Pseudo-panel methods and an example of application to Household Wealth data

Marine Guillerm *

Pseudo-panel methods are an alternative to using panel data for estimating fixed effects models when only independent repeated cross-sectional data are available. They are widely used to estimate price or income elasticities and carry out life-cycle analyses, for which long-term data are required, but panel data have limits in terms of availability over time and attrition.

Pseudo-panels observe cohorts, i.e. stable groups of individuals, rather than individuals over time. Individual variables are replaced by their intra-cohort means. Due to the linearity of this transformation, the linear model with individual fixed effect corresponds to its pseudo-panel data counterpart. The individual fixed effect is replaced by a cohort effect and the model is particularly simple to estimate if the cohort effect can be itself considered as a fixed effect. The criteria for forming the cohorts must therefore take into account a number of requirements. It must obviously be observable for all the individuals and form a partition of the population (each individual is classified into exactly one cohort); beyond this, it must correspond to a characteristic of the individuals that will not change over time (e.g. year of birth). Finally, the size of the cohorts results from a trade-off between bias and variance. It must be large enough to limit the extent of measurement error on intra-cohort variable means, that generates bias and imprecise estimators of the model parameters. However, increasing the size of the cohorts decreases the number of cohorts observed, which makes estimators less precise.

The extension to non-linear models is not direct and only introduced here. Finally, the article provides an application to the French Household Wealth Survey (*enquête Patrimoine*).

JEL codes: C21, C23, C25, D91.

Keywords: pseudo-panel, grouped data, fixed effects models, repeated cross-sectional data.

* *Insee-DMCSI-DMS (Department of Statistical Methods – Applied Methods for Econometrics and Evaluation Unit) at the time of writing. (marine.guillerm@travail.gouv.fr)*

This article received comments, corrections and remarks from several people. The author wishes to thank them, and particularly Pauline Givord for her help and support throughout this project, as well as Simon Beck, Didier Blanchet, Richard Duhautois, Bertrand Garbinti, Stéphane Gregoir, Ronan Le Saout, Simon Quantin and Olivier Sautory. Any remaining errors are the author's sole responsibility. She would also like thank Pierre Lamarche for his assistance on the French Household Wealth surveys.

This article is translated from « Les méthodes de pseudo-panel et un exemple d'application aux données de patrimoine ».

Behavioural economics is generally confronted by the fact that many dimensions of the information needed to analyse behaviours cannot be observed in the available data. For example, consumer behaviours depend on individual preferences that are only imperfectly captured in statistical data. Income elasticity estimates are therefore biased. Sometimes it is difficult to dissociate the effects of several variables even though they are observed at the same time. Although age and generation are usually available, it will be impossible to distinguish what derives from one or the other on the basis of cross-sectional data (at a given date). This is particularly detrimental for life-cycle analysis. Take the example of examining variation in wage trajectories over the lifecycle. Cross-sectional data would provide observations on individuals of different ages and for this reason, at various stage of their careers. However, it is not possible, on the basis of this information, to establish that differences observed in wage trajectories result from an effect of age (or professional experience), rather than an effect of generation. The generation effect partially determines the time individuals spend on their education, the job market conditions when they begin their career, which are factors that also influence the wage.

It is standard to use panel data to answer these questions, using observations repeated over time for identical units with the aim of neutralising potentially specific individual characteristics. This usually involves introducing individual "fixed effects" to capture these specific characteristics. Repeated observations of the same variables at different dates helps also to address, at least partly, the aforementioned identification problems. Age varies with time, unlike the generation, which means that the same generation can be observed at different ages. However, this type of data is rare and often limited to small samples and covers short time periods, (this reduces their relevance for life-cycle analysis for example). This type of data is also subject to attrition or non-response problems, making it difficult to follow the same individuals over a long period of time. Over time, the representativeness of panel data can become problematic.

Pseudo-panel methods are one way of making up for the lack of panel data. Their use dates back to Deaton (1985), who was the first to suggest using panel methods on repeated cross-sectional data. The advantage of these data is their availability and the fact that they can cover long periods of time, many surveys being carried out at regular intervals over time. They generally include independent repeated cross-sections, i.e. different samples. Panel methods cannot be directly applied as the observed individuals change at each date. And even with exhaustive sources such as census surveys or certain administrative data, it is not always possible to follow individuals over time for reasons such as confidentiality. However, when the same individuals cannot be followed, types of individuals, generally referred to as "cohorts" or "cells" can be followed. These cohorts are identified by a set of observed characteristics that are stable over time (such as the generation or gender). In the estimations, this make it possible to capture, by a fixed "cohort" effect, some unobserved characteristics that could result in biased estimations. Pseudo-panels have been used to model a wide range of topics, including investment (Duhautois, 2001), consumption (Gardes, 1999; Gardes et al., 2005; Marical & Calvet, 2011), or long-term behavioural changes, such as wage trajectories (Koubi, 2003), women's participation in the labour market (Afsa & Buffeteau, 2005), subjective well-being (Afsa & Marcus, 2008) or living standards (Lelièvre et al., 2010), to mention just the most recent research. In practice, the use of these methods depends on the way in which cohorts are defined. In the case of linear models, standard estimation methods using panel data can be adapted quite easily.

This article provides an introduction to these techniques with an emphasis on practical aspects. After a brief recap of fixed effects models on panel data, it focuses on the principles that should guide the criteria applied for the definition of cohorts. The second part presents estimation methods. These first two sections only cover the case of linear models. The third part provides additional technical information and evokes the extension to dichotomous models. Finally, the last section provides a case study with an application to the French Household Wealth surveys (*enquêtes Patrimoine*).

Issues of implementation of statistical software are not addressed in the articles. Examples of SAS, R and Stata programmes are provided in Guillerm (2015), on which the article is based.

General principle: from individual fixed effects to cohort effects

Why use panel data and what to do when they are not available

The starting point for pseudo-panel models are fixed effects linear models, typically used with panel data. It is therefore useful to present them (for a more detailed presentation, see Magnac, 2005). In general, we want to model the influence of one or more explanatory variables on a variable of interest. We consider here the case of continuous variables of interest. For binary variables, specific methods need to be used (see section on "Estimation of dichotomous models"). The difficulty of estimating these types of models usually stems from the fact that the determinants of the variable of interest are not all observed. If these unobserved determinants are partially correlated with the explanatory variables of the model, there is a risk of incorrectly attributing part of their effect to these explanatory variables.

A classic illustration of this problem is the estimation of the income elasticity of a consumer good. For example, the actual price of food consumption is imperfectly observed: the time spent on the preparation and consumption of meals, which is not valued in the same way by each household, needs to be added to the price of the goods themselves. The value of time increases with income (Gardes et al., 2005). Not taking this value into account results in underestimating the income-elasticity of food consumption.

A typical solution is to use panel data (i.e. repeated observations of the same individuals over time), in order to control factors whose effect is supposed to be constant over time. An individual fixed effect is therefore added to the standard linear model, in order to capture the effect of individual characteristics that are constant over time on the variable of interest[1]:

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

where y_{it} is the variable of interest (in the example, the level of consumption of the good), x_{it} is a vector (line) of K explanatory variables observed for the individual i on the date t (in the example, individual or household income, age, etc.), β is the effect of these variables (i.e. a vector of parameters of dimension K).

α_i is the individual fixed effect. It captures all the determinants of the variable of interest that are fixed over time. Only the parameters associated with variables that are not constant over time can be identified if a fixed effect is introduced into the model. For example, an estimate of the intrinsic effect of gender cannot be obtained if the model includes a fixed effect. Finally, ε_{it} is a residual term, i.e. anything that the model does not take into account. Ignoring the fixed effect in the estimation leads to biased estimators of the effect of the explanatory variables considered when these variables are correlated with the fixed effect.

With repeated observations, the impact of explanatory variables can be estimated using the linear model by neutralising the impact of individual fixed effects. In practice, this can be done by using a transformation of the variables instead of their level, in order to eliminate the individual fixed effect. The most commonly used estimator (as it is the most efficient under certain assumptions) is obtained by carrying out a "within" transformation: at each date we use observations centred on the individual mean over the period, i.e. the transformed variables $z_{it} - \overline{z}_i$, where $\overline{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}$ is the mean of individual values of z over the entire observation period. Another solution would be to directly estimate the fixed effects as model parameters. However, this implies estimating a very large number of parameters (a fixed effect for each of the individuals observed in addition to explanatory variable parameters), which has no real interest for the interpretation[2].

This "within" estimator converges towards the true values of the parameters of interest insofar as the explanatory variables are not correlated

1. Random effects models are another type of modelling traditionally used on panel data. These models also include an individual effect and are another way of taking into account the fact that unobserved characteristics of the individual that are fixed over time have an effect on the variable of interest in modelling. However, unlike fixed effects models, they are based on the assumption that the individual effect is not correlated with the explanatory variables (the individual effect takes into account the correlation of different observations associated with a single individual without overestimating the precision of estimators). If we are able to make such an assumption, there is no point in using pseudo-panels. With independent cross-sections, there is no correlation between the observations, as each individual is only observed once. Models can therefore be estimated directly based on stacked individual data.

2. Especially since if few temporal observations per individual are available, fixed effects estimation lacks precision.

with the remaining residual terms. In other words, the individual impacts at each date, for a given individual, must not be linked to the realisation of any of the explanatory variables included in the model[3].

However, panel methods are based on the observation of the same individuals at different dates, which is rare. In many cases, we have repeated independent cross-sectional data. The principle of pseudo-panels is to follow cohorts (i.e. groups of individuals sharing a set of characteristics that are fixed over time), rather than individuals over time. The model will be considered in terms of these cohorts of individuals rather than the individuals in them. In practice, this means that the observed variables are replaced by the means of these variables within each cohort. These data are treated as panel data and, when possible, panel data estimation techniques are applied.

Life-cycle analysis is another example, as already mentioned with the estimation of income-elasticity and price-elasticity, where pseudo-panel methods are frequently used. If we want to study the accumulation of household wealth over the life cycle, a naïve analysis would study differences in wealth according to age using observations at a given date. However, many other individual characteristics explain the differences in wealth between individuals, such as variations in wage and career, education level, family resources, propensity to save, etc. Some characteristics are correlated with age. For example, this would be the case if some generations had experienced more favourable conditions than others at the beginning of their careers. Failing to take these determinants into account can lead to biased estimations of the effect of age on household wealth. A typical solution is to include these additional aspects (the effect of these variables is "controlled") in a linear model. However, although some of these determinants are usually available in most surveys, this is not always the case. It is therefore easy to obtain measures of age, education level or current salary, but it is more difficult to obtain precise information over the entire career, or on inherited assets, let alone determine if they are "ants" or "grasshoppers" in terms of their propensity to save. As described above, one solution is to estimate a fixed effects model similar to (1).

Life-cycle analysis and the estimation of income or price-elasticity are two examples of issues where pseudo-panels are often used for lack of panel data. Life-cycle analysis requires data over particularly long periods of time, and series of cross-sections provide this time dimension more often than panel data. This justifies the use of pseudo-panel estimations even when panel data are available. For example, Antman and McKenzie (2005) use a rotating panel to assess earnings mobility. Keeping only the new observations entering the panel each quarter (one fifth of the sample) provides them with a long-term data, while they would have been limited to a period of five quarters if they had used the panel. Furthermore, unlike panels, pseudo-panels do not raise issues of sample attrition associated with following households. In the example of earnings mobility, attrition raises problems because it may be related to a move, which itself may result from a change in earnings. Using panel data, Gardes et al. (2005) carry out an estimation of income-elasticity on panel data and pseudo-panels. In the example they use, they show that the estimations are quite close.

Formally, we are interested in $y^*_{ct} = E\left(y_{it} | i \in c, t\right)$, the expectation of the variable of interest in cohort c at date t. The following is obtained from the previous model (by its integration conditional to the date and cohort):

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

where for each variable z, $z^*_{ct} = E(z_{it} | i \in c, t)$.

Like the initial model at the individual level, the pseudo-panel model (2) is linear in its parameters, which means that, in principle, standard estimation techniques can be used for panel data. However, in practice, things are a little more complicated.

First, the "true" values y^*_{ct} and x^*_{ct} are not known. We only have an estimation, their empirical counterpart within the observed cohort: $\bar{y}_{ct} = \frac{1}{n_{ct}} \sum_{i \in c,t} y_{it}$ and $\bar{x}_{ct} = \frac{1}{n_{ct}} \sum_{i \in c,t} x_{it}$ (i.e., at each date, the means of observed values for the individuals of the sample belonging to the cohort). The estimation on this sub-sample of individuals may not correspond exactly with "true" values. Fluctuations in the sampling of individuals from a same cohort from one date to another are another problem. Since the observed individuals are not the same at each date, the

3. In the fixed effects model, this residual term represents all the individual factors that are variable over time and not observed.

mean of fixed effects $\bar{\alpha}_{ct}$ may vary over time, although in theory, it is constant.

Measurement errors raise different difficulties for estimating model (2), depending on whether they affect the covariates or the variable of interest. Measurement errors on the covariates result in biased estimators (for further details, see "Measurement error model" and Appendix B). The good thing is that the higher the number of individuals of the cohort in the sample, the closer the estimation will be to the true value and the higher the precision of the estimators of the mean values will be, making it possible to neglect measurement errors in the model. On the other hand, measurement errors on the variable of interest and the temporal variability of the cohort effect reduce the precision of estimators and lead to a problem of efficiency if the measurement error is heteroscedastic. Finally, the problem of the variability of cohort effects over time can also stem from how cohorts are defined: beforehand, the effects α^{*}_{ct} must be able to be considered as constant, otherwise there is a risk of producing biased estimators. These remarks guide the criteria that will be used when defining the cohorts of individuals.

Constructing cohorts

Firstly, the selection criterion must be observable for all the individuals and form a partition of the population (each individual is classified into exactly one cohort). Beyond this obvious point, the criteria for defining the cohorts must not be chosen at random. It must aim to make plausible the assumption that the cohort terms $\bar{\alpha}_{ct}$ are fixed over time. Two distinct factors can call this assumption into question. With survey data, only one sample of the true cohorts is observed. The first source of variation of $\bar{\alpha}_{ct}$ comes from sampling fluctuations: $\bar{\alpha}_{ct}$ corresponds to the mean of fixed effects on the observations of cohort *c* from the sample available at date *t*. It is an estimator of the true value α^{*}_{ct}, which is not observed. Even if the true cohort is stable, the individuals that represent it change over time. α^{*}_{ct} can also vary if the true cohort is made up of a population itself unstable over time, especially if the criterion adopted does not correspond to a characteristic of the individuals that is stable over time. This is the second potential source of variation of $\bar{\alpha}_{ct}$.

A stable criterion on a stable population

Choosing a selection criterion that makes α^{*}_{ct} constant over time eliminates one of the sources of variation of $\bar{\alpha}_{ct}$, to a certain extent. α^{*}_{ct} is fixed when the true cohorts contain the same individuals at each date. Two conditions are required: that cohorts are constructed on a stable population and on the basis of a stable criterion (otherwise it would mean that the profile of the individuals might change over time).

Year of birth is obviously an example of a selection criterion that corresponds to a stable characteristic of the individuals. In this case, generations of individuals are followed. This criterion is frequently used in pseudo-panel estimations. The term cohort does not imply that only this criterion is valid (some authors use the term "cell"). Other groupings are possible and several criteria can be combined. For example, Bodier (1999) constructs cohorts based on the generation and higher education level to study the effects of age on the level and structure of household consumption. Conversely, a selection criterion based on earnings or the labour market status would not be relevant *a priori* because, for a given individual, it is likely to change over time [4].

However this condition of criterion stability at the individual level is not sufficient. The cohort itself must not change over time either. This issue is particularly crucial for repeated survey data on different samples. In a survey, individuals with a particular profile form a sample of the entire cohort of interest. However in some cases, their representation in the survey may vary depending on the criteria applied to construct the cohort. For example, let us assume that cohorts are defined on the basis of the year of birth. Depending on the date of the survey, the different generations will be represented to varying degrees. They will progressively enter the cohort as they reach the minimum age required to be surveyed (or when young people form new households), whereas the oldest individuals will gradually leave (death, entry into retirement homes or care institutions if out of the scope of the survey). It is important to be aware of these composition effects for analysis if they are linked to the variable of interest. For example, let us assume that we are interested in

4. In practice, there are cases where pseudo-panels have been constructed using criteria that are unstable over time. The relevance of such pseudo-panels must be discussed on a case by case basis. For instance, Marical and Calvet (2011) construct a pseudo-panel based on household age to estimate fuel price elasticities. As age is not a stable characteristic of individuals, even with panel data, the cohorts would not contain the same individuals. However, a pseudo-panel by age can be used to follow households that do not age, and where the family composition (which is linked to fuel consumption) changes little over time.

the profile of the income of successive generations. Life expectancy and income are partially correlated (for example, see Blanpain, 2011). At an advanced age, individuals with the highest income are therefore overrepresented among the "surviving" individuals of a single generation. A cohort analysis that following a generation could suggest that the income of individuals from this generation increases with age, which might not be the case. In practice, a case by case analysis is necessary to assess whether the cohorts represent a stable population over time, even if it means limiting the scope of the analysis. For example, in a study on the effects of age and generation on the level and structure of consumption, Bodier (1999) limited the population to individuals aged 25 to 84, considering that households composed of people beyond these limits may no longer be representative of the population of their generation.

It has to be underlined that this problem is not specific to pseudo-panels, but it is particularly obvious when cohorts are followed over long periods where these entry and exit phenomena (entries onto the labour market, leaving the parents' home, business creation, death, migration, etc.) are likely to occur. However, unlike traditional panel data, attrition problems associated with the difficulty of following identical individuals over time (due e.g. to moving, refusal to answer the next wave of a survey,…) are not an issue.

Large enough cohorts…

The principle of pseudo-panels is to construct cohorts, i.e. profiles, that group together individuals with behaviours considered to be similar. This assumption is even more plausible if precise profiles are defined. However, this can come at a cost, especially with survey data. The smaller the cohort, the greater the extent of errors when measuring empirical means $\bar{y}_{ct}$ and $\bar{x}_{ct}$ and the greater the temporal variability of the means of individual effects $\bar{\alpha}_{ct}$. There will also be even more bias and imprecision issues with the standard estimator (within estimator) covered earlier (for further details, see "Measurement error model" and Appendix B).

Bias and imprecision of estimators can be limited by increasing the size of cohorts. In practice in empirical studies, it is generally considered that 100 individuals per cohort is enough to ignore sampling errors (and therefore simplify the estimation). This choice is based in particular on the studies of Verbeek and Nijman (1992, 1993). Using simulated data, they conclude that the assumption is reasonable (in the sense that the resulting bias is not too high) for categories with at least 100 individuals. However, they recommend cohorts twice as large to significantly reduce the risk of bias.

…while conserving variability

The larger the cohorts, the lower the extent of measurement errors and the bias and imprecision of the estimators that they generate. But the cohorts' size is not the only parameter to be taken into account. It is quite easy to see that for a given sample size, forming large cohorts means that the number of observations used for the pseudo-panel model will be reduced. For example, let us assume that the cohort is built on the criterion of the year of birth but that the repeated cross-sectional data contain few people from one generation at each date. To reduce potential sample fluctuations, one typical solution is to increase the size of the cohorts by broadening the generations (e.g. by five-year age brackets). However in this case, the variability of observations at a given date is reduced, as the final number of useful observations decreases. Grouping close but different generations also means that the variability of these means is reduced over time. These two elements (number of observations used for the estimation, low variability) are both factors that traditionally reduce the precision of the final estimator. Intuitively, the smaller the number of observations, the less precise the estimation is. However, it is also necessary to observe different values of the variables of interest (that is, to be able to observe their variation over time), in order to assess how strongly they are correlated. This reflects a classic bias-variance tradeoff. Forming large cohorts limits the bias of the estimator but causes variability to be lost, which reduces the precision of the estimators. Verbeek and Nijman (1992) show that the bias of the within estimator traditionally used (see below) can be large if the inter-temporal variability is low in relation to the measurement errors, even when the cohorts are large.

In short, a good selection criterion must: (1) be a characteristic that does not change over time on an individual basis, define a stable (sub-) population, and result from a tradeoff so that (2) large enough cohorts can be formed (3) without losing too much variability. These various constraints highly limit the choice of cohort selection criteria. In practice, many studies use the year of birth as this criterion meets many of

these requirements, is often available in survey data and is stable. Furthermore, depending on the size of cross-sectional samples, close generations can be grouped to create larger or smaller cohorts. Finally, it is important to remember that this dimension is of interest itself in many studies. The cohort effect can then be directly interpreted as a generation effect, which can be interesting to study. In life-cycle analysis in particular, grouping individuals by generation preserves variability on the "age" variable.

Estimation of pseudo-panel models

When the cohort selection criterion has the qualities required to consider model (2) as a fixed effects model, the parameters are generally estimated based on standard panel data estimation techniques. In practice, the estimated model is therefore:

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

We apply a within transformation evoked above, in which, for each cohort, the various variables are centred on the mean of the observed values for the cohort, for all the observation dates. We therefore regress $\bar{y}_{ct} - \bar{y}_c$ on $\bar{x}_{ct} - \bar{x}_c$, where for each variable z, $\bar{z}_c = \frac{1}{T}\sum_{t=1}^{T}\bar{z}_{ct}$. The within estimator is obtained:

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

This allows us to deduce the following cohort effect estimator:

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

In practice, the within estimator is obtained by carrying out first a within transformation then calculating the least squares estimator on these centred variables. However, this has to be done carefully because, as transformed variables are being used, the standard estimator of the variance obtained with the ordinary least squares procedure does not correspond directly to the unbiased estimator of the variance of the within model. It underestimates it. A multiplying factor $(CT - K) / (CT - C - K)$ needs to be taken into account, where C is the number of cohorts, T the number of observation dates and K the number of explanatory variables. In SAS, the Bwithin macro written by Duguet (1999) takes this problem into account (for other Stata and R procedures, see Guillerm, 2015).

The within estimator is obtained in the same way – either by including cohort dummies or via instrumentation. Including cohort dummies in model (3) can be used to directly obtain the fixed effects estimators[5], which sometimes are of interest in themselves. In a life-cycle analysis where cohorts consist of generations, the generation effect could be estimated directly. Again, it is important to be careful, since the estimation of these fixed effects will lack precision if the number of periods is not large enough.

Moffitt (1993) proposes an alternative estimation method using instrumentation. He shows that the within estimator (4) of the pseudo-panel model technically corresponds to the two-stage least squares estimator on individual data (explanatory variables and cohort dummies), where all cohort-time interaction dummies would be used as the instrument. The formal proof is provided in Appendix A. In order to understand the intuition, remember that in the first step of the two-stage least squares procedure the explanatory variables are projected onto the instruments. The projection of x_{it} onto cohort - date interaction dummies corresponds exactly to the empirical mean $\bar{x}_{ct}$, where c is the cohort to which individual i belongs. The second step involves replacing the instrumented variables in the initial model with their projection, in this case regressing y_{it} on $\bar{x}_{ct}$ and the cohort dummies. The estimator obtained is the same as the within estimator (4).

This can simplify the estimation, because we are working directly on the individual data. This analogy also serves as a basis for extending pseudo-panels to dichotomous models (see "Estimation of dichotomous models"). Another advantage of this approach is that other types of more parsimonious instruments can be used. For example, if the year of birth is adopted, a function of the year of birth (e.g. a polynomial function) can be used to build the instrument

5. Direct estimation of the fixed effects is not recommended with individual data as it requires estimating an extensive number of parameters. For pseudo-panels, the number of cohorts is generally limited. If each cohort has approximately 100 individuals, the number of fixed effects to estimate in the pseudo-panel model is divided by as much in relation to the panel model.

rather than dummy variables associated with a partition of the years of birth.

This approach can also be used to find the criteria for grouping individuals in cohorts[6]. Two conditions are required to construct a good instrument. It must first be correlated with the explanatory variables. This is due to the fact that cohorts must have enough variability to allow the estimation of the model at the aggregated level of cohorts. To understand the underlying intuition, we can use the extreme case where these cohort-date interaction dummies would be completely independent from the model's explanatory variables (i.e. that the distribution of these explanatory variables is identical at each date and from one cohort to another). In this case, the empirical means of these variables at a date and cohort level are very similar, which means that the model cannot be estimated. The other feature of a valid instrument is that it must not be correlated with the unobserved determinants of the variable of interest. Moffitt shows that this property is proven if the cohorts are constructed on the basis of a stable criterion and when the size of the cohorts tends to infinity.

Beyond the estimation itself, several remarks can be made, The first of which concerns the choice of explanatory variables. In the standard fixed effects linear model, only the parameters associated with variables that are not constant over time can be identified: the fixed effect "absorbs" the effect of constant variables. In a pseudo-panel model, the aggregation into cohorts artificially creates variability and gives the impression that the parameters associated with the fixed characteristics are identifiable. For example, a variable that is constant on an individual level such as the dummy variable "being a woman" becomes "the proportion of women in the cohort c on the date t" in the pseudo-panel data. The observed temporal variations (normally low) are only due to sampling errors. Introducing these types of variables in the analysis is therefore not recommended.

Some additional technical points

This section provides two extensions to the standard way of handling technical issues with pseudo-panel estimations: taking into account (1) the heteroscedasticity of residual terms and (2) the heteroscedasticity of measurement errors in the estimation. The models presented so far are only suitable if the variable of interest is continuous. With discrete variables, specific methods need to be used. An introduction to this aspect is presented in a third section.

Heteroscedasticity in pseudo-panels

In practice, cohorts vary in size from one to the other and for a given cohort, between one date and another. These size variations may result in heteroscedasticity in model (2). As the precision of the estimator directly depends on this number, varying degrees of error terms are introduced depending on the cohorts. In the presence of heteroscedasticity, the within estimator (4) is unbiased but the estimator of its precision is biased and the statistical tests are therefore invalid.

The efficient within estimator is obtained by weighting the observations by the cohort's size, which means a least squares estimation of the following model:

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

Just as with the homoscedastic model, $K + C$ parameters need to be estimated. This estimation is easy to implement unless the number of cohorts is too large, in which case a within transformation is generally used with the aim of eliminating the fixed effects before estimation. However in this model, a standard within transformation will not eliminate the cohort dummies because the weight assigned to each cohort (n_{ct}) varies over time. Gurgand et al. (1997) show that in this case the efficient within estimator is:

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

where X a matrix of dimension CT × K stacks the line vectors $\bar{x}_{ct}$, y a vector of dimension CT stacks the values $\bar{y}_{ct}$, $(WDW)^{-}$ is the generalised inverse of the matrix WDW, with W the standard within matrix of dimension CT and D the diagonal matrix where the diagonal elements are $1/n_{ct}$.

Measurement error model

The estimation methods presented in the section above do not take into account the fact that the true intra-cohort means noted y^{*}_{ct} and x^{*}_{ct}

6. *For more information, see Moffitt (1993) and Verbeek (2008).*

are measured with errors using the means calculated on the sample (noted $\bar{y}_{ct}$ and $\bar{x}_{ct}$). As stated above, these measurement errors pose two problems: Error in the explanatory variables, results in biased estimators and error in the variable of interest as well as the variability of the cohort effect over time reduce the precision of the estimators. The estimation techniques presented above are implicitly based on the assumption that measurement errors can be overlooked. Otherwise, appropriate techniques are required. The estimators of model (2) proposed by Deaton (1985) therefore rely on measurement error models that take this problem into account. He adapts Fuller' theory (1986) to pseudo-panel estimation.

We write u_{ct} and v_{ct}, the measurement errors:

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

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

When they are integrated into model (2), we obtain:

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \alpha_c + \tilde{\varepsilon}_{ct} \quad (8)$$
$$c = 1,\ldots,C \quad t = 1,\ldots,T$$

where $\tilde{\varepsilon}_{ct} = \varepsilon^*_{ct} + u_{ct} - v_{ct}\beta$. We show that this residual value is correlated to $\bar{x}_{ct}$.

The estimator of the parameter β proposed by Verbeek and Nijman (1993) relies on a parametric specification of the measurement error and its correlation with the variable of interest (for more information, see Appendix B). It gives:

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

$\sum$ and σ correspond, respectively, to the variance-covariance matrix of measurement errors in x^*_{ct} and to the covariance between measurement errors in x^*_{ct} and y^*_{ct}. They are generally not known. Deaton suggests estimating them on the individual data:

$$\hat{\Sigma} = \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} \hat{\Sigma}_{ct} \quad (10)$$

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

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

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

Several types of convergence can be considered in the case of pseudo-panel estimations as several parameters come into play: N the number of individuals observed at each date, C the number of cohorts, n_{ct} the size of cohorts and T the number of observation dates.

Intuitively when the cohorts' size increases, the larger the cohorts, the more the intra-cohort means – that is, the estimators of the true intra-cohort means – are precise. Measurement errors become negligible and we find the standard within estimator.

The within estimator has an asymptotic bias when the size of cohorts is fixed but a lower variance than the Verbeek and Nijman estimator (for more information, see Verbeek & Nijman, 1993). This reflects again a classic bias-variance tradeoff.

Estimating dichotomous models

The previous estimators are only suitable for linear models and not when the variable of interest is binary. For this, specific estimation techniques need to be used. With panel data, switching from linear to non-linear estimation of a fixed effects model is in itself difficult. The use of pseudo-panels makes the estimation even more complex. To date, few studies have implemented the estimation methods developed for such models. Only the broad principles are given here.

The model to be estimated appears in the following form:

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

where $\tilde{y}_{it}$ is a latent variable (unobserved). The value of the observed binary variable y_{it} is 1 if $\tilde{y}_{it}$ is positive and 0 otherwise. x_{it} is a vector of explanatory variables, α_i is an individual fixed effect and ε_{it} an error term which

is generally assumed to follow a logistic or a normal distribution.

As in the linear case, the goal is to estimate a fixed effects model. With panel data, there are two standard estimation techniques: the conditional *logit* which consists in transforming data to eliminate the fixed effect (see, for example, Davezies, 2011) or the Chamberlain approach, (Chamberlain, 1984).

The Chamberlain approach is the starting point of the estimation method using pseudo-panel data proposed by Collado (1998). It consists in writing the relationship between the individual fixed effect and the covariates:

$$\alpha_i = x_{i1}\lambda_1 + \ldots + x_{iT}\lambda_T + \theta_i \qquad (13)$$

where $E(\theta_i \mid x_{i1},\ldots,x_{iT}) = 0$.

Substituting (13) into (12) gives the reduced form:

$$\tilde{y}_{it} = x_{i1}\pi_{t1} + \ldots + x_{iT}\pi_{tT} + \theta_i + \varepsilon_{it} \qquad (14)$$

$i = 1,\ldots,N \qquad t = 1,\ldots,T$

where $\pi_{ts} = \beta + \lambda_s$ if $s = t$ or $\pi_{ts} = \lambda_s$ otherwise. The error term $\theta_i + \varepsilon_{it}$ is not correlated with the covariates.

In the absence of panel data, the complete series of covariates is not available for a single individual. Model (14) can therefore not be directly estimated. Collado (1998) suggests estimating this model by replacing in (14) each individual value of the covariates x_{it} with the cohort mean of the individual's cohort, i.e. $\bar{x}_{ct}$. Here the cohorts are constructed following the same rules as those presented in the linear framework (see above). It should be noted that the variable of interest y_{it} is not aggregated.

Substituting individual observations with the intra-cohort means of explanatory variables introduces measurement errors into the model (the sum of the individual deviation, the intra-cohort mean and the sampling error mean) and a correlation between the error term and the covariates. Collado proposes two estimators for the β parameter. These estimators are calculated in two steps. The first, applied to both estimators, involves a quasi-maximum likelihood estimate of the π_{ts} parameters. The two proposed estimators of the β parameter are then deduced from the estimator of the π_{ts} parameters. One is calculated by minimum distance and the other by doing a within transformation on the data. The within estimator has the advantage of being easier to calculate but is not efficient, unlike the minimum distance estimator.

Moffitt (1993) proposes an alternative estimation technique, based on the parallel drawn between pseudo-panel estimation and instrumentation (see above). In the linear framework, estimating the model using the pseudo-panel method is equivalent to instrumenting using cohort-date interaction dummies. Moffitt proposes this same instrumentation to estimate model (12).

An example of pseudo-panel application: effect of age and generation on household wealth

There are many examples of pseudo-panels being used in econometric work on consumption (e.g. Gardes et al., 2005; Marical & Calvet, 2011) and in life-cycle analysis (see box). Here we propose a basic application of pseudo-panel methods to estimate age effects on household wealth. This application is highly simplified with respect to the issue of wealth accumulation and is only meant to provide a practical example of these methods. A more comprehensive analysis of this issue can be found in Lamarche and Salembier (2012).

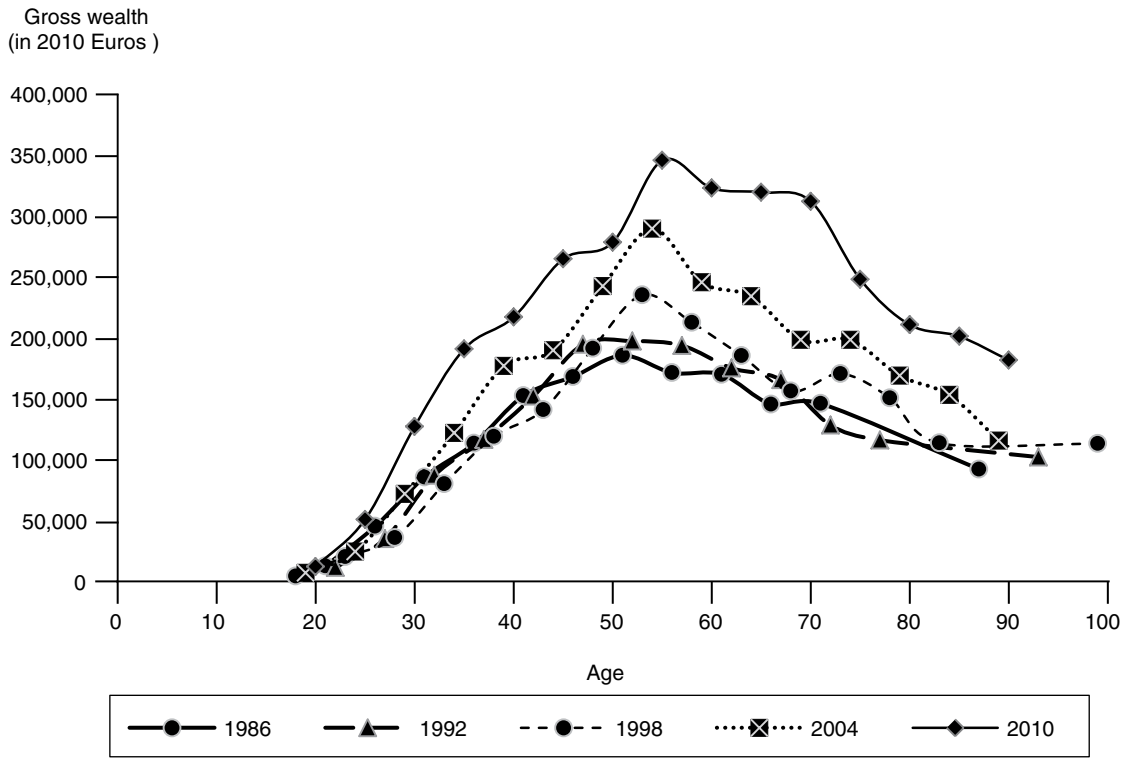
We will use the French Household Wealth surveys (*enquête Patrimoine*), conducted every six years since 1986[7], which provides five observation dates (1986, 1992, 1998, 2004 and 2010). In the survey, households are asked about their real estate, financial and professional assets. The sum of these assets provide the gross wealth (calculated in constant 2010 Euros). In 2010, the survey underwent major changes to better assess households' wealth. In particular, the categories of households with the highest wealth were oversampled and assets such as cars, household equipment, jewellery, and artwork were taken into account. To avoid biasing the changes between 2004 and 2010, these methodological changes were for the most part neutralised in the wealth calculations.

7. *In 1986 and 1992, the name of the French Household Wealth survey was* enquête Actifs financiers.

To briefly describe the issue, the aim is to study saving patterns at different ages. In the initial version put forward by Modigliani and Brumberg (1954), the life-cycle theory assumes that people adopt an intertemporal approach to allocating their income. Over their lifetimes, they experience three periods during which their earnings, and their savings and consumption behaviours differ. At the beginning of their career their income tends to be low and they spend more than they earn (dissaving). Then, throughout their career, their income increases, they save and accumulate wealth as they prepare for their income to drop when they retire. Wealth accumulation therefore follows a bell curve pattern with age. It is difficult to test the life-cycle theory, by estimating for instance changes in wealth with age. This type of estimation would require the same individuals to be followed over a very long period, which is quite impossible. As stated earlier, a cross-sectional estimation would not be relevant since it does not allow the distinction to be made between the effects of age and generation. With this very simple case of estimating the effect of age, the next two graphs can be used as a starting point for a typical exploratory approach. Each Household Wealth survey is used to represent the change in mean gross wealth according to age (Figure I). The profiles obtained seem to confirm the life-cycle theory beyond a doubt. A bell curve is obvious with an increase in gross wealth until about 60 years of age, followed by a drop. However, part of this profile can be explained by the fact that different generations are observed at each date. Economic context, the age at which people begin working, and taxes are all characteristics shared by the individuals of a same generation that have an effect on accumulated household wealth. They also explain differences in wealth at the same age between different generations. Long term data are required to separate these two effects.

To attempt to capture this "generation" dimension, all the surveys are stacked so as to obtain observations for individuals from identical generations at different dates (and therefore different ages). We obtain five observations,

Figure I
Household wealth according to age in 1986, 1992, 1998, 2004 and 2010



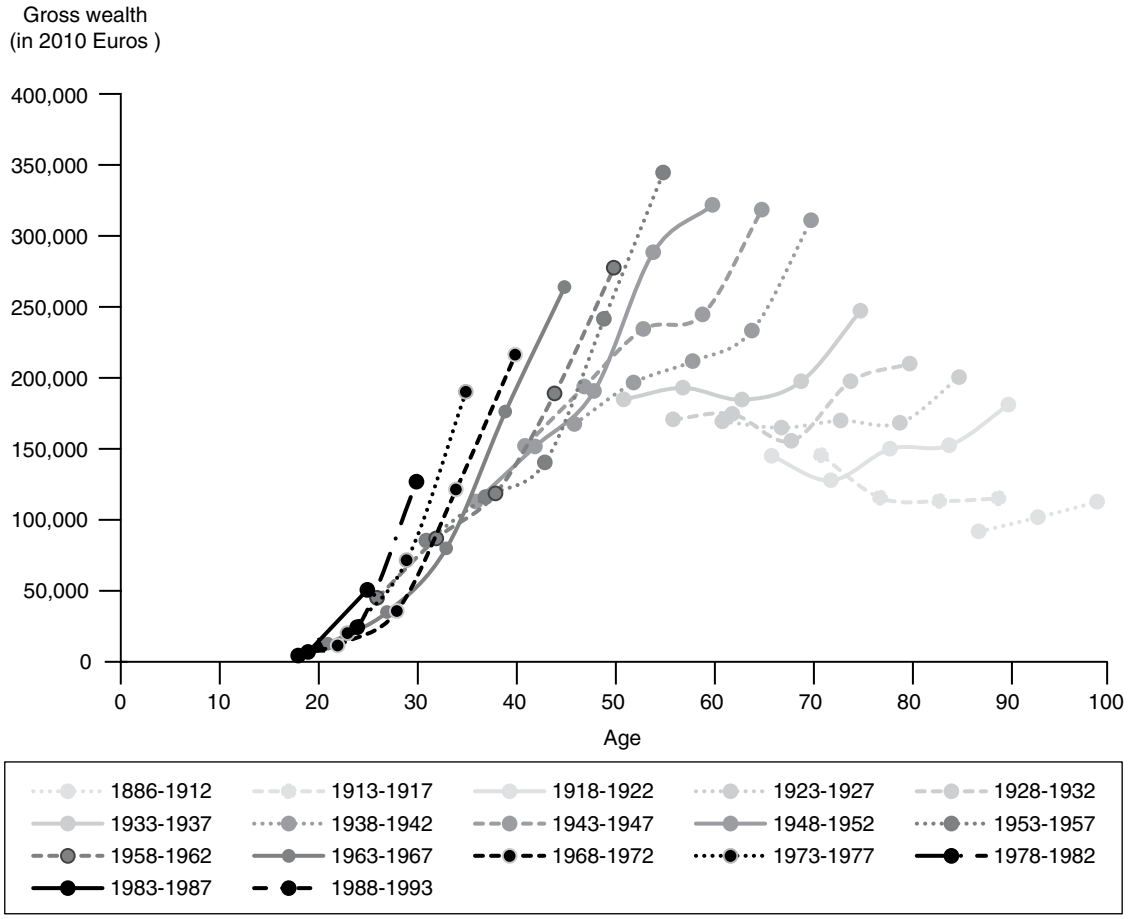
Reading note: respondents of the 2010 Household Wealth Survey had an average wealth of 278 156 at age 48 to 52. The centre of each age group is represented on the x-axis (e.g. 65 for the 63-67 age group).
Coverage: households residing in France (excluding Mayotte).
*Source: Insee, Household Wealth Surveys (*enquêtes Patrimoine*), 1986 to 2010.*

corresponding to the average wealth at five different ages for almost all the generations (except for the youngest or oldest). In theory, one profile for all the generations, defined by year of birth, could be represented. However in practice, we are confronted with the problem that in a survey sample, the number of individuals from a given generation is not very high. These estimations are therefore very imprecise. To offset this problem, we define cohorts as the grouping of adjacent generations (five in Figure II).

Figure II shows, for each cohort, the profile of wealth accumulation by age. It is very different from the profile presented using only the cross-sectional dimension. Contrary to what Figure I suggests, wealth continues to grow well over the age of 60. As underlined by Lamarche and Salembier (2012), several factors explain this stylised fact. Even beyond retirement, households may want to save in order to leave an inheritance or simply build up contingency savings (should they become dependent). Furthermore, the most elderly may decide not to sell their real estate assets to avoid moving and the particularly high cost that this entails (see Angelini & Laferrère, 2012). It should also be highlighted that the accumulation of wealth with age partially results from changes in generation composition observed at extreme ages. The scope of the survey only examines private households and therefore does not include elderly people in retirement homes. Wealthier households also have a longer life expectancy than others (and likely more assets).

Figure II compares the average wealth of different cohorts at the same age. There are

Figure II
Household wealth according to age from one generation to another



Reading note: respondents of the 2010 Household Wealth Survey had an average wealth of 278 156 at age 48 to 52. The centre of each age group is represented on the x-axis (e.g. 65 for the 63-67 age group).
Coverage: households residing in France (excluding Mayotte).
*Source: Insee, Household Wealth Surveys (*enquêtes Patrimoine*), 1986 to 2010.*

sometimes significant differences. The vertical deviation between the curves corresponds to the generation effect and a period effect. For example, let us assume that these period effects, which correspond to the increase in household wealth over time (again, we are working in constant 2010 Euros to avoid including inflation) are negligible. This resolves the problem of identifying age, cohort and period effects (see Box). Under this assumption, the graph suggests that, at the same age, each generation has accumulated more wealth than the previous. The difference is considerable between generations born in the 1950s who experienced the post-World War II economic boom (1945-1975) and previous wartime generations. The decrease in wealth after the age of 60 observed in Figure I likely stems more from significant differences in wealth between these two generations than dissaving at retirement.

Pseudo-panel econometric modelling provides a more accurate quantification of the age effects seen in Figure II. It is based on a model written out on an individual basis, as follows:

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

$\log Pat_{it}$ is the logarithm for the wealth of the individual i on date t, age_{it} is their age on date t. Let us assume here that the effect of age on wealth is identical for all generations and that it has a quadratic profile[8]. α_i is an individual fixed effect. It estimates the impact of unobserved fixed characteristics of individual i on his/her wealth.

The pseudo-panel model that is estimated in practice is as follows:

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

where for each variable z, $z_{gt} = E(z_{it} \mid i \in g,t)$. These values are not observed. They are estimated by the intra-cohort means $\bar{z}_{gt} = \frac{1}{n_{gt}} \sum_{i \in g,t} z_{it}$ calculated from available data, where n_{gt} is the number of individuals of cohort g observed on date t.

Two practical remarks need to be made. The first concerns the composition of the sample. The estimation relies on the fact that $\bar{\alpha}_{gt}$ is fixed over time. This can be called into question. As mentioned above, for the oldest generations, two composition effects come into play. First, the wealthiest households have a longer average life expectancy, and secondly,

8. The accumulation of wealth with age between different generations only differs in level. The model could be made more complex by integrating interaction terms between age and generation.

Box

AGE, COHORT AND PERIOD EFFECTS

Simultaneously estimating an age, cohort and period effect is a recurring problem that already existed before pseudo-panels, but which is raised in the same way for individual data and pseudo-panel data. The difficulty stems from the collinearity between the three variables (age + cohort = period), i.e. from the fact that individuals of the same age and the same generation cannot be observed at different dates.

It is generally resolved by treating age, cohort and period effects as additive. The model therefore simply includes a set of age, cohort and period dummies without interaction terms. This additivity assumption is significant. It leads us to assume that the age effect, for instance, is common to all generations. In the case of this model, the literature proposes two primary solutions for resolving the identification problem. The first involves imposing identifying constraints on the model (in addition to the nullity of a coefficient for each dimension and an identifying constraint in the presence of a constant in the model). Mason et al. (1973) show that we can simply assume that two coefficients from a single dimension (age, cohort, or period) are equal. Different identifying constraints lead to different estimations and must be discussed on a case by case basis. Rodgers (1982) disagrees with this practice and proposes replacing one of the effects with variables that correlate with it, for example macro-economic variables in the place of the period effect. Readers interested in this issue may refer to Hall et al. (2007) for a literature review on the subject, or Yang and Land (2013).

the Household Wealth survey does not survey people in retirement homes. On the other end, the Household Wealth survey only includes a small number of very young households, which are probably very specific. To work on a stable population, we limit ourselves to households over the age of 26 and under 80[9]. The second remark concerns the size of cohorts. Cohorts group together several successive generations. Limiting the number of these successive generations reduces the risk of aggregating heterogeneous behaviours. However this means that estimations are based on very few observations per cohort and therefore risk being very imprecise. To illustrate this issue, the model was estimated using relatively broad cohorts (three, five and ten years) (Table C1 in the Appendix).

The table below shows the results of pseudo-panel estimations. For comparative purposes, the results obtained from cross-sectional regression (the data from the five successive surveys are stacked) and the estimations taking measurement errors into account are also presented.

Figure III shows the effect of age on wealth as estimated using both the cross-sectional and pseudo-panel approaches[10]. The two estimations show a bell curve relationship between wealth and age. From cross-sectional data, we estimate that wealth begins to decrease at age 58. The pseudo-panel estimation gives a much higher turning point, around age 70. So when the generation effect is taken into account, the decrease in wealth is observed much later than a cross-section approach suggests.

As the model is log-linear, 100 x [exp(α_g) - 1], where α_g is the coefficient associated with the generation *g* in the model (Table C2 in the Appendix and Figure IV below), corresponds to the effect on wealth (measured in %) of belonging to generation *g* rather than to the

9. Furthermore, as means are sensitive to extreme values, some very high net worth households were removed from the analysis. The few observations corresponding to zero net worth were also removed since logarithm modelling is used.
10. The polynomial of degree 2 is therefore represented: $\beta_0 + \beta_1 age + \beta_2 age^2$ where the coefficients are estimated from cross-sectional data and pseudo-panel data.

Table
Estimation of age effects

		Pseudo-panel Estimations		
	Cross-sectional data	3-year generations	5-year generations	10-year generations
		Within estimator		
Intercept	4.59***	4.80***	4.65***	4.89***
	(0.127)	(0.383)	(0.437)	(0.542)
Age	0.223***	0.197***	0.199***	0.193***
	(0.0052)	(0.0142)	(0.016)	(0.0212)
Age^2	- 0.0019***	- 0.00140***	- 0.00136***	- 0.00136***
	(0.0000493)	(0.000135)	(0.000145)	(0.0002)
		Measurement error model		
		Verbeek and Nijman estimator (9)		
Intercept		4.63***	5.05***	5.63***
		(0.279)	(0.307)	(0.398)
Age		0.203***	0.187***	0.162***
		(0.0104)	(0.0127)	(0.0172)
Age^2		- 0.00143***	- 0.00128***	- 0.00102***
		(0.000092)	(0.00012)	(0.00016)
Number of observations	43 117	94	57	31

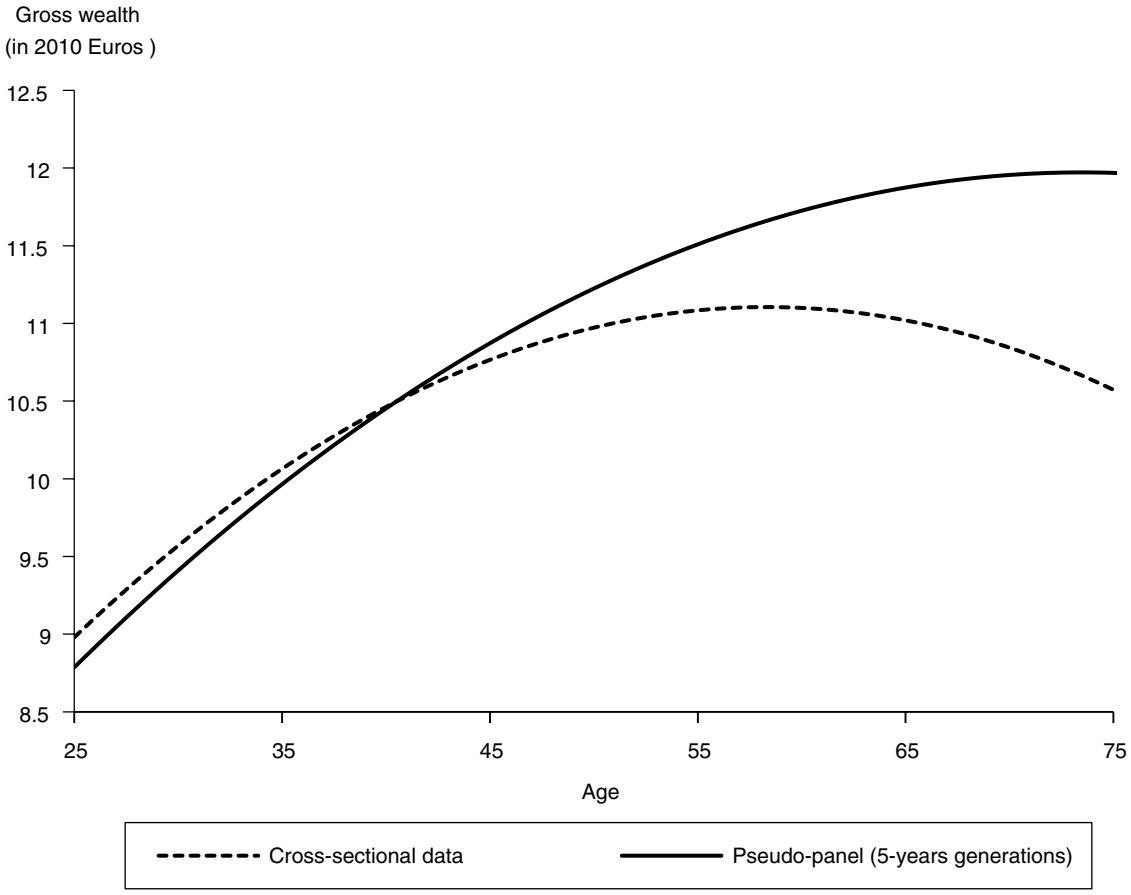
Note: the constant is calculated using the birth years 1951-1953 as the baseline generation for 3-year generations, 1953-1957 for 5-year generations and 1953-1962 for 10-year generations. Standard deviations were calculated by bootstrapping for the measurement error model.
****. **. * indicate the significance level of the coefficients at 1%, 5% and 10% respectively. The number of individuals observed in the different generations is presented in Table C2 in the Appendix.*
Coverage: households residing in France (excluding Mayotte).
*Source: estimation based on the French Household Wealth Surveys (*enquêtes Patrimoine*).*

1951-1953 generation (generation of reference). For example, being born between 1939 and 1941 rather than between 1951 and 1953 has a negative effect on household wealth, estimated at 100 x [exp(– 0.44) – 1] = – 35.6%. We estimate that between the 1939-1941 and 1951-1953 generations, household wealth increased on average by 3.7% annually. Its growth then slowed down.

The sensitivity of estimations to the cohort grouping criteria does not seem too high in this case. Figure IV shows the generation effects estimated using the pseudo-panel methods based on three ranges chosen to construct the generations. Unsurprisingly, the greater the range, the smoother the profile. In all cases, we observe a significant increase in the wealth of successive generations until the baby-boom generations, followed by stagnation. For the youngest generations, the diagnosis seems to diverge depending on the selection criteria, but these changes are never significant (see Table C2 in the Appendix). This uncertainty stems from the fact that estimations are based on smaller samples (these generations are not observed in the older surveys), as shown in Table C1 (Appendix C). It can also be seen that, as expected, the precision of the estimators of coefficients β_1 and β_2 is greater for three-year generations than for five or ten-year generations.

The Verbeek and Nijman estimator, which takes into account measurement errors, was also calculated directly with the estimator formula. As the estimator formula suggests, in theory, the direction of bias is not known and changes depending on the range adopted to define the generations. The estimations differ little from those obtained with the within estimator, except for the 10-year generations. □

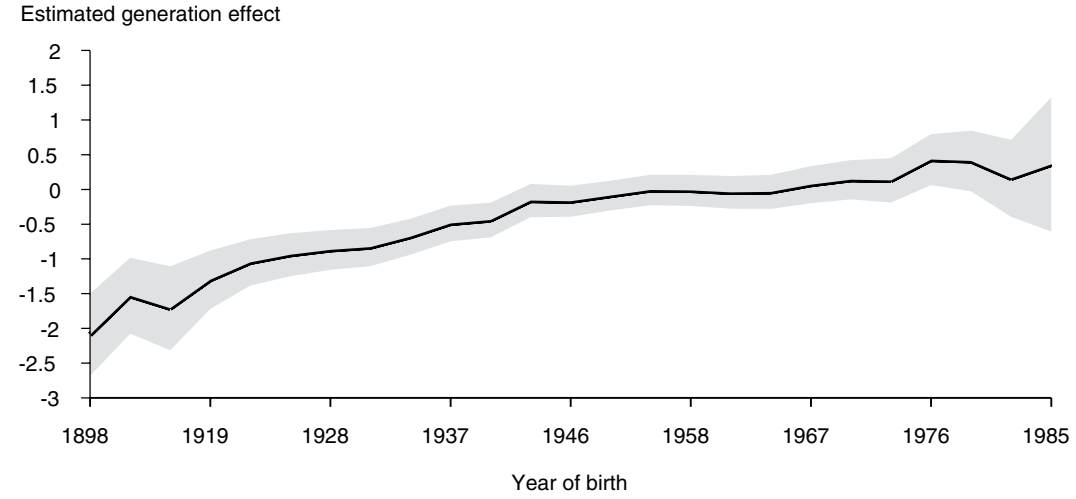
Figure III
Household wealth according to age as estimated by models



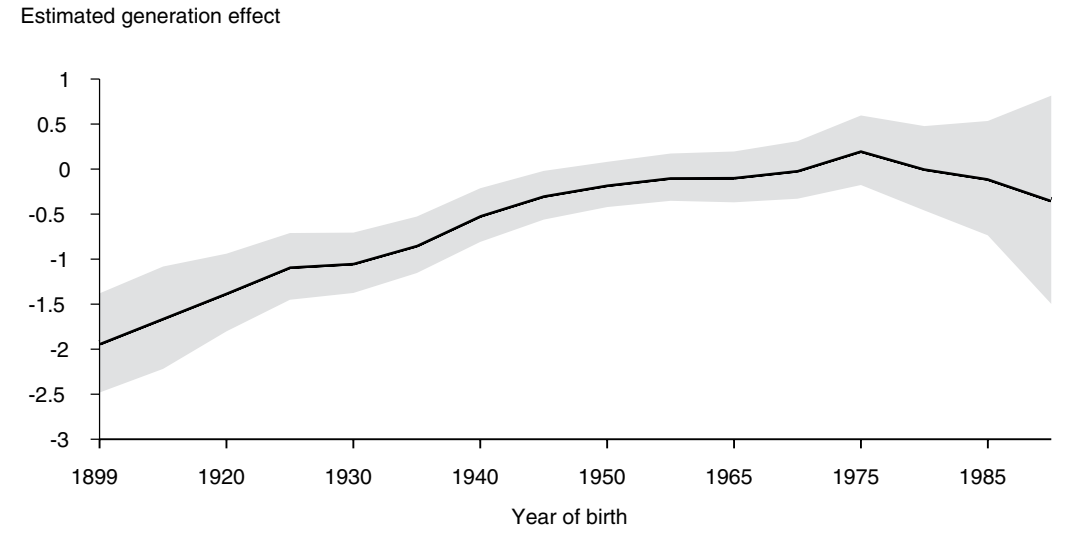
Reading note: at 65, the gross wealth logarithm as estimated by the pseudo-panel model is 11.87.
Coverage: households residing in France (excluding Mayotte).
*Source: estimation based on the French Household Wealth Surveys (*enquêtes Patrimoine*).*

Figure IV
Generation effects estimated using the pseudo-panel method

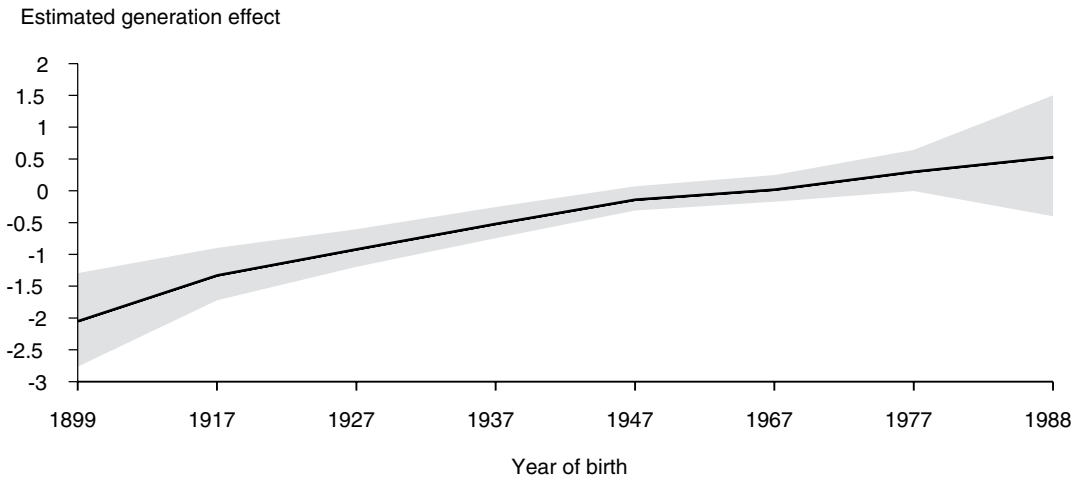
(a) 3-years generations



(b) 5-years generations



(c) 10-years generations



Reading note: the generation effect estimated by the pseudo-panel model (3-year generation) for the 1939-1941 generation is – 0.44, which corresponds to 35.6% lower gross wealth than the baseline generation (1951-1953). The grey area corresponds to a 5% confidence interval.
Coverage: households residing in France (excluding Mayotte).
*Source: estimation based on the French Household Wealth Surveys (*enquêtes Patrimoine*).*

BIBLIOGRAPHY

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.

APPENDIX

A. PSEUDO-PANEL AND INSTRUMENTATION

Moffitt (1993) shows that estimation using the pseudo-panel approach and estimation by instrumenting using cohorts-date interaction dummies provide the same estimator.

Estimation via two-stage least squares follows the following two steps:

Step 1: Projection of explanatory variables onto the instrument.

If the individual fixed effect α_i is written as the sum of a fixed effect cohort α_c and an individual deviation $v_i = \alpha_i - \alpha_c$, model (1) would be as follows:

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

x_{it} is potentially correlated to v_i. Therefore, x_{it} is instrumented using cohort indicators in interaction with the time indicators. The first step is to project x_{it} onto the instrument. The predicted value of x_{it} in this regression corresponds to the intra-cohort mean $\bar{x}_{ct}$.

Step 2:

x_{it} is replaced by its predicted value in (17). y_{it} is therefore regressed on $\bar{x}_{ct}$ and the cohort indicators, which gives the same estimator as the within estimator (4).

B. DETAILS ON THE ESTIMATION OF THE PARAMETERS OF A MEASUREMENT ERROR MODEL

$\bar{x}_{ct}$ and $\bar{y}_{ct}$ are observations with errors of the true intra-cohort means x^*_{ct} and y^*_{ct}. u_{ct} and v_{ct} are the measurement errors:

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

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

They are assumed to be normally distributed:

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

where n is the size of the cohorts.

Integrating (18) and (19) into model (2) gives:

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

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

The correlation between this residual value and the covariates gives:

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

In general it is not zero. The estimator of the least squares of $\bar{y}_{ct}$ on $\bar{x}_{ct}$ is therefore biased.

Model (21) is a fixed effects model. After a within transformation, model (21) becomes:

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

We show that:

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

From this equation, an expression of β is deduced:

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

Estimator (9) is the empirical counterpart of this expression.

When only the explained variable is observed with error, the within estimator is without bias but it is less precise than a model without measurement errors. When the measurement error only relates to the explanatory variables, it leads to an attenuation bias (the absolute value of the within estimator converges towards a lower value than the absolute value of parameter β).

C. APPLICATION OF PSEUDO-PANELS TO THE FRENCH HOUSEHOLD WEALTH SURVEY (*ENQUÊTE PATRIMOINE*)

Table C1
Cohorts' size

3-year generations

Generation (year of birth)	Year				
	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

Reading note: in the 1986 French Household Wealth Survey, 373 individuals born between 1954 and 1956 were surveyed.
Coverage: households residing in France (excluding Mayotte).
*Source: Insee, French Household Wealth surveys (*enquêtes Patrimoine*).*

5-year generations

Generation (year of birth)	Year				
	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

Reading note: in the 1986 French Household Wealth Survey, 615 individuals born between 1953 and 1957 were surveyed.
Coverage: households residing in France (excluding Mayotte).
*Source: Insee, French Household Wealth surveys (*enquêtes Patrimoine*).*

10-year generations

Generation (year of birth)	Year				
	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

Reading note: in the 1986 French Household Wealth Survey, 863 individuals born between 1953 and 1962 were surveyed.
Coverage: households residing in France (excluding Mayotte).
*Source: Insee, French Household Wealth surveys (*enquêtes Patrimoine*).*

Table C2
Estimated generation effects

3-year generations	
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	*ref.*
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)

5-year generations	
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	*ref.*
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)

10-year generations	
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	*ref.*
1963-1972	0.038
	(0.107)
1973-1982	0.32*
	(0.165)
1983-1993	0.55
	(0.485)

*Reading note: The estimated coefficient of the 1939-1941 generation in the model is – 0.44, which means that being born between 1939 and 1941 rather than between 1951 and 1953 (reference generation) has a negative effect on wealth, estimated at 100 x [exp(– 0.44) – 1] = – 35.6 %. ***, **, * indicate a significance level of the coefficients at 1%, 5% and 10% respectively.*
Coverage: households residing in France (excluding Mayotte).
*Source: Insee, French Household Wealth surveys (*enquêtes Patrimoine*).*