y_{it} n'est pas agrégée.

Substituer les moyennes intra-cohortes des variables explicatives aux observations individuelles

introduit des erreurs de mesure dans le modèle (la somme de la déviation individuelle à la moyenne intra-cohorte et de l'erreur liée à l'échantillonnage) et de la corrélation entre le terme d'erreur et les covariables. Collado propose deux estimateurs du paramètre β. Ces estimateurs sont calculés en deux étapes. La première, commune aux deux estimateurs, consiste à estimer par pseudo-maximum de vraisemblance les paramètres π_{ts}. Les deux estimateurs du paramètre β proposés se déduisent ensuite de l'estimateur des paramètres π_{ts}. L'un est calculé par distance minimale, l'autre en procédant à une transformation *within* sur les données. L'estimateur *within* présente l'avantage d'être plus simple à calculer mais n'est pas efficace contrairement à l'estimateur par distance minimale.

Une autre technique d'estimation est proposée par Moffitt (1993). Elle repose sur le parallèle établi entre estimation en pseudo-panel et instrumentation (voir *supra*). Pour rappel, dans le cas linéaire, l'estimation du modèle par pseudo-panel est équivalente à instrumenter par les indicatrices de cohortes croisées avec les indicatrices de date d'observation. Cette même instrumentation est proposée par Moffitt pour estimer le modèle (12).

Un exemple d'application des pseudo-panels : effet d'âge et de génération sur le niveau de patrimoine

On trouve de nombreux exemples d'applications des pseudo-panels dans l'économétrie de la consommation (par exemple, Gardes et al., 2005, ou Marical & Calvet, 2011) et dans les analyses en cycle de vie (voir encadré). Nous proposons ici une application élémentaire des méthodes de pseudo-panel à l'estimation d'un effet âge sur le patrimoine détenu par les ménages. Cette application très simplifiée par rapport à la problématique de l'accumulation patrimoniale ne vise qu'à illustrer la mise en œuvre concrète de ces méthodes. On trouvera dans Lamarche et Salembier (2012) une analyse plus complète de cette question.

On utilise les différentes enquêtes *Patrimoine*. Cette enquête est menée tous les six ans depuis 1986[7]. Nous disposons de cinq dates d'observation (1986, 1992, 1998, 2004 et 2010). Les ménages sont interrogés sur leur détention de biens immobiliers, financiers et professionnels. La somme de ces trois patrimoines constitue le patrimoine brut (calculé en euros constants 2010). En 2010, l'enquête a connu des évolutions importantes visant à mieux évaluer le patrimoine des ménages. En particulier, les catégories de ménages ayant les plus hauts patrimoines ont été surreprésentées dans l'échantillon sélectionné et des éléments du patrimoine comme la voiture, l'équipement de la maison, les bijoux, les œuvres d'art ont été pris en compte. Pour ne pas biaiser les évolutions entre 2004 et 2010, ces évolutions méthodologiques ont été en grande partie neutralisées dans les calculs de patrimoine.

Pour décrire très brièvement la problématique, il s'agit d'étudier les logiques d'épargne aux différents âges. Dans sa version initiale formulée par Modigliani et Brumberg (1954), la théorie du cycle de vie prévoit que les individus procèdent à une affectation intertemporelle de leurs revenus. Au cours de leur vie, ils connaissent trois périodes durant lesquelles leurs revenus, leurs comportements d'épargne et de consommation diffèrent. Le début de leur vie active est marqué par des revenus faibles et une désépargne. Ensuite, au cours de leur vie active, leur revenu augmentant, ils épargnent et se constituent un patrimoine, en prévision d'une baisse de revenu au moment de leur retraite. Le patrimoine suit ainsi une évolution en cloche avec l'âge. Il est difficile de tester l'hypothèse du cycle de vie, en estimant par exemple comment le patrimoine évolue avec l'âge. Une telle estimation requièrerait de disposer du suivi des mêmes personnes sur très longue période, ce qui n'est pas possible. Comme on l'a déjà souligné, une estimation en coupe ne serait pas satisfaisante, car elle ne permet pas de distinguer les effets de l'âge de ceux de la génération. Sur ce cas très simple d'estimation d'un effet âge, il est possible de commencer par une démarche exploratoire très classique avec les deux graphiques suivants. Chaque enquête *Patrimoine* permet de représenter l'évolution des patrimoines bruts moyens en fonction de l'âge (figure I). Les profils obtenus semblent conforter sans restriction la théorie du cycle de vie. On observe bien en effet une courbe en cloche, avec une croissance du patrimoine brut jusqu'à près de 60 ans et une baisse au-delà. Cependant, une partie de ce profil s'explique par le fait que l'on compare à chaque date des générations distinctes. Le contexte économique, l'âge d'entrée dans la vie active, la fiscalité sont autant de caractéristiques

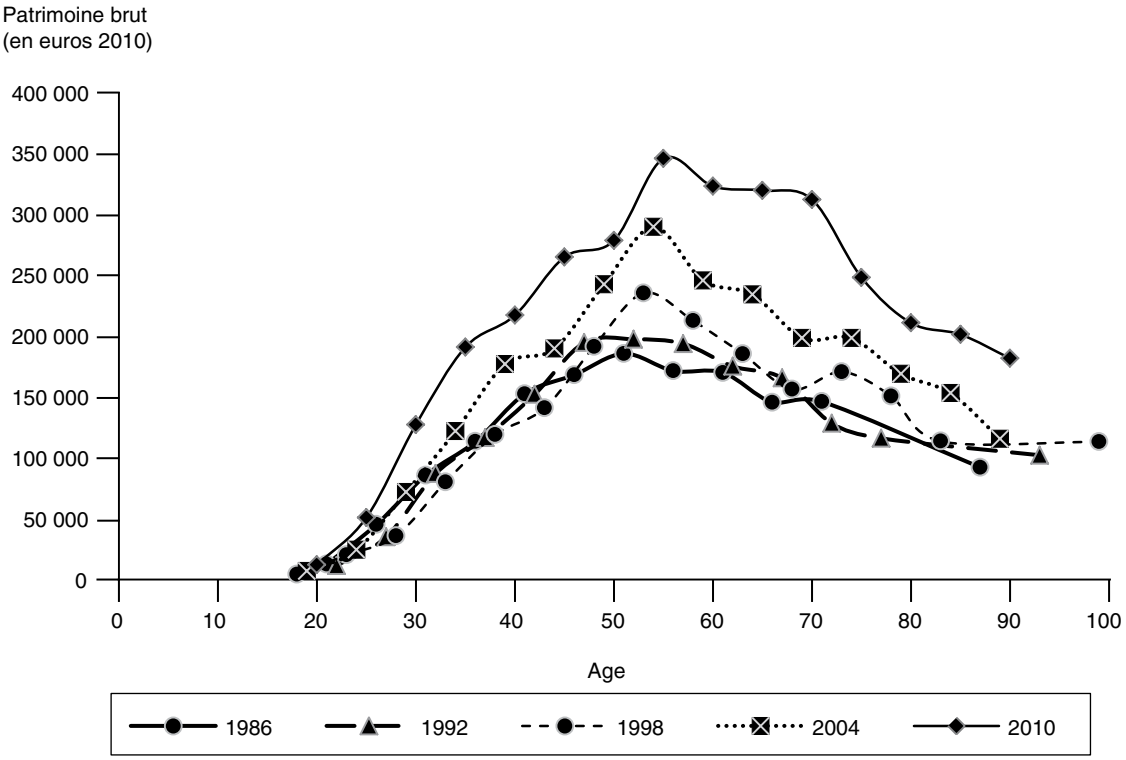
7. *En 1986 et 1992, il s'agissait de l'enquête* Actifs financiers.

partagées par les individus d'une même génération qui ont un effet sur le patrimoine accumulé et expliquent des différences de patrimoine à âge égal entre les différentes générations. Séparer ces deux effets demande un suivi sur longue période de ces générations.

Pour tenter de capter cette dimension « génération », on empile donc toutes les enquêtes pour disposer de l'observation de personnes de générations identiques à des dates différentes (et donc des âges différents). On dispose donc de cinq observations, correspondant au patrimoine moyen à cinq âges différents, pour presque toutes les générations (sauf pour les plus jeunes ou les plus âgées). En principe, on pourrait représenter un profil pour toute génération, définie par l'année de naissance. En pratique cependant, on est confronté au problème que dans l'échantillon d'une enquête, le nombre d'individus d'une génération donnée n'est pas très élevé. Ces estimations sont donc très imprécises. Pour pallier ce problème, on définit des cohortes comme le regroupement de générations adjacentes (cinq sur la figure II).

La figure II représente pour chaque cohorte le profil d'accumulation du patrimoine par âge. Il est très différent de celui présenté en utilisant uniquement la dimension en coupe. Contrairement à ce que la figure I suggère, le patrimoine continue de croître bien après 60 ans. Comme souligné par Lamarche et Salembier (2012), ce fait stylisé s'explique par plusieurs facteurs. Même au-delà de la retraite les ménages peuvent être incités à épargner, dans l'idée de transmettre un patrimoine ou simplement pour constituer une épargne de précaution (liée aux risques de dépendance). Par ailleurs, les plus âgés peuvent renoncer à se séparer de leur patrimoine immobilier, souvent synonyme de déménagement, en raison de son coût particulièrement élevé (voir par exemple Angelini & Laferrère, 2012). Il faut aussi souligner que la croissance du patrimoine avec l'âge traduit en partie des changements de composition des générations observées aux âges extrêmes. Le champ de l'enquête ne porte que sur les ménages ordinaires et exclut donc les personnes âgées en maison de retraite. De plus, les ménages aisés ont une espérance de vie plus

Figure I
Patrimoine en fonction de l'âge pour chacune des cinq enquêtes *Patrimoine*



Lecture : le patrimoine des individus interrogés dans l'enquête Patrimoine *2010, à un âge compris entre 48 et 52 ans est en moyenne de 278 156 euros. Chaque classe d'âge est représentée sur l'axe des abscisses en son centre (par exemple 65 ans pour la classe d'âge 63-67 ans).*
Champ : ménages résidant en France (hors Mayotte).
Source : enquêtes Patrimoine *1986 à 2010, Insee.*

élevée que les autres (et aussi probablement un patrimoine supérieur).

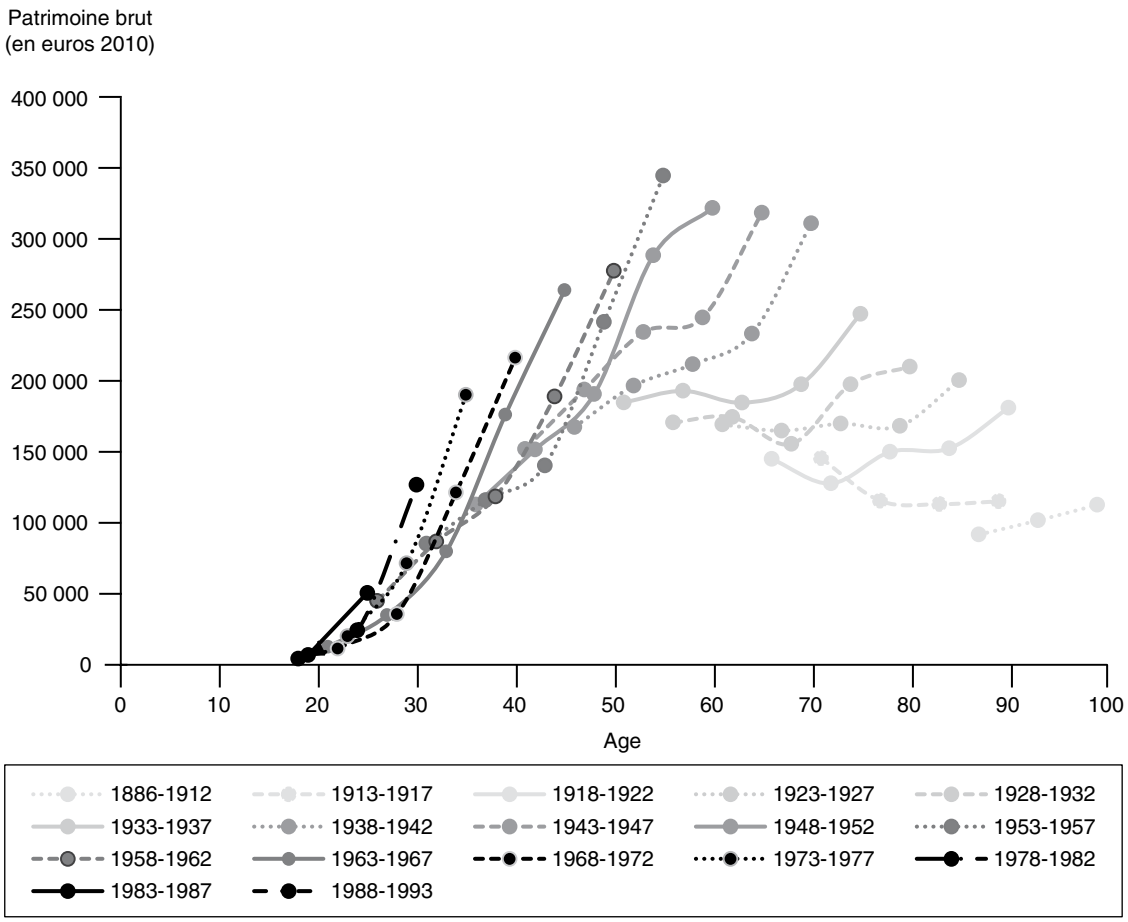
La figure II permet de comparer le patrimoine moyen des différentes cohortes au même âge. On observe des écarts parfois conséquents. L'écart vertical entre les courbes correspond à l'effet de génération, ainsi qu'à un effet période. Pour illustrer, supposons que ces effets périodes, qui correspondent à l'augmentation au cours du temps des patrimoines (rappelons qu'on travaille en euros constants 2010 afin de ne pas intégrer l'inflation), soient négligeables. Cela permet de résoudre le problème d'identification des effets d'âge, de cohorte et de période (cf. encadré). Sous cette hypothèse, le graphique suggère que, à âge égal, chaque génération a accumulé plus de patrimoine que la précédente. L'écart est particulièrement élevé entre les générations nées dans les années 50 qui ont connu les Trente Glorieuses et les précédentes qui ont connu la guerre. La décroissance du patrimoine après 60 ans observée sur la figure I tient ainsi certainement plus à des écarts de richesse importants entre ces deux générations qu'à une désépargne au moment de la retraite.

La modélisation économétrique en pseudo-panel permet de quantifier plus précisément les effets d'âge qui ressortent sur la figure II. Elle s'appuie sur un modèle qui s'écrit au niveau individuel :

$$\log Pat_{it} = \beta_1 age_{it} + \beta_2 age_{it}^2 + \alpha_i + \varepsilon_{it} \qquad (15)$$
$$i = 1,\ldots,N \quad t = 1,\ldots,T$$

$\log Pat_{it}$ est le logarithme du patrimoine de l'individu i à la date t, age_{it} son âge à la date t. On suppose ici que l'effet de l'âge sur le patrimoine

Figure II
Patrimoine en fonction de l'âge selon la génération



Lecture : le patrimoine des individus interrogés dans l'enquête Patrimoine *2010, à un âge compris entre 48 et 52 ans est en moyenne de 278 156 euros. Chaque classe d'âge est représentée sur l'axe des abscisses en son centre (par exemple 65 ans pour la classe d'âge 63-67 ans).*
Champ : ménages résidant en France (hors Mayotte).
Source : enquêtes Patrimoine *1986 à 2010, Insee.*

est identique pour toutes les générations, et qu'il a un profil quadratique[8]. α_i est un effet fixe individuel. Il estime l'impact des caractéristiques fixes inobservées de l'individu *i* sur son patrimoine.

Le modèle en pseudo-panel que l'on estime en pratique s'écrit :

$$\left(\log Pat\right)_{gt} = \beta_1 age_{gt} + \beta_2 age_{gt}^2 + \alpha_g + \varepsilon_{gt} \qquad (16)$$
$$g = 1,\ldots,G \quad t = 1,\ldots,T$$

où pour chaque variable z, $z_{gt} = E(z_{it} \mid i \in g,t)$. Ces valeurs ne sont pas observées. Elles sont estimées par les moyennes intra-cohortes $\overline{z}_{gt} = \frac{1}{n_{gt}} \sum_{i \in g,t} z_{it}$ calculées sur les données disponibles, avec n_{gt} le nombre d'individus de la cohorte *g* observés à la date *t*.

Deux remarques pratiques doivent être faites. La première porte sur la constitution de l'échantillon. L'estimation repose sur le fait que $\overline{\alpha}_{gt}$ est fixe dans le temps. Ceci peut être remis en cause. Comme discuté plus haut, pour les générations les plus âgées deux effets de composition jouent : d'une part, les ménages les plus aisés ont en moyenne une longévité supérieure et, d'autre part, l'enquête *Patrimoine* n'interroge pas les individus en maison de retraite. À l'autre extrême, l'enquête *Patrimoine* ne comprend que quelques ménages très jeunes, qui sont sans doute très spécifiques. Pour travailler sur une population stable, on se restreint aux ménages de plus de 26 ans et de moins de 80 ans[9]. La deuxième remarque porte sur la taille des cohortes. Les cohortes regroupent plusieurs générations successives. Restreindre le nombre de ces générations successives réduit le risque d'agréger des comportements hétérogènes mais au prix d'estimations reposant sur un nombre d'observations par cohorte très faible : elles risquent donc d'être très imprécises. Pour illustrer cette question, on a estimé le modèle en prenant des cohortes plus ou moins larges (trois, cinq et dix années) (tableau C1 en annexe).

On présente dans le tableau ci-après les résultats des estimations en pseudo-panel. À titre de comparaison, on présente également les résultats obtenus par une régression en coupe (on empile les données des cinq enquêtes successives) et les estimations en tenant compte des erreurs de mesure.

La figure III représente l'effet de l'âge sur le patrimoine tel qu'il est estimé en coupe d'une

8. L'accumulation du patrimoine ne diffère avec l'âge entre les différentes générations qu'en niveau. Le modèle pourrait être complexifié en intégrant des termes d'interaction entre l'âge et la génération.

9. Par ailleurs, les moyennes étant sensibles aux valeurs extrêmes, certains ménages aux patrimoines très élevés ont été écartés de l'analyse. De même, on supprime les quelques observations correspondant à un patrimoine nul, car on utilise une modélisation en logarithme.

Encadré

EFFETS D'ÂGE, DE COHORTE ET DE PÉRIODE

L'estimation simultanée d'un effet d'âge, de cohorte et de période est un problème récurrent antérieur aux pseudo-panels mais qui se pose de la même manière sur des données individuelles et sur des données de pseudo-panel. La difficulté vient de la colinéarité entre les trois variables (âge + cohorte = période) ou, dit autrement, du fait qu'il n'est pas possible d'observer des individus de même âge et de même génération à des dates différentes.

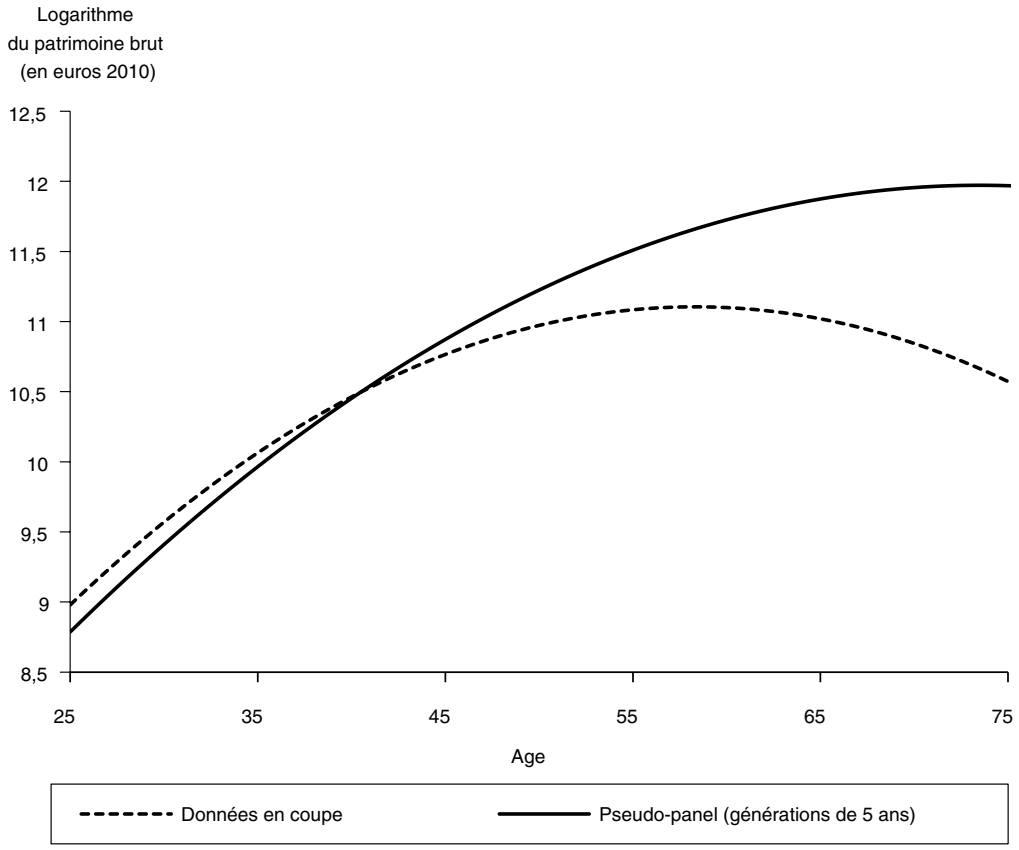
On se place en général dans le cas où les effets d'âge, de cohorte et de période sont additifs. Le modèle inclut alors simplement un ensemble d'indicatrices d'âge, de cohorte et de période, sans termes d'interaction. Cette hypothèse d'additivité n'est pas anodine. Elle conduit à supposer que l'effet âge, par exemple, est commun à toutes les générations. Dans le cadre de ce modèle, deux solutions principales ont été proposées dans la littérature pour résoudre le problème d'identification. La première solution consiste à imposer des contraintes identifiantes au modèle (en plus de la nullité d'un coefficient pour chaque dimension et d'une contrainte identifiante en présence d'une constante dans le modèle). Mason et al. (1973) montrent qu'il suffit de supposer que deux coefficients d'une même dimension (âge, cohorte ou période) sont égaux. Des contraintes identifiantes différentes conduisent à des estimations différentes et doivent être discutées au cas par cas. Rodgers (1982) s'oppose à cette pratique et propose de remplacer l'un des effets par des variables qui lui sont corrélées, par exemple des variables macro-économiques à la place de l'effet période. Le lecteur intéressé par cette problématique peut se référer par exemple à Hall et al. (2007) pour une revue de littérature sur le sujet ou à Yang et Land (2013).

Tableau
Estimation des effets de l'âge

		Estimations par pseudo-panel		
	Données en coupe	Générations de 3 ans	Générations de 5 ans	Générations de 10 ans
		Estimateur *within*		
Constante	4.59*** (0.127)	4.80*** (0.383)	4.65*** (0.437)	4.89*** (0.542)
Age	0.223*** (0.0052)	0.197*** (0.0142	0.199*** (0.016)	0.193*** (0.0212)
Age^2	- 0.0019*** (0.0000493)	- 0.00140*** (0.000135	- 0.00136*** (0.000145)	- 0.00136*** (0.0002)
		Modèle à erreurs de mesure Estimateur de Verbeek et Nijman (9)		
Constante		4.63*** (0.279)	5.05*** (0.307)	5.63*** (0.398)
Age		0.203*** (0.0104)	0.187*** (0.0127)	0.162*** (0.0172)
Age^2		- 0.00143*** (0.000092)	- 0.00128*** (0.00012)	- 0.00102*** (0.00016)
Nombre d'observations	43 117	94	57	31

*Note : la constante est calculée en prenant comme générations de référence les années de naissance 1951-1953 pour les générations de 3 ans. 1953-1957 pour celles de 5 ans et 1953-1962 pour celles de 10 ans. Les écarts-types ont été calculés par bootstrap pour le modèle à erreurs de mesure. ***. **. * indiquent respectivement une significativité des coefficients à 1 %. 5 % et 10 %. Le nombre d'individus observés dans les différentes générations est présenté dans le tableau C2 en annexe.*
Champ : ménages résidant en France (hors Mayotte).
Source : estimation à partir des enquêtes Patrimoine*.*

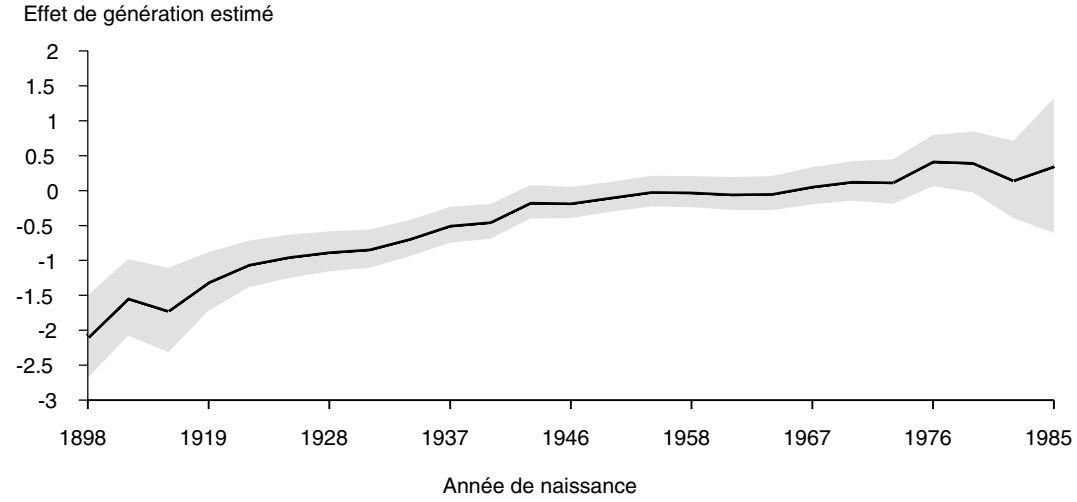
Figure III
Patrimoine en fonction de l'âge tel qu'estimé par les modèles



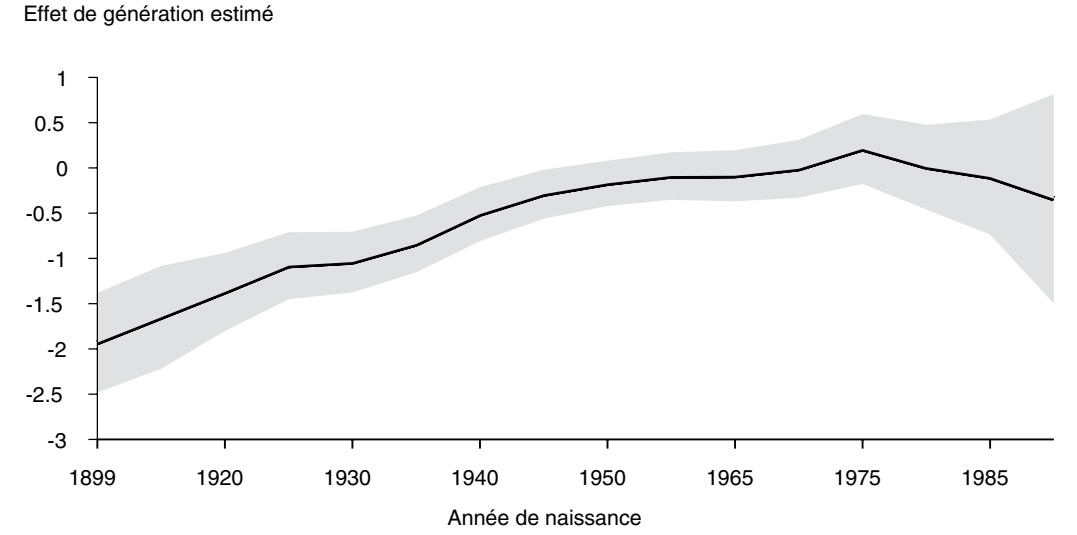
Lecture : à 65 ans, le logarithme du patrimoine brut tel qu'estimé par le modèle en pseudo-panel est de 11.87.
Champ : ménages résidant en France (hors Mayotte).
Source : estimation à partir des enquêtes Patrimoine*.*

Figure IV
Effets générations estimés par pseudo-panel

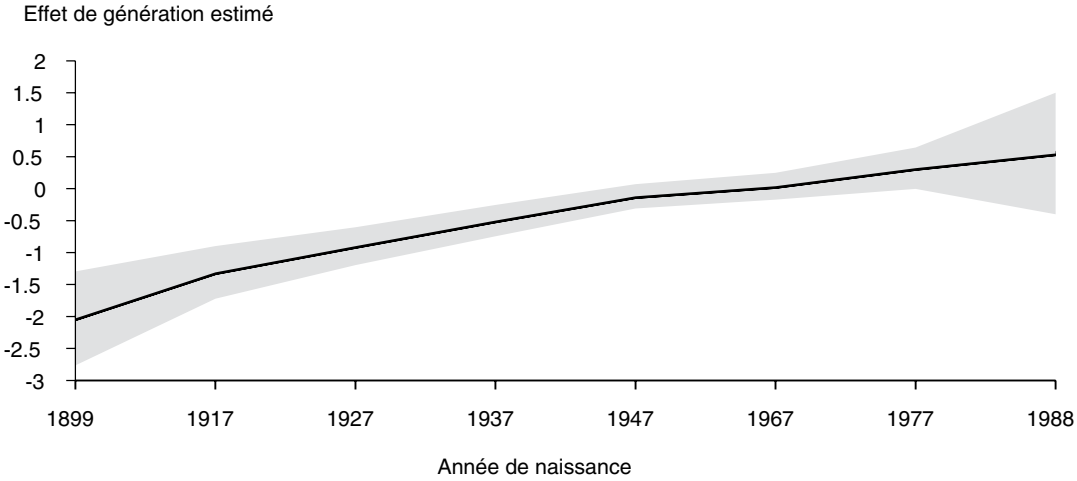
(a) Générations de 3 ans



(b) Générations de 5 ans



(c) Générations de 10 ans



Lecture : l'effet génération estimé par le modèle de pseudo-panel (générations de 3 ans) pour la génération 1939-1941 est de – 0.44, ce qui correspond à un patrimoine brut 35.6 % moins élevé que la génération de référence 1951-1953. La zone grisée correspond à l'intervalle de confiance à 5 %.
Champ : ménages résidant en France (hors Mayotte).
Source : estimation à partir des enquêtes Patrimoine*.*

part et en pseudo-panel d'autre part[10]. Les deux estimations montrent une relation entre patrimoine et âge en forme de cloche. Sur données en coupe, on estime que le patrimoine commence à décroître à partir de l'âge de 58 ans. Dans l'estimation par pseudo-panel, cet âge est beaucoup plus avancé : il se situe à 70 ans. Lorsque l'on tient compte de l'effet génération, la baisse du patrimoine est ainsi beaucoup plus tardive qu'une coupe transversale le suggère.

Le modèle étant log-linéaire, 100 x [exp(α_g) - 1], où α_g est le coefficient associé à la génération *g* dans le modèle (tableau C2 en annexe et figure IV *infra*), correspond à l'effet en pourcentage sur le patrimoine d'appartenir à la génération *g* plutôt qu'à la génération 1951-1953 (génération de référence). Par exemple, être né entre 1939 et 1941 plutôt qu'entre 1951 et 1953 a un effet négatif sur le patrimoine, estimé à 100 x [exp(– 0,44) – 1] = – 35.6 %. On estime ainsi qu'entre les générations 1939-1941 et 1951-1953, le patrimoine a cru de 3.7 % en moyenne annuelle. Ensuite, la croissance a ralenti.

La sensibilité des estimations au critère de regroupement des cohortes n'apparaît pas trop élevée ici. La figure IV représente les effets générations estimés par pseudo-panel selon trois largeurs choisies pour définir les générations. Sans surprise, plus la largeur est élevée et plus le profil est lisse. On observe dans tous les cas, une augmentation importante du patrimoine des générations successives jusqu'à celles du baby-boom, et une stagnation ensuite. Pour les générations les plus jeunes, le diagnostic semble diverger selon le critère de regroupement, mais ces évolutions ne sont jamais significatives (cf. tableau C2 en annexe). Cette incertitude tient au fait que les estimations sont effectuées sur des échantillons plus réduits (ces générations ne sont pas observées dans les enquêtes les plus anciennes), comme illustré dans le tableau C1 (annexe, point C). On remarque également que, comme attendu, la précision des estimateurs des coefficients β_1 et β_2 est plus importante pour des générations de trois ans que pour des générations de cinq ou dix ans.

L'estimateur de Verbeek et Nijman qui tient compte des erreurs de mesure a également été calculé directement avec la formule de l'estimateur. Comme la formule de l'estimateur le suggère, le sens du biais n'est pas connu a priori et change selon la largeur retenue pour la définition des générations. Les estimations diffèrent peu de celles obtenues avec l'estimateur *within*, sauf pour les générations de 10 ans. □

10. On représente donc le polynôme de degré deux : $\beta_0 + \beta_1 age + \beta_2 age^2$ dont les coefficients sont estimés sur données en coupe d'une part et par pseudo-panel d'autre part.

BIBLIOGRAPHIE

Afsa, C. & Buffeteau, S. (2005). L'évolution de l'activité féminine en France : une approche par pseudo-panel. Insee, *Document de travail-DESE* G2005/02.

Afsa, C. & Marcus, V. (2008). Le bonheur attend-il le nombre des années ? Insee, *France, portrait social*, 163–174.

Angelini, V. & Laferrère, A. (2012). Residential mobility of the European elderly. *CESifo Economic Studies*, 58(3), 544–569.

Antman, F. & McKenzie, D. (2005). Earnings mobility and measurement error: a pseudo-panel approach. The World Bank, *Policy Research Working Paper Series* 3745.

Blanpain, N. (2011). L'espérance de vie s'accroît, les inégalités sociales face à la mort demeurent. *Insee Première* N° 1372.

Bodier, M. (1999). Les effets d'âge et de génération sur le niveau et la structure de la consommation. *Economie et Statistique*, 324-325, 163–180.

Chamberlain, G. (1984). Panel data. In: Z. Griliches & M. D. Intriligator (Eds), *Handbook of Econometrics*, vol. 2. Elsevier Science Publishers BV, Ch. 22, pp. 1247–1318.

Collado, M. D. (1998). Estimating binary choice models from cohort data. *Investigaciones Economicas*, 22(2), 259–276.

Davezies, L. (2011). Modèles à effets fixes, à effets aléatoires, modèles mixtes ou multi-niveaux : propriétés et mises en œuvre des modélisations de l'hétérogénéité dans le cas de données groupées. Insee, *Document de travail DESE* G2011/03.

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics,* 30(1-2), 109–126.

Duguet, E. (1999). Macro-commandes SAS pour l'économétrie des panels et des variables qualitatives. Insee, *Document de travail DESE* G 9914.

Duhautois, R. (2001). Le ralentissement de l'investissement est plutôt le fait des petites entreprises tertiaires. *Economie et Statistique*, 341-342, 47–66.

Fuller, W. A. (1986). *Measurement Error Models*. New-York (NY): John Wiley & Sons, Inc.

Gardes, F. (1999). L'apport de l'économétrie des panels et des pseudo-panels à l'analyse de la consommation. *Economie et Statistique*, 324-325, 157–162.

Gardes, F., Duncan, G. J., Gaubert, P., Gurgand, M. & Starzec, C. (2005). Panel and pseudo-panel estimation of cross-sectional and time series elasticities of food consumption: The case of U.S. and Polish data. *Journal of Business and Economic Statistics*, 23, 242–253.

Guillerm, M. (2015). Les méthodes de pseudo-panel. Insee, *Document de travail Méthodologie et Statistique-DMCSI,* M 2015/02.

Gurgand, M., Gardes, F. & Bolduc, D. (1997). Heteroscedasticity in pseudo-panel. Université de Paris I, *Cahier de Recherche Lamia*, unpublished Working Paper.

Hall, B. H., Mairesse, J. & Turner, L. (2007). Identifying age, cohort, and period effects in scientific research productivity: Discussion and illustration using simulated and actual data on French physicists, *Economics of Innovation and New Technology*, 16(2), 159–177.

Koubi, M. (2003). Les carrières salariales par cohorte de 1967 à 2000. *Economie et Statistique*, 369-370, 149–170.

Lamarche, P. & Salembier, L. (2012). Les déterminants du patrimoine : facteurs personnels et conjoncturels. *Insee Références - Les revenus et le patrimoine des ménages*, 23–41.

Lelièvre, M., Sautory, O. & Pujol, J. (2010). Niveau de vie par âge et génération entre 1996 et 2005. *Insee Références - Les revenus et le patrimoine des ménages*, 23–35.

Magnac, T. (2005). Économétrie linéaire des panels : une introduction. Insee, *Neuvièmes Journées de Méthodologie Statistique*.

Marical, F. & Calvet, L. (2011). Consommation de carburant : effets des prix à court et à long terme par type de population. *Economie et Statistique*, 446, 25–44.

Mason, K. O., Mason, W. M., Winsborough, H. H. & Poole W. K. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review,* 38(2), 242–258.

Modigliani, F. & Brumberg, R. (1954). Utility analysis and the consumption function: An interpretation of cross-section data. In: Kurihara K. K. (Ed). *Post-Keynesian Economics*. New Brunswick: NJ. Rutgers University Press, pp. 388–436.

Moffitt, R. (1993). Identification and estimation of dynamic models with a time series of repeated cross-sections. *Journal of Econometrics*, 59(1-2), 99–123.

Rodgers, W. L. (1982). Estimable functions of age, period, and cohort effects. *American Sociological Review*, 47(6), 774–787.

Verbeek, M. (2008). Pseudo-panels and repeated cross-sections. In: Mátyás L. & Sevestre P. (Eds), *The Econometrics of Panel Data, Advanced Studies in Theoretical and Applied Econometrics*, vol. 46, Berlin Heidelberg: Springer, pp. 369–383.

Verbeek, M. & Nijman, T. (1992). Can cohort data be treated as genuine panel data ? *Empirical Economics,* 17(1), 9–23.

Verbeek, M. & Nijman, T. (1993). Minimum MSE estimation of a regression model with fixed effects from a series of cross-sections. *Journal of Econometrics*, 59(1-2), 125–136.

Yang, Y. & Land, K. C. (2013). *Age-period-cohort analysis: New models, methods, and empirical applications.* CRC Press.

ANNEXE

A. PSEUDO-PANEL ET INSTRUMENTATION

Moffitt (1993) montre que l'estimation par pseudo-panel et l'estimation en instrumentant par les indicatrices de cohortes croisées avec celles de la date d'observation fournissent le même estimateur.

Une estimation par les doubles moindres carrés suit les deux étapes suivantes :

Première étape : projection des variables explicatives sur l'instrument.

Si on décompose l'effet fixe individuel α_i en un effet fixe cohorte α_c et une déviation individuelle $v_i = \alpha_i - \alpha_c$, le modèle (1) se réécrit :

$$y_{it} = x_{it}\beta + \alpha_c + v_i + \varepsilon_{it} \qquad (17)$$

x_{it} est potentiellement corrélé à v_i. Aussi x_{it} est instrumenté par les indicatrices de cohortes en interaction avec les indicatrices de temps. La première étape consiste à projeter x_{it} sur l'instrument. La valeur prédite de x_{it} dans cette régression correspond à la moyenne intra-cohorte $\bar{x}_{ct}$.

Deuxième étape :

x_{it} est remplacé par sa valeur prédite dans (17). On régresse ainsi y_{it} sur $\bar{x}_{ct}$ et les indicatrices de cohortes, ce qui fournit le même estimateur que l'estimateur *within* (4).

B. DÉTAILS SUR L'ESTIMATION DES PARAMÈTRES D'UN MODÈLE À ERREURS DE MESURE

$\bar{x}_{ct}$ et $\bar{y}_{ct}$ sont des observations avec erreurs des vraies moyennes intra-cohortes x^*_{ct} et y^*_{ct}. u_{ct} et v_{ct} sont les erreurs de mesure :

$$\bar{y}_{ct} = y^*_{ct} + u_{ct} \qquad (18)$$

$$\bar{x}_{ct} = x^*_{ct} + v_{ct} \qquad (19)$$

Elles sont supposées être normalement distribuées :

$$\begin{pmatrix} u_{ct} \\ v'_{ct} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \frac{1}{n} \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right) \qquad (20)$$

avec n la taille des cohortes.

En intégrant (18) et (19) dans le modèle (2), on obtient :

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \alpha_c + \tilde{\varepsilon}_{ct} \qquad c = 1,\ldots,C \qquad t=1,\ldots,T \qquad (21)$$

avec $\tilde{\varepsilon}_{ct} = \varepsilon^*_{ct} + u_{ct} - v_{ct}\beta$.

La corrélation entre ce résidu et les covariables vaut :

$$E\left(\bar{x}'_{ct}\tilde{\varepsilon}_{ct}\right) = \frac{1}{n}(\sigma - \Sigma\beta)$$

Elle n'est pas nulle en général. L'estimateur des moindres carrés de $\bar{y}_{ct}$ sur $\bar{x}_{ct}$ est donc biaisé.

Le modèle (21) est un modèle à effets fixes. Après transformation *within*, le modèle (21) devient :

$$\bar{y}_{ct} - \bar{y}_c = (\bar{x}_{ct} - \bar{x}_c)\beta + \tilde{\varepsilon}_{ct} - \bar{\tilde{\varepsilon}}_c \text{ où } \bar{\tilde{\varepsilon}}_c = \frac{1}{T}\sum_{t=1}^{T}\tilde{\varepsilon}_{ct} \qquad (22)$$

On montre que :

$$E\left(\bar{x}_{ct} - \bar{x}_c\right)'\left(\bar{y}_{ct} - \bar{y}_c\right) =$$
$$E\left(\bar{x}_{ct} - \bar{x}_c\right)'\left(\bar{x}_{ct} - \bar{x}_c\right)\beta + \frac{T-1}{T} \times \frac{1}{n}(\sigma - \Sigma\beta)$$

De cette équation, on déduit une expression de β :

$$\beta = \left[E\left(\bar{x}_{ct} - \bar{x}_c\right)'\left(\bar{x}_{ct} - \bar{x}_c\right) - \frac{T-1}{T} \times \frac{1}{n}\Sigma \right]^{-1}$$
$$\left[E\left(\bar{x}_{ct} - \bar{x}_c\right)'\left(\bar{y}_{ct} - \bar{y}_c\right) - \frac{T-1}{T} \times \frac{1}{n}\sigma \right]$$

L'estimateur (9) est la contrepartie empirique de cette expression.

On remarque que lorsque seule la variable expliquée est observée avec erreur, l'estimateur *within* est sans biais mais sa précision est détériorée par rapport à un modèle sans erreur de mesure. Lorsque l'erreur de mesure porte uniquement sur les variables explicatives, on a un biais d'atténuation (la valeur absolue de l'estimateur *within* converge vers une valeur plus faible que la valeur absolue du paramètre β).

C. APPLICATION DES PSEUDO-PANELS AUX DONNÉES DES ENQUÊTES *PATRIMOINE*

Tableau C1
Effectifs des cohortes

Générations de 3 ans

Génération (année de naissance)	Année de l'enquête *Patrimoine*				
	1986	1992	1998	2004	2010
1886-1911	267				
1912-1914	191	124			
1915-1917	109	132			
1918-1920	179	268	153		
1921-1923	321	431	375		
1924-1926	278	502	397	228	
1927-1929	305	544	440	421	
1930-1932	301	498	469	444	336
1933-1935	282	522	512	468	555
1936-1938	287	426	456	413	593
1939-1941	284	430	488	445	569
1942-1944	317	481	502	392	704
1945-1947	372	614	654	467	804
1948-1950	408	727	728	562	894
1951-1953	391	683	680	570	838
1954-1956	373	731	626	554	756
1957-1959	292	704	652	560	774
1960-1962	77	569	582	544	723
1963-1965		407	582	552	743
1966-1968		124	465	506	654
1969-1971			463	511	599
1972-1974			132	426	541
1975-1977				367	414
1978-1980				112	396
1981-1983					290
1984-1986					85

Lecture : dans l'enquête Patrimoine *1986 le nombre d'individus interrogés nés entre 1954 et 1956 est de 373.*
Champ : ménages résidant en France (hors Mayotte).
Source : enquêtes Patrimoine*.*

Générations de 5 ans

Génération (année de naissance)	Année de l'enquête *Patrimoine*				
	1986	1992	1998	2004	2010
1886-1912	344				
1913-1917	223	256			
1918-1922	391	551	395		
1923-1927	477	831	672	359	
1928-1932	516	861	767	734	336
1933-1937	476	787	804	744	964
1938-1942	478	742	815	707	954
1943-1947	588	944	993	734	1307
1948-1952	678	1181	1192	938	1457
1953-1957	615	1213	1068	964	1295
1958-1962	248	1020	1008	888	1233
1963-1967		531	915	877	1209
1968-1972			727	842	1000
1973-1977				643	742
1978-1982				112	598
1983-1987					173

Lecture : dans l'enquête Patrimoine *1986 le nombre d'individus interrogés nés entre 1953 et 1957 est de 615.*
Champ : ménages résidant en France (hors Mayotte).
Source : enquêtes Patrimoine*.*

Générations de 10 ans

Génération (année de naissance)	Année de l'enquête *Patrimoine*				
	1986	1992	1998	2004	2010
1886-1912	344				
1913-1922	614	807	395		
1923-1932	993	1692	1439	1093	336
1933-1942	954	1529	1619	1451	1918
1943-1952	1266	2125	2185	1672	2764
1953-1962	863	2233	2076	1852	2528
1963-1972		531	1642	1719	2209
1973-1982				755	1340
1983-1993					173

Lecture : dans l'enquête Patrimoine *1986 le nombre d'individus interrogés nés entre 1953 et 1962 est de 863.*
Champ : ménages résidant en France (hors Mayotte).
Source : enquêtes Patrimoine*.*

Tableau C2
Effets de génération estimés

Générations de 3 ans	
1886-1911	- 2.09***
	(0.302)
1912- 1914	- 1.53***
	(0.279)
1915-1917	- 1.71***
	(0.308)
1918-1920	- 1.30***
	(0.214)
1921-1923	- 1.05***
	(0.170)
1924-1926	- 0.94***
	(0.158)
1927-1929	- 0.87***
	(0.146)
1930-1932	- 0.83***
	(0.140)
1933-1935	- 0.68***
	(0.132)
1936-1938	- 0.49***
	(0.131)
1939-1941	- 0.44***
	(0.127)
1942-1944	- 0.16
	(0.122)
1945-1947	- 0.17
	(0.114)
1948-1950	- 0.088
	(0.109)
1951-1953	*réf.*
1954-1956	- 0.0078
	(0.112)
1957-1959	- 0.014
	(0.114)
1960-1962	- 0.042
	(0.120)
1963-1965	- 0.035
	(0.125)
1966-1968	0.069
	(0.136)
1969-1971	0.14
	(0.144)
1972-1974	0.13
	(0.163)
1975-1977	0.43
	(0.187)
1978-1980	0.41
	(0.223)
1981-1983	0.16
	(0.283)
1984-1986	0.36
	(0.493)

Générations de 5 ans	
1886-1912	- 1.93***
	(0.281)
1913-1917	- 1.65***
	(0.290)
1918-1922	- 1.37***
	(0.220)
1923-1927	- 1.08***
	(0.189)
1928-1932	- 1.04***
	(0.171)
1933-1937	- 0.84***
	(0.160)
1938-1942	- 0.51***
	(0.152)
1943-1947	- 0.29**
	(0.138)
1948-1952	- 0.17
	(0.128)
1953-1957	*réf.*
1958-1962	- 0.089
	(0.134)
1963-1967	- 0.0086
	(0.144)
1968-1972	- 0.0089
	(0.163)
1973-1977	0.21
	(0.197)
1978- 1982	0.0098
	(0.239)
1983-1987	- 0.10
	(0.324)
1988-1993	- 0.34
	(0.590)

Générations de 10 ans	
1886-1912	- 2.03***
	(0.375)
1913-1922	- 1.31***
	(0.210)
1923-1932	- 0.90***
	(0.151)
1933-1942	- 0.50***
	(0.125)
1943-1952	- 0.12
	(0.097)
1953-1962	*réf.*
1963-1972	0.038
	(0.107)
1973-1982	0.32*
	(0.165)
1983-1993	0.55
	(0.485)

Lecture : le coefficient estimé de la génération 1939-1941 dans le modèle est – 0.44, ce qui signifie qu'être né entre 1939 et 1941 plutôt qu'entre 1951 et 1953 (génération de référence) a un effet négatif sur le patrimoine, estimé à 100 x [exp(-0.44) - 1] = - 35.6 %.
****, **, * indiquent respectivement une significativité des coefficients à 1 %, 5 % et 10 %.*
Champ : ménages résidant en France (hors Mayotte).
Source : enquêtes Patrimoine*.*