Moreno, Héctor; Bourguignon, François; Dang, Hai-Anh

Working Paper

On Synthetic Income Panels

GLO Discussion Paper, No. 809

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Moreno, Héctor; Bourguignon, François; Dang, Hai-Anh (2021) : On Synthetic Income Panels, GLO Discussion Paper, No. 809, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/232302

On Synthetic Income Panels

Héctor Moreno, François Bourguignon and Hai-Anh Dang*

March 30th, 2021

Abstract

In many developing countries, the increasing public interest in monitoring economic inequality and mobility is hindered by the scarce availability of longitudinal data. Synthetic panels based on matching individuals with the same time-invariant characteristics in consecutive cross-sections have been recently proposed as a substitute to such data. We extend the methodology to construct such synthetic panels in several directions by: a) explicitly assuming the unobserved or time variant determinants of (log) income are AR(1) and relying on pseudo-panel procedures to estimate the corresponding auto-regressive coefficient; b) abstracting from (log) normality assumptions; c) generating a close to perfect match of the terminal year income distribution and d) considering the whole income mobility matrix rather than mobility in and out of poverty. We exploit the cross-sectional dimension of a national-representative Mexican panel survey to evaluate the validity of this approach. With the median estimate of the AR coefficient, the income mobility matrix in the synthetic panel closely approximates that of the genuine matrix observed in the actual panel, except for out-lying values of the AR coefficient.

Keywords: Synthetic panel, income mobility, Mexico.

JEL codes: D31, I32.

*Moreno (ahector.morenom@gmail.com) is with Oxford Poverty and Human Development Initiative, University of Oxford; Bourguignon (francois.bourguignon@psemail.eu) is with Paris School of Economics; Dang (hdang@worldbank.org) is with GLO and the World Bank. We gratefully acknowledge comments on previous versions from Andrew Clark, Dean Jolliffe, Marc Gurgand, Jaime Montaña, Christophe Muller, Umar Serajuddin, Elena Stancanelli, and attendants at internal seminars at the Paris School of Economics. Moreno acknowledges financial support from CONACYT, CEEY and from the French *Agence Nationale de la Recherche* within the JPI MYBL framework (Award no. ANR-16-MYBL-0001-02). Dang acknowledges financial support from DFID through its Knowledge for Change Program (KCP) with the World Bank. Earlier versions of this paper circulated with the title "On the Construction of Synthetic Panels" presented at NEUDC 2015 at Brown University (USA), and Moreno (2018).

1. Introduction

Measuring income mobility is inextricably linked to the measurement of inequality and poverty. The incomes of two persons A and B may be very different at times t and t'. But can this difference be truly considered as inequality if they switch their incomes between t and t'? Likewise, should a person with an income above the poverty line in period t be considered as non-poor if this same individual's income falls below the line in period t'? Clearly, this may depend on how much above the poverty line she was in the first period and how much below in the second period. Measuring inequality and poverty may thus be misleading if one uses only a snapshot of income disparities at a single point of time instead of considering a sequence of individual incomes over multiple periods.

Longitudinal or panel data that permit the analysis of dynamics of individual incomes are seldom available in developing countries. Yet, snapshots of income distributions can be estimated from repeated cross-sectional household surveys that have become increasingly available. A recent idea was to construct *synthetic* panel data based on these data by appropriately matching individuals in the two cross-sections with the same time invariant characteristics but with the appropriate age difference in (two) consecutive cross-sections. Such synthetic panels potentially offer advantages over real panel data. They may cover a larger number of periods and they suffer much less from typical panel data problems like attrition, non-response and, to some extent, measurement error (Verbeek, 2007). But their reliability depends on the quality of the matching method.

This approach has received much attention recently following two strands of the literature.[1] The early literature followed Dang *et al.* (2014) methodology that allows computing bounds of income mobility (i.e., in and out of poverty). This procedure matches individuals with identical time-invariant characteristics and assumes that part of the (log) income that is independent of these characteristics is normally distributed across two periods with a theoretical correlation coefficient equal to 0 or 1 for the upper and lower bounds respectively.

[1] Bourguignon *et al.* (2004) was an earlier attempt in the same direction using the first two moments of the income distribution to estimate rho. More recently, Kraay and van der Weide (2017) use the first two moments of aggregate data to provide bounds of individual level income mobility over long periods.

While this study also suggests narrowing the bounds by obtaining empirical values for this correlation coefficient from auxiliary datasets, a more recent strand of the literature follows methodological refinements in Dang and Lanjouw (2013) that collapse these bounds of mobility into point estimates based on a correlation coefficient estimated through pseudo-panel techniques. [2]

Unsurprisingly, the properties of such synthetic panels strongly depend on the assumptions being made and the way key parameters are estimated. In the methodology designed by Dang and Lanjouw (2013), for instance, the bi-normality assumption made on the joint distribution of initial and final (log) incomes – conditionally on time invariant characteristics - and the way the associated coefficient of correlation is estimated strongly influences the synthetic income (and poverty) mobility matrix. As this coefficient is bound to have a strong impact on the extent of estimated mobility, the estimation method and its precision clearly are of first importance.

The present paper improves on previous works by offering a more rigorous treatment of the estimation of the correlation coefficient that explicitly relies on an AR(1) specification. It also improves previous methodologies by departing from the normality assumption and offers a quasi-perfect fit to both the initial and final cross-sectional distributions. Finally, we significantly extend the focus in existing studies on poverty mobility to the whole income mobility matrix. These various extensions could enable us to provide richer and more accurate analysis using synthetic panels than are available in existing studies.

The validity and the precision of the synthetic panels constructed with our proposed method are tested by comparing the synthetic mobility matrix obtained from the initial and terminal cross-sections of a Mexican panel household survey between 2002 and 2005 and the observed actual matrix in that survey. Although no formal test is possible on a single observation, the results are encouraging as the synthetic joint distribution of initial and final incomes is rather close to the joint distribution in the actual (genuine) panel. However, simulations performed by allowing the AR(1) coefficient to vary within its estimated confidence interval show

[2] These synthetic panel techniques have been applied to survey data in a number of countries from Africa, Europe, Latin America, and Asia (e.g., Ferreira *et al.* (2013); Cruces *et al.* (2015); Beegle *et al.* (2016); OECD (2018)). Dang *et al.* (2019) offer a recent review of studies that employ these synthetic panel techniques.

somewhat high variability of the synthetic mobility matrix and associated income mobility measures. This suggests that caution should be exercised in analyzing income mobility based on synthetic panel techniques.

This paper consists of five sections. We describe our proposed methodology to construct synthetic panels based on AR(1) stochastic income processes in the next section. To highlight the new contributions, we also offer some brief comparison between our method and previous studies in this section. We subsequently describe in Section 3 the survey data that we analyse, the setup of the income model, and the calibration method for the correlation coefficient. Section 4 presents the key results and compares the central estimate of the synthetic income mobility matrix and various mobility measures to those obtained from the authentic panel. We also offer robustness checks in this section. We finally conclude in Section 5.

2. The construction of a synthetic income panel

2.1 Matching techniques and the synthetic panel approach

Consider two rounds of independent cross-section data at time t and t', and let $y_{i(\tau)\tau'}$ denote the (log) income in period τ' of an individual i observed in period τ.[3] From these repeated cross sections, we only observe $y_{i(t)t}$ and $y_{i(t')t'}$, the incomes of different individuals in each period. Constructing the synthetic panel is somehow 'inventing' a plausible value for $y_{i(t)t'}$, which is the unobserved (log) income in period τ' of the same individual i in period τ.

The first step is to account for the way in which time-invariant individual attributes, z, may be remunerated in a different way in periods τ and τ'. To do so, an income model defined exclusively on the time-invariant attributes observed in the two cross-sections is estimated with the OLS method

$$y_{i(\tau)\tau} = z_{i(\tau)}\beta_\tau + \varepsilon_{i(\tau)\tau} \quad \tau = t, t' \qquad (1)$$

[3] This notation is borrowed from Moffit (1993).

where β_τ represents the vector of 'returns' to time-invariant individual attributes, z, and $\varepsilon_{i(\tau)}$ denotes a 'residual' that stands for the effect of time-varying individual characteristics and other unobserved time-invariant attributes. Time-invariant attributes may include years of birth, regions of birth, education, and parent's education. (We return to discuss these variables this in Section 3.2). For now, it is just enough to stress that it would not make sense to introduce time-varying characteristics in the income model (1). Some of them may be observed either in the initial or the terminal survey period, but their values for both periods are essentially unknown in the repeated cross sections.

Let $\hat{\beta}_\tau$, $\hat{\varepsilon}_{i(\tau)\tau}$, and $\hat{\sigma}_\tau^2$ respectively denote the vector of estimated returns, the corresponding residuals, and their variance as obtained from the following OLS regression:

$$y_{i(\tau)\tau} = z_{i(\tau)}\hat{\beta}_\tau + \hat{\varepsilon}_{i(\tau)\tau} \quad \tau = t, t' \qquad (2)$$

Consider an individual i observed in the first period t. Since the individual attributes $z_{i(\tau)}$ are time-invariant, we have $z_{i(t)} \equiv z_{i(t')}$ by definition and can replace $z_{i(\tau)}$ with either of these two terms. Part of the dynamics of her income between t and t' stems from the change in the returns of fixed attributes, or $z_{i(t)}(\hat{\beta}_{t'} - \hat{\beta}_t)$ and can be inferred from the OLS estimates. The remaining is the change in the residual term: $\hat{\varepsilon}_{i(t)t'} - \hat{\varepsilon}_{i(t)t}$. However, the first term in this difference is not observed. The challenge in constructing the synthetic panel is thus to find a plausible value for it. Let $\tilde{\varepsilon}_{i(t)t'}$ be that 'virtual' residual. At this stage, the only information available is its distribution in the population.

2.2 Previous approaches

In their earlier attempt at constructing synthetic panels, Dang *et al.* (2014) assume the virtual residual at time t' and the initial residual at time t are jointly normally distributed with an arbitrary correlation coefficient ρ. The synthetic income mobility process can be described by the joint cdf:

$$\Pr\left(y_{i(t)t} \leq Y; y_{i(t)t'} \leq Y'\right) = N\left[\frac{Y - z_{i(t)}\hat{\beta}_t}{\hat{\sigma}_t}, \frac{Y' - z_{i(t)}\hat{\beta}_{t'}}{\hat{\sigma}_{t'}}; \rho\right]$$

where $N(.)$ is the cumulative probability function of a bi-normal distribution with correlation coefficient ρ.

Dang *et al.* (2014) consider the two extreme cases of $\rho = 0$ and $\rho = 1$, so as to obtain an upper and a lower limit on mobility. Applying this approach to the probability of moving in or out of poverty in Peru and in Chile, the corresponding ranges proved, not surprisingly, to be rather broad. In other words, the change $\left(\hat{\beta}_{t'} - \hat{\beta}_t\right)$ in the returns to the time-invariant attributes was playing a limited role in explaining income mobility.

In a later, unpublished paper, Dang and Lanjouw (2013) generalize the preceding approach by considering a point estimate rather than a range for the correlation between the initial and terminal residuals. Their method consists of approximating the correlation between the (log) individual incomes in the two periods t and t', ρ^y, by the correlation between the mean incomes of cohorts in the two samples, ρ^{y_c}, as in previous pseudo-panel analysis. The covariance between (log) incomes is approximated by $cov_y = \rho^{y_c}.\sigma_{y_t}\sigma_{y_{t'}}$ where $\sigma^2_{y_\tau}$ is the variance of (log) income at time τ. Combining this result with the two equations in (2) gives

$$cov_y = \beta_t' Var(z)\beta_{t'} + \rho.\sigma_{yt}\sigma_{yt'} \qquad (3)$$

where $Var(z)$ is the variance-covariance matrix of the time-invariant characteristics, z, and cov_ε the covariance between the residual terms. With an approximation of cov_y, and estimates of β_t and $\beta_{t'}$, as well as of the variance of the residual terms, it is possible to obtain an approximation of the correlation coefficient between the residuals.

This appears a handy way of getting an estimate of the correlation coefficients between initial and terminal cross-section (log) income residuals by relying on their pseudo-panel dimension and cross-sectional variance. Yet, it will be seen below that this method tends to overestimate the correlation coefficient.[4]

2.3. Synthetic panels with AR(1) residuals

[4] Dang and Lanjouw (2013) also propose another way to estimate the correlation coefficient between the residuals that avoid these issues.

The methodology proposed in this paper assumes explicitly that the residual in the income model (2) for a given individual $i(t)$ follows a first order auto-regressive process, AR(1), between the initial and the final period. If the income of an individual were observed at the two time periods t and t', it would obey the following dynamics:

$$y_{i(t)t'} = z_{i(t)}\beta_{t'} + \varepsilon_{i(t)t'} \quad with \quad \varepsilon_{i(t)t'} = \rho\varepsilon_{i(t)t} + u_{i(t)t'} \qquad (4)$$

where the 'innovation term', $u_{i(t)t'}$, is assumed to be orthogonal to $\varepsilon_{i(t)t}$ and i.i.d. with zero mean and variance σ_u^2.

The autoregressive nature of the residual of the basic income model can be justified in different ways. The time-varying income determinants may be AR(1), the returns to the unobserved time invariant characteristics may themselves follow an autoregressive process of first order or, finally, stochastic income shocks may be characterized by this kind of linear decay. It is reasonably assumed that the auto-regressive coefficient, ρ, is positive.

Consider now the construction of the synthetic panel when the parameters of the AR(1) model in equation (4) are all known. The issue of how to estimate these parameters will be tackled in the next section. As described in the previous section, we regress income on the time-invariant attributes in the two periods following equation (2). We can use Equation (4) to obtain estimates of the residual of the income model, $\tilde{\varepsilon}_{i(t)t'}$, in time t' for observation $i(t)$:

$$\tilde{\varepsilon}_{i(t)t'} = \rho\hat{\varepsilon}_{i(t)t} + \tilde{u}_{i(t)t'}$$

In this equation, $\tilde{u}_{i(t)t'}$ has to be drawn randomly within the distribution of the innovation term, the cdf of which will be denoted $G_{t'}^{u}$. If estimates (or approximates) of ρ and the distribution $G_{t'}^{u}$ are available, the virtual income of individual *i(t)* in period t' can be simulated as:

$$\tilde{y}_{i(t)t'} = z_{i(t)}\hat{\beta}_{t'} + \rho\hat{\varepsilon}_{i(t)t} + G_{t'}^{u^{-1}}(p_{i(t)}) \qquad (5)$$

where $p_{i(t)}$ are independent draws within a (0,1) uniform distribution. After replacing $\hat{\varepsilon}_{i(t)t}$ by its expression in (2), this is equivalent to:

$$\tilde{y}_{i(t)t'} = \rho y_{i(t)t} + z_{i(t)}\left(\hat{\beta}_{t'} - \rho\hat{\beta}_t\right) + G_{t'}^{u^{-1}}\left(p_{i(t)}\right) \qquad (6)$$

Thus the virtual income in period *t'* of individual *i(t)* observed in period *t* depends on her observed income in period *t*, $y_{i(t)t}$, her observed time-invariant attributes, $z_{i(t)}$, and a random term drawn from the distribution $G_{t\prime}^{u}$. Because those virtual incomes are drawn randomly for each individual observed in period t, the income mobility measures derived from this exercise necessarily depends on the number of draws. That is, we will need to implement a large number of draws to accurately estimate the expected value of these measures and, most importantly, their distributions.

The two unknowns, ρ and $G_{t'}^{U}(.)$ must be approximated or 'calibrated' in such a way that the distribution of the virtual period t' income, $\tilde{y}_{i(t)t'}$, coincides with the distribution of $y_{i(t')t'}$ observed in the period *t'* cross-section. We discuss next the estimation of the auto-regressive coefficient, ρ, through pseudo-panel techniques before discussing the calibration of $G_{t'}^{U}(.)$.

2.3.1 Estimating the autocorrelation coefficients

The estimation of pseudo-panel models using repeated cross-sections has been analysed in detail since the pioneering papers by Deaton (1985) and Browning *et al.* (1985), and in particular Moffit (1993), McKenzie (2004) and Verbeek (2007). We closely follow the methodology proposed by the latter studies when estimating dynamic linear models on repeated cross-sections. Note, however, that in comparison with this literature, a new feature of the present methodology (as also proposed in Dang *et al.* (2014) and Dang and Lanjouw (2013)) is to construct the synthetic panels on only two rounds, rather multiple rounds, of cross-sections.

With repeated cross-sections, the estimation of an AR(1) process at the individual level can be done by aggregating individual observations into *groups* defined by some common time invariant characteristic such as year of birth, region of birth, school achievement, and gender. In defining these groups, it is important that the AR(1) coefficient as well as the variance of the innovation term, σ_u^2, should reasonably be assumed to be identical among them.

If G groups g have been defined overall, one could think of estimating the auto-regressive correlation coefficient ρ by running OLS on the group means of residuals:

$$\bar{\hat{\varepsilon}}_{gt'} = \rho\bar{\hat{\varepsilon}}_{gt} + \eta_{gt'} \qquad (7)$$

where $\bar{\hat{\varepsilon}}_{g\tau}$ is the mean OLS residual of (log) income for individuals belonging to group g at time τ, and $\eta_{gt'}$ is an error term orthogonal to $\bar{\hat{\varepsilon}}_{gt}$ with variance σ_u^2/n_{gt} where n_{gt} is the number of observations in group g. The estimating equation (7) raises a major difficulty since the group means of residuals of OLS regressions are asymptotically equal to zero at both dates t and t' so that equation (7) is essentially indeterminate.

There are two solutions to this indeterminacy. The first one is to work with second rather than first moments. Taking variances on both sides of the AR(1) equation:

$$\varepsilon_{i(t)t'} = \rho\varepsilon_{i(t)t} + u_{i(t)t'}$$

for each group g leads to:

$$\sigma_{\varepsilon gt'}^2 = \rho^2.\sigma_{\varepsilon gt}^2 + \sigma_{ugt'}^2$$

where $\sigma_{\varepsilon g\tau}^2$ is the variance of the OLS residuals within group g in the cross-section τ and $\sigma_{ugt'}^2$ is the unknown variance of the innovation term in group g. As discussed above, the expected value of that variance within group g is σ_u^2/n_{gt}. ρ can thus be estimated through non-linear GLS across groups g according to:

$$\sigma_{\varepsilon gt'}^2 = \rho^2.\sigma_{\varepsilon gt}^2 + \sigma_u^2/n_{gt} + \omega_{ut'} \qquad (8)$$

where $\omega_{ut'}$ stands for the deviation between the group variance of the innovation term and its expected value. Thus, it can be assumed to be zero mean, identically and independently distributed, and with a variance inversely proportional to n_{gt}.

The second approach to the estimation of ρ is to estimate the full dynamic equation in (log income) given by (3) across groups g. Using the same steps as those that led to (5), this equation can be written as:

$$\bar{y}_{gt'} = \rho\bar{y}_{gt} + \bar{z}_{gt}\gamma + \bar{u}_{gt'} \qquad (9)$$

where it has been reasonably assumed that $\bar{z}_{gt}$ and $\bar{z}_{gt'}$ are identical, which is only asymptotically correct,[5] so that the coefficient γ actually stands for $\beta_{t'} - \rho\beta_t$. In any case, ρ can be consistently estimated through GLS applied to (8), keeping in mind that the residual term $\bar{u}_{gt'}$ is heteroskedastic with variance σ_u^2/n_{gt}.

Note that this approach departs from Dang and Lanjouw (2013)'s first method to approximate ρ. As discussed above, they derive the covariance of residuals from the covariance of (log) incomes through (3). The latter is estimated through OLS applied to

$$\bar{y}_{gt'} = \delta\bar{y}_{gt} + \theta_{gt'} \qquad (10)$$

and $cov_y = \hat{\delta}\sigma_{yt}\sigma_{yt'}$. As can be seen from (9), however, a term in $\bar{z}_{gt}$ is missing on the RHS of (10), which means that the residual term $\theta_{gt'}$ is not independent of the regressor $\bar{y}_{gt}$. If the missing variable is positively correlated with the regressor, it could be the case that $\hat{\delta}$ is biased upward, the same being true of the covariance of (log) incomes (see Basu (2020) for other conditions).

The two approaches proposed above can be combined by estimating (8) and (9) simultaneously to get an unbiased estimate of the auto-regressive coefficient ρ.[6] As this is essentially adding information, moving from G to 2G observations, this joint estimation should yield more robust estimators.

Note finally, that it is possible to obtain additional degrees of freedom in the construction of the synthetic panel by assuming that the auto-regressive coefficient differs across several g-

[5] This requires the additional assumption of no difference in the sampling procedure between the (two) round of repeated cross-sections.

[6] A similar approach is followed by Kraay and van der Weide (2017).

groupings. For instance, there may be good reasons to expect that ρ declines with age. But this would require that individuals are described by enough time-invariant attributes and that there are enough observations in the whole sample so that a large number of 'groups' with a minimum number of observations can be defined.[7]

2.3.2. Calibrating the distribution of the innovation terms

In theory, once an estimate of the autoregressive coefficient ρ is available, the distribution $G_{t'}^{U}(\,)$ of the innovation terms, $u_{i(t)t'}$, can be recovered from the data.

The AR(1) specification implies:

$$\tilde{\varepsilon}_{i(t)t'} = \hat{\rho}\hat{\varepsilon}_{i(t)t} + \tilde{u}_{i(t)t}$$

where $\hat{\rho}$ is the pseudo-panel correlation coefficient obtained from (8) and (9), $\tilde{\varepsilon}_{i(t)t'}$ are the virtual residuals, and $\tilde{u}_{i(t)t'}$ are the randomly generated innovation terms. The challenge is to find the distribution $G_{t'}^{U}(\,)$ of the innovation terms such that the distribution of the virtual residuals $\tilde{\varepsilon}_{i(t)t'}$ be the same as the distribution of the observed OLS residuals $\hat{\varepsilon}_{i(t')t'}$ obtained from the income regression (1). Assuming $G_{t'}^{U}(\,)$ is a continuous function, it must satisfy the following functional equation:

$$F_{t'}(X) = \int_{-\infty}^{+\infty} F_t[(X-u)/\hat{\rho}] \cdot g_{t'}^{u}(u)du \qquad (11)$$

where $F_\tau(\,)$ is the cdf of the observed residuals $\hat{\varepsilon}_{i(\tau)\tau}$ and $g_{t'}^{u}$ the density of the innovation term. Hence, knowing the distribution of the residuals in the two periods and the autocorrelation coefficient, it is theoretically possible to recover the distribution of the innovation terms that make the distribution of the synthetic panels identical to the observed distributions at the two points of time.

The functional equation (11) is not simple. Known as the Fredholm equation, it can be solved through numerical algorithms, which are rather intricate. A simpler parametric method was

[7] This would pose fewer concerns with cross-sectional survey data samples, which are typically much larger than the panel data sample that we use in this paper to test the synthetic panel construction procedure.

chosen instead, based on the approximation that the distribution $G_{t'}^{u}$ is a mixture of two normal variables with parameters $\theta = \{p, \mu_1, \sigma_1, \mu_2, \sigma_2\}$ - p being the probability that the distribution is $N(\mu_1, \sigma_1)$ and $(1-p)$ the distribution $N(\mu_2, \sigma_2)$. These parameters may be calibrated by minimizing the square of the difference between the two sides of (11). Although it is only an approximation, this method turned out to give rather satisfactory results. In addition, it is also less restrictive than the normality assumption made in Dang *et al.* (2014).[8] The details of the calibration of the distribution $G_{t'}^{u}$ with a mixture of two normal distributions are provided in Appendix 1.

2.4 Practical summary

Practically, the whole procedure leading to the construction of a synthetic income panel under the assumption that the income residuals follow an AR (1) process and with the constraint that the initial and terminal distribution of income match the corresponding cross-sections may be summarized as follows.

1. *Income model*
    a. Define a set of time-invariant attributes, *z*, to be used in the (log) income model.
    b. For each period, run OLS on (log) income with *z* as regressors and store both vectors of residuals, $\hat{\varepsilon}_{i(t)t}$ and $\hat{\varepsilon}_{i(t')t'}$, and the returns to time invariant attributes, $\hat{\beta}_t$ and $\hat{\beta}_{t'}$.
2. *Autoregressive parameter*
    a. Define a number of groups g based on time invariant attributes with enough observations for group means to be precise enough.
    b. Average the (log) income and the time invariant characteristics for each group and compute the variance of the OLS residuals of the models estimated in 1.a).

[8] A mixture of normal variables is also used in the parametric representation of the dynamics of income proposed by Guvenen et al. (2015).

c. Estimate the residual auto-correlation coefficient $\hat{\rho}$ through the joint pseudo-panel equations (8) and (9)

3. *Distribution of the innovation terms*. Calibrate the set of parameters, θ, of the distribution of the innovation term, which are assumed to be a mixture of two normal variables, as described in Appendix 1.
4. *Synthetic panel*. For each observation in the initial cross-section, *t*, draw randomly a value in the preceding distribution and compute the virtual income in period *t'* using equation (6). Evaluate income mobility matrices and mobility measures based on that drawing.
5. *Simulation*. Repeat 4 to obtain the expected value and distribution of the mobility matrices and measures.

3. Construction and validation of the synthetic income panel

The procedure detailed above is now applied to construct synthetic incomes in 2005 for households surveyed in 2002. We analyse two survey rounds from a panel survey implemented in Mexico pretending that these rounds are cross sections. The transition matrix between the initial and terminal years observed in the panel will be replaced by the procedure described in the preceding section. The genuine matrix in the original panel data will be used essentially for evaluating its precision. While our procedure can be conducted either at the household level or the individual level, we focus on households as observational units, since these tend to offer a wider perspective on wellbeing.

3.1 Data and Income definition

We use the *Mexican Family Life Survey* (hereafter referred to as MxFLS), which is representative at the national, regional, and urban-rural levels. This longitudinal survey gathers information on socioeconomic indicators, migration, demographics, and health indicators for the Mexican population. It is expected to track the Mexican population throughout a period of at least ten years.

The first and second waves, conducted in 2002 and 2005 respectively, rely on a baseline sample size of 8,400 households and collect data on the socio-demographic characteristics of each household member, individual occupation and earnings, household income and expenditures, and assets ownership. The sample in 2005 was expanded to compensate for attrition, which amounted to 10% of the original sample. We use the common sample between the two rounds that did not attrite, that is, the set of households observed both in 2002 and 2005. Due to confidentiality, information on the sample design (sampling units) is not publicly available.[9]

Household income data follow the official definition for computing income poverty in Mexico. They include both monetary and non-monetary resources. The former comprise receipts from employment, own businesses, rents from assets, and public and private transfers. Non-monetary income includes in-kind gifts received and the value of services provided within the household, such as the rental value of owner occupied dwelling or self-consumption.[10] Total income is divided by the household size in order to obtain per capita income and is deflated by the Consumer Price Index (anchored to the prices in August 2005) to make 2002 and 2005 data comparable.

To ensure stable household formation, as with traditional pseudo-panel methods, we restrict the sample to households with heads age 25 to 62 years old in 2002, hence 28 to 65 years old in 2005, and with non-missing income in both years. In addition, to overcome biased estimates due to outlier observations, four percent of the observations were discarded (two percent each at the bottom and at the top of the income distribution).

3.2 Time-invariant attributes and the income models

Time-invariant attributes could be determined based on several different criteria. In particular, individual and deterministic attributes like the year of birth, sex, educational achievement, and ethnicity are the most natural candidates. Depending on the issue of interest, the time horizon and the specific country under study, other household

[9] See Rubalcava and Teruel (2006).

[10] This definition changed to introduce a multidimensional poverty approach in 2008.

characteristics can be used such as household size (after taking into account the probability of a new-born during the considered time interval) or the area of residence (after considering migration). [11]

More time-invariant attributes can help improve the goodness-of-fit of the income model, which can result in better synthetic panel approximation. However, a longer time interval between the cross-sections could negatively affect time invariability due to changes with the sampled populations (e.g., the educational level can improve over a long period of time for the whole population). While variables that are not strictly time-invariant should be discarded (e.g., current employment status and occupation), they should be considered in the particular case of the country under analysis. Some other variables could be considered time-invariant under reasonable circumstances such as marital status.

We use two model specifications, each with different degrees of time invariability, to assess the sensitivity of the selected variables. The first specification (Model 1) uses the household head's characteristics such as gender, formal years of schooling, birth year, and the household composition by age groups. This includes a dummy variable to account for the presence of a less than 3-year old child in the terminal year to account for the probability of a new-born in this three-year period. It also includes other variables including the area of residence (urban/rural), marital status, and regions (northeast, west, centre, northwest, and south-southeast). An alternative specification (Model 2) includes long-lasting productive assets such as real estate and farming assets (land for agricultural production and cattle), dwelling ownership, and the possession of other dwellings other than the one in use. **Table A1** in Appendix 2 show the descriptive statistics and OLS estimates for the income model in equation (1).

It is useful to briefly note some restrictions on the variables we select for the income model. The survey collect data on ethnicity, religious conviction, and household head literacy. It also contains data on retrospective data including the size of the birth city, the year of marriage, the education of household head's parents, place of birth, and migration records. Those

[11] A potential issue with using the residence area as a time-invariant variable is migration. But this should pose no concern during a short period. According to census data, the internal migration rate in Mexico was around 2% in the period 2000-2005 (Chávez-Juárez and Wanner, 2012).

attributes are not included in these income model specifications due to high prevalence of missing data or extremely low sampled observations. We do not observe statistical differences on most of the variables across these two years for the selected variables, except for the dummy variables indicating the presence of children below three years old and long-lasting assets (farming assets and dwellings property). Consequently, Model 1 is our preferred model specification.

Although the proposed method does not assume normality for the residuals, neither for the initial nor for the final year, we tested this assumption in our income models. For illustrative purposes, **Graph 1** shows the kernel distribution of (log) income residuals in both years, and compares it with the normal distribution. These graphs and the Skewness and Kurtosis tests, along with the Shapiro-Wilk normality test, confirm that the normality assumption with the distribution of residuals is strongly rejected. [12]

[Graph 1. Income models' residuals: kernel density by year and model]

3.3 The autocorrelation coefficient and calibration parameters

Estimating the autocorrelation coefficient is a central step in the construction of synthetic panels. Firstly, household observations were grouped by some common characteristics to create a pseudo panel. In our case, thirty-two clusters were obtained by the interaction of eight birth-year cohorts, of 5 years interval each, and four groups of education: incomplete primary education, complete primary but incomplete secondary education, complete secondary education but incomplete high school, and complete high school or more.[13] For instance, one such group comprise households whose heads were born between 1974 and 1978 with incomplete primary education.

[12] The Skewness and Kurtosis tests rejects the null hypothesis of normality in 2005 and 2002 respectively. The Shapiro-Wilk W test also rejects the hypothesis that both residuals are normally distributed.

[13] Other studies working with pseudo panel methods use age interactions with other characteristics like manual or non-manual occupations as in Browning *et al.* (1985), regions as in Propper *et al.* (2001), sex (see Cuesta *et al.* (2007)), or education levels as in Blundell *et al.* (1998). Proper, Rees, and Green (2001) use cells of around 80 observations whereas Alessie, Devereux, and Weber (1997) use cells of more than 1,000 observations. Antman and Mackenzie (2007b) and Antman and Mackenzie (2007) used 100 observations as a reference. In our case the vast majority of the groups possess no less than one hundred observations.

We subsequently estimate equations (8) and (9) with the resulting pseudo panel. The AR (1) coefficient ρ is estimated at 0.26 using the actual panel data (Model 1), which serves as the benchmark. Regardless of the equation being used, the estimates in **Table 1** have the expected signs and do not significantly deviate from this value. However, the combined use of these two approaches, through a non-linear equation system, delivers a more accurate point estimate of 0.25. Unsurprisingly, its confidence interval is substantially broader than, but also fully consistent with, that of the estimate based on the actual panel. It will be seen later how this lack of precision of the estimated auto-regressive coefficient, ρ, leads to a lack of precision of synthetic income mobility estimates.[14]

[Table 1. Pseudo panel AR(1) estimates by method, 2002-2005

The estimate of ρ and its corresponding 95% confidence intervals now enables us to calibrate the parameters that characterize the distribution of innovation terms. This is implemented using two different regimes. Regime 1 uses the point estimate of $\hat{\rho}$, in Table 1, to obtain a unique set of parameters $\theta(\hat{\rho}) = \{p, \mu_1, \sigma_1, \mu_2, \sigma_2\}$ of the distribution of the innovation term $G_{t'}^{u}(.)$, employing the procedure described in Appendix 1. The calibration parameters for model 1 are $\theta(\hat{\rho} = 0.25) = \{p = 0.33, \mu_1 = 0.007, \sigma_1 = 1.5, \mu_2 = -0.003, \sigma_2 = 0.94\}$.[15] A value of this innovation term is then drawn from that distribution for every observation in the initial year to obtain a synthetic panel. However, because the random drawing introduces noise, the procedure is repeated 500 times. We report the mean values of each mobility measures with their 95% confidence intervals.

In regime 2, the imprecision of the estimate of $\hat{\rho}$ is fully accounted for by repeating the preceding exercise over a sample of ρ values spanning its most likely range of variation. First, we randomly draw 100 correlation coefficients from a normal distribution within its 95% confidence interval. These intervals are obtained from the estimates using the system of equations (8, 9) in Table 1. We then use these coefficients as in regime 1, except that we

[14] If the estimation of the correlation coefficient through pseudo-panel techniques is not very accurate, it must be kept in mind that the coefficient estimated on the genuine panel data is certainly not as precise as it appears in table 1. In fact, measurement errors are known to result in biased estimates based on actual panel data. Measurement errors are less of a problem in the pseudo-panel approach since they are averaged out when considering groups of households. The price to pay with the pseudo-panel approach, however, is less precision.

[15] Note that $p\mu_1 + (1 - p)\,\mu_2$ is practically zero, as could be expected since the mean residual is zero.

repeat the procedure 50 times for each one of the correlation coefficients (rather than the 500 repetitions using a single ρ value). The mean value of mobility measures with regime 2 are therefore obtained from 5,000 repetitions (50 times100), although with different values of ρ. **Graph 2** shows a graphic description of the resulting parameters for each model.

[Graph 2. Distribution of the calibration parameters conditional on rho]

The mean of mobility indicators is not expected to be very different between these regimes but we do expect some difference in their precision, particularly in regime 2 that reflects the imprecision in the estimates of ρ. This is to be observed in the 5%-95% confidence intervals reported in the tables below. Note that the estimates of mobility indicators derived from the actual panel are themselves subject to sampling errors, which we address by bootstrapping when computing their confidence intervals.

4. Synthetic panel results

We now examine income mobility based on the synthetic panel and compare it to the genuine panel. Note that the Mexican economy witnessed some growth between 2002 and 2005, which could affect some of the mobility measures that we analyse.[16] We first compare the shape of the (log) synthetic distribution, for each model and regime, with the genuine (log) income distribution in 2005. **Graph 3** shows the kernel densities of both the genuine and virtual income distributions. The synthetic income distribution is derived from averaging all the repetitions implemented in the calibration procedure. The graph provides a first visual assessment of the fit of the synthetic estimates and shows that all the model specifications reasonably reproduce the shape of the actual income distribution, except for a small discrepancy in the bottom of the distribution. The mixture of normal variables used to approximate this distribution necessarily has smooth tails and cannot account for such irregularity in the actual distribution.

[Graph 3. Genuine and synthetic income density by regime and model, 2005]

[16] The real GDP per capita grew by 0.21%, 2.60%, and 0.92% in 2003, 2004 and 2005 with respect to the previous year according to the World Bank’s World Development Indicators.

We then examine the transition matrix associated with the synthetic panel using model 1 with both regimes. To increase the sensibility of income changes over the terminal income distribution, the transition matrix is defined in absolute terms using real-income thresholds. These thresholds are defined using the income quintile limits observed in the baseline (2002) and remain fixed in the terminal year (2005). The marginal distribution in the baseline shows 20% of the population in each income bracket by construction.

We plot three transition matrixes in a single table to facilitate the comparison of both regimes with the actual panel. The upper and lower parts of **Table 2** correspond to regime 1 and 2 respectively while the middle part shows the genuine matrix with bootstrapped confidence intervals (see **Table A3** in Appendix 2 for model 2). The synthetic transition probabilities for regime 1 appear close to the genuine ones in the sense that their confidence intervals most often contain the observed probabilities, 15 cases out of 25 indicated with an asterisk, and do substantially overlap - 24 cases out of 25. As expected, working with a wider set of rho values from regime 2 tend to deliver slightly larger, although not always, confidence intervals. Note that both the synthetic and the actual panel reflect the same pattern of pro-poor growth (i.e. a smaller share of population in the bottom bracket of 2005 as compared with that of 2002).

[Table 2. Transition matrix by 2002 income *quintiles* with authentic and synthetic panel, MXFLS 2002-2005]

We also use the Mann-Whitney test to evaluate the goodness of fit between the synthetic and the genuine 2005 income distributions conditional on the ventiles of the 2002 income distribution. In fact, this test is equivalent to comparing the confidence intervals in the synthetic and genuine transition matrix in **Table 2** for each cell but using ventiles, rather than quintiles, for the 2002 income to increase the sensitivity of this test.[17]

Graph 4 summarizes these results for model 1 and regime 1. The graph displays the share of the 500 draws that pass the test for $\rho = 0.25$ and shows how satisfactorily the synthetic panel reproduces the dynamics in the genuine panel. It can be seen that the fit is satisfactory in

[17] This test utilizes information regarding the rank order and constitutes an alternative for the two-sample t-test of independent samples.

practically all the ventiles of the baseline income distribution, except for the poorest and the richest ventiles. On average, more than 90% of the samples passed this test on the rest of the baseline distribution. This comparison is extended for the two theoretical upper and lower bound values for ρ, 0 and 1 (as in Dang *et al.* (2014)). The results are much poorer with these extreme values. Note that assuming a perfect correlation of residuals in both the initial and terminal years delivers the poorest performance in this setting.

[Graph 4. Mann-Whitney test. Shares of samples that pass the test]

These results highlight the importance of identifying the appropriate estimates of ρ in the construction of synthetic estimates. This is not surprising. At the same time, however, it is worth mentioning that the difference of fit using the end values of the confidence interval in the pseudo-panel estimate (i.e. $\rho = 0.15$ and $\rho = 0.45$, see **Table A4** in Appendix 2) is not large, except for the extreme deciles of the initial distribution.

Since poverty dynamics has been the main focus for empirical applications of synthetic panels, we compute two sets of poverty transitions, based on the upper limits of the first two income quintiles in direct reference to the 'shared prosperity' goal adopted by the World Bank. As discussed above, we employ alternative values of ρ, using model 1 with both regimes. **Table 3** shows that the estimated persistent poverty rate (i.e. being poor in both periods), using the first poverty line, is 6.7% and 6.8% respectively for regime 1 and regime 2 with the central value of ρ. Both these figures are very close to the persistent poverty rate of 6.3% based on the actual panel and fall inside the 95% confidence intervals (indicated by an asterisk). A similar result holds for the estimates using the second poverty line. Larger differences are found for downward mobility (i.e., from non-poor to poor) with the first poverty line only. On the other hand, the table shows that poverty mobility estimates based on synthetic panels are sandwiched between the two extreme theoretical values of ρ (i.e., 0 and 1). Note that the discrepancy between the synthetic and genuine estimates with these values increase with more distant values of ρ . This finding further illustrates the sensitivity of poverty mobility estimates based on synthetic panels.

[Table 3. Poverty dynamics, 2002-2005]

We show next a simple measure of absolute income mobility -the fractions of households with higher and lower income in the terminal year as an additional validation check. **Graph 5** shows the shares of households with positive and negative income growth for model 1 with regime 1, which add up to 100%. Both the actual and synthetic panels show a clear pattern of progressive growth incidence where the poorest groups display the largest growth gains while the richest groups assembled the largest losses. The differences in these absolute mobility measures are essentially not significant (i.e., the 95% confidence intervals for the synthetic and actual panels overlap substantially). For instance, both the synthetic and the genuine figures show that around 90% of households in the poorest quintile experience a positive growth rate, while the remaining 10% experience a negative income growth.

[Graph 5. Absolute Mobility]

We supplement these results with Non-Anonymous Growth Incidence Curves (NAGIC). These curves plot individual income growth rates over the rank of the initial distribution.[18] **Graph 6** employs deciles of the genuine and synthetic income with their corresponding 95% confidence intervals using model 1 and regime 1. These downward sloping NAGIC charts are remarkably similar in terms of their level and shape, confirming a pattern of progressive growth. Again, the differences with these estimates are not statistically significant given that in most cases the genuine estimates fall within the synthetic 95% confidence intervals. There is also an ample overlap between the confidence intervals along the whole income distributions (produced by bootstrapping for the genuine estimates).

[Graph 6. Non-Anonymous Growth Incidence Curves]

[18] Alternatively, the Anonymous Growth Incidence Curve (AGIC) shows the change in average income per current decile, rather than per decile of initial income. The difference between AGIC and NAGIC is precisely that the latter account for mobility – see Bourguignon (2011), and Bergman and Bourguignon (2019). The AGIC are not shown here because by construction of the synthetic panels cross-sectional distributions are identical for both the initial and the final period- up to the approximation to meet that constraint.

5. Concluding remarks

This paper proposed methodological improvements in the construction of synthetic income panels based on repeated cross-sections. These innovations include explicitly assuming that the unobserved or time variant determinants of (log) income follows an auto-regressive process of first order. The proposed approach then relies on pseudo-panel procedures to estimate the corresponding auto-regressive coefficient. Furthermore, the use of calibration techniques allows abstracting from oft-used (log) normality assumptions, generating a close to perfect match to the terminal year income distribution. These improvements allow considering the whole income mobility matrix rather than mobility in and out of poverty which has been the main focus for empirical applications of synthetic panels.

We perform an empirical validation by using two consecutive cross-sections of income based on a genuine panel survey in Mexico. Income mobility indicators are reasonably similar for the synthetic and genuine panels and are most often not statistically different. Yet, the validation also showed the sensitivity of particular indicators to the value of the auto-regressive coefficient used in modelling the effects of the unobserved determinants of (log) income. This exercise points to the important role that the value of the auto-regressive coefficient plays in the accuracy of synthetic panels and calls for caution in analysing income mobility with synthetic panels.

An original pseudo-panel method developed in this paper yield an estimate of that coefficient which is theoretically unbiased but comes out with a rather broad interval of confidence. A Monte-Carlo approach where a large number of synthetic panels are generated, each one based on a value of the auto-regressive coefficient drawn from the distribution of its pseudo-panel estimators, seems the proper way of dealing with that estimation imprecision. The resulting confidence intervals of income mobility measures prove sizable, even though they very much overlap with those found for the genuine panel.

This conclusion is important for a possibly systematic use of synthetic panels. If successive synthetic panels were to be used to examine how income mobility, for instance mobility in and out of poverty changes over time or differs across countries, then the synthetic panel approach could only reveal sizable changes or differences. Yet, it bears to say that more experience than with a single country and a single period is needed to get a better knowledge of the degree of precision that it is possible to reach with this synthetic approach to income mobility.

References

Arellano, M., and Bond, S. 1991. “Some tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations”. Review of Economic Studies, 58, 277-297.

Antman, F., and McKenzie, D. 2007. “Poverty traps and nonlinear Income Dynamics with Measurement Error and Individual Heterogeneity” Journal of Development Studies, Vol. 43, No. 6 (August 2007). Routledge.

Antman, F., and McKenzie. 2007b. “Earnings Mobility and Measurement Error: A Pseudo Panel Approach” Economic Development and Cultural Change, Vol. 56, No. 1 (October 2007). The University of Chicago Press.

Basu, D. 2020. "Bias of OLS Estimators due to Exclusion of Relevant Variables and Inclusion of Irrelevant Variables," Oxford Bulletin of Economics and Statistics, 82(1), 209-234.

Beegle K., Christiaensen, L., Dabalen, A. and Gaddis, I. 2016. “*Poverty in a Rising Africa”*. Washington, DC: The World Bank.

Berman, Y., and Bourguignon, F. 2019. “Anonymous and Non-Anonymous Growth Incidence Curves: Evidence from the United States over the Last 50 Years.” Mimeo.

Bourguignon F. 2011. “Non-anonymous growth incidence curves, income mobility and social welfare dominance”. Journal of Economic Inequality, 9(4), 605-627.

Bourguignon F., Goh, C., and Kim, D. 2004. “Estimating individual vulnerability to poverty with pseudo panel-data”. In Morgan, Grusky and Fields (eds), Mobility and Inequality; Frontiers of Research in Sociology and Economics, Stanford University Press.

Browning, M., Deaton, A., & Irish, M. 1985. A Profitable Approach to Labor Supply and Commodity Demands over the Life-Cycle. Econometrica, 53(3). May 1985.

Chávez-Juárez, F., and Wanner, P. 2012. “Determinants of Internal Migration in Mexico at an Aggregated and a Disaggregated Level” (March 26, 2012). Available at SSRN: http://ssrn.com/abstract=1978806

Cruces, G., Lanjow P., Lucchetti, L., Perova, E., Vakis, R., and Viollaz, M. 2015. “Intra-generational Mobility and Repeated Cross-Sections: A three country validation exercise”. Journal of Economic Inequality, 13:161–179.

Dang, H., and Lanjouw, P. 2013. “Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections”, June. The World Bank. Policy Research Paper, 6504.

Dang, H., Jolliffe, D., and Carletto, C. 2019. "Data Gaps, Data Incomparability, and Data Imputation: A Review of Poverty Measurement Methods for Data-Scarce Environments". *Journal of Economic Surveys*, 33(3): 757-797.

Dang, H., Lanjouw, P., Luoto, J., and McKenzie, D. 2014. “Using Repeated Cross-Sections to Explore Movements in and out of Poverty”. *Journal of Development Economics*, 107: 112-128.

Deaton, A. 1985. “Panel Data from Times Series of Cross-Sections,” Journal of Econometrics, 30.

Ferreira, F., Messina, J., Rigolini J., López-Calva, L., Lugo, M., and Vakis, R. 2013. “Economic Mobility and the Rise of the Latin American Middle Class”. Washington, DC: World Bank.

Fields, G., 2012. “Does Income mobility equalize longer-term incomes? New measures of an old concept”. Journal of economic inequality 8(4), 409-427.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. 2015. “What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?”. NBER Working Paper No. 20913. January 2015

Jäntti, M. and Jenkins, S. 2015. “Income mobility”. In Bourguignon and Atkinson (2014). “Handbook of Income Distribution”, Volume 2A. Chapter 10. Elsevier.

Kraay, A., and Van Der Weide, R. 2017. “Approximating income distribution dynamics using aggregate data”. Policy Research working paper; no. WPS 8123. Washington, D.C. : World Bank Group.

McKenzie, D. 2004. “Asymptotic theory for heterogeneous dynamic pseudo-panels”. Journal of Econometrics, Volume 120, Issue 2, 2004, Pages 235-262.

Moffit, R. 1993. “Identification and Estimation of Dynamic Models with time series of Repeated Cross-sections”. Journal of Econometrics, 59, 99-123.

Moreno, H. 2018. “Long Run Economic Mobility”. Doctoral dissertation (Paris School of Economics). Université Panthéon-Sorbonne - Paris I, 2018. English.

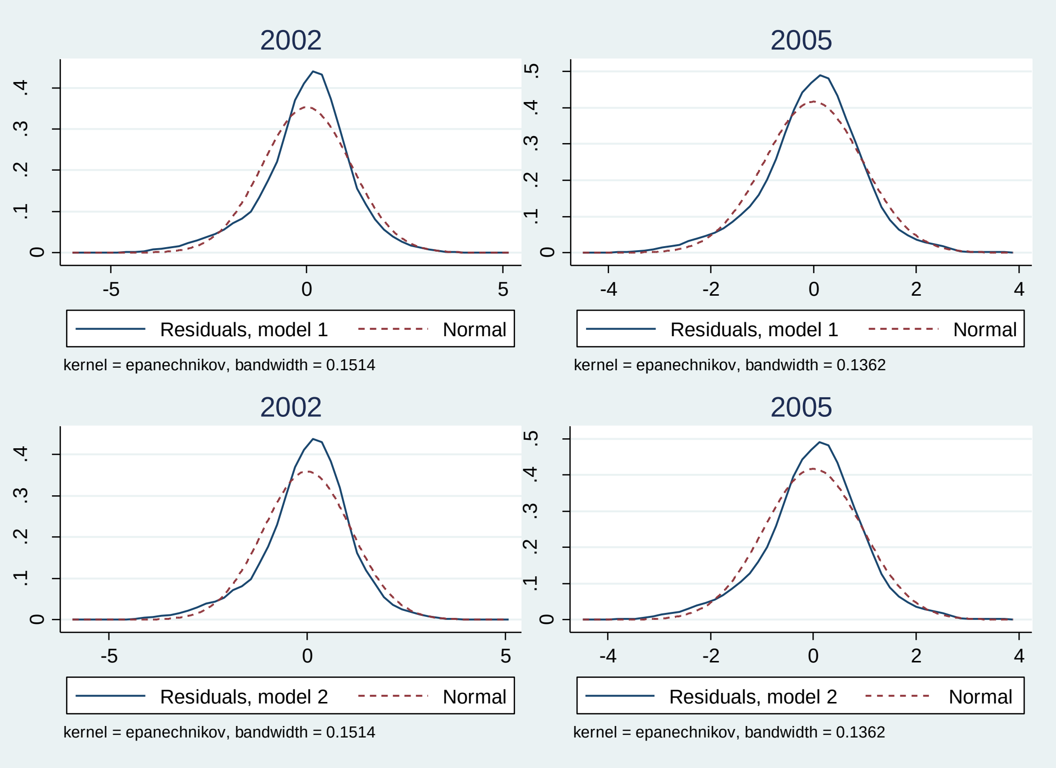
OECD. 2018. "A Broken Social Elevator? How to Promote Social Mobility". OECD Publishing. Paris.

Rubalcava, L., and Teruel, G. 2006. “Mexican Family Life Survey, First Wave”, Working Paper, www.ennvih-mxfls.org.

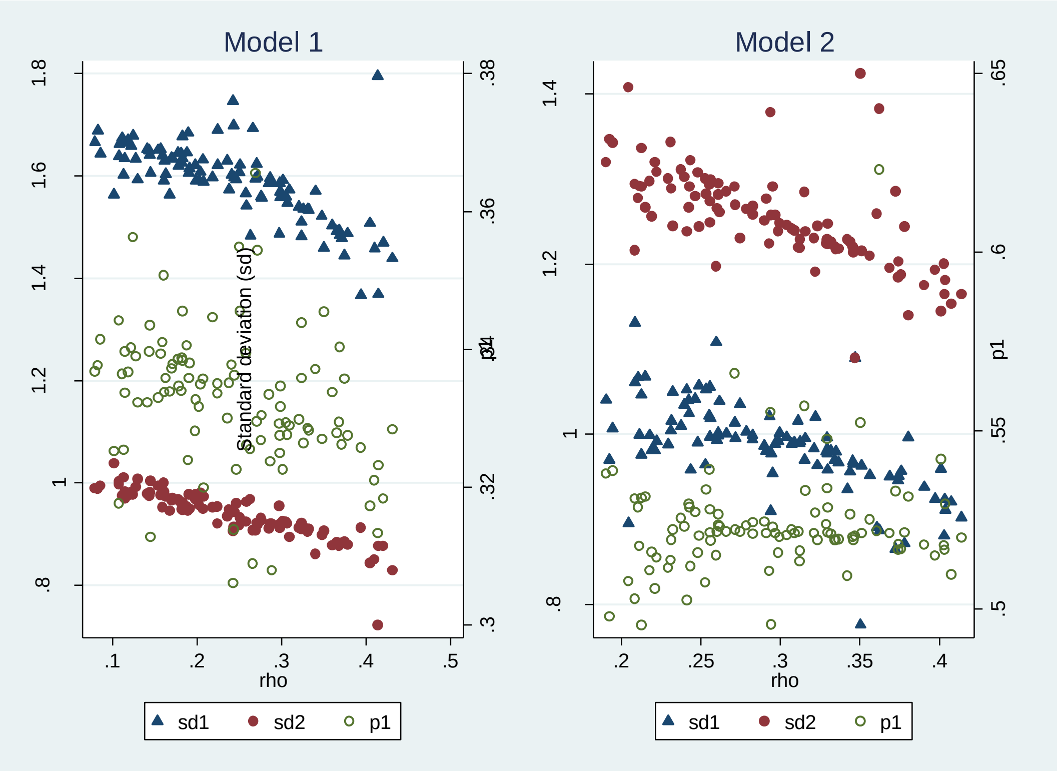
Rubalcava, L., and Teruel, G. 2008. “Mexican Family Life Survey, Second Wave”, Working Paper, www.ennvih-mxfls.org.

Verbeek, M. 2008. “Pseudo panels and repeated cross-sections”. Chapter 11 in Mátyás and Sevestre, eds., 2008, “The Econometrics of Panel Data”, Springer-Verlag Heidelberg.

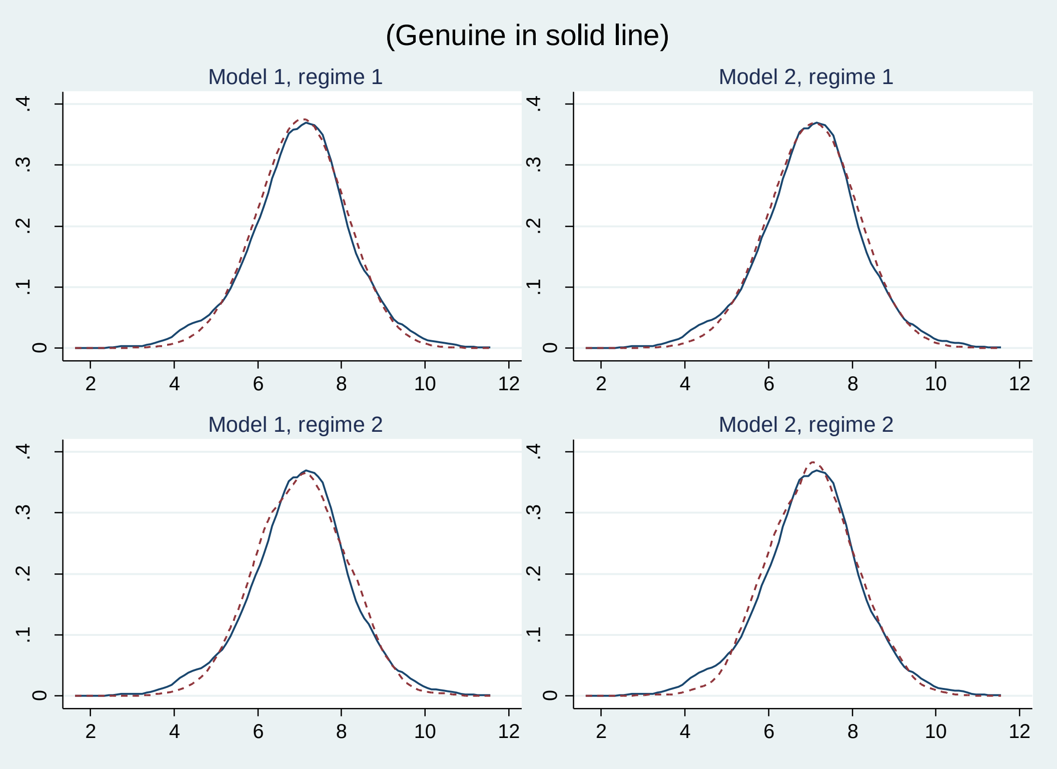
Graph 1. Income models' residuals: kernel density by year and model



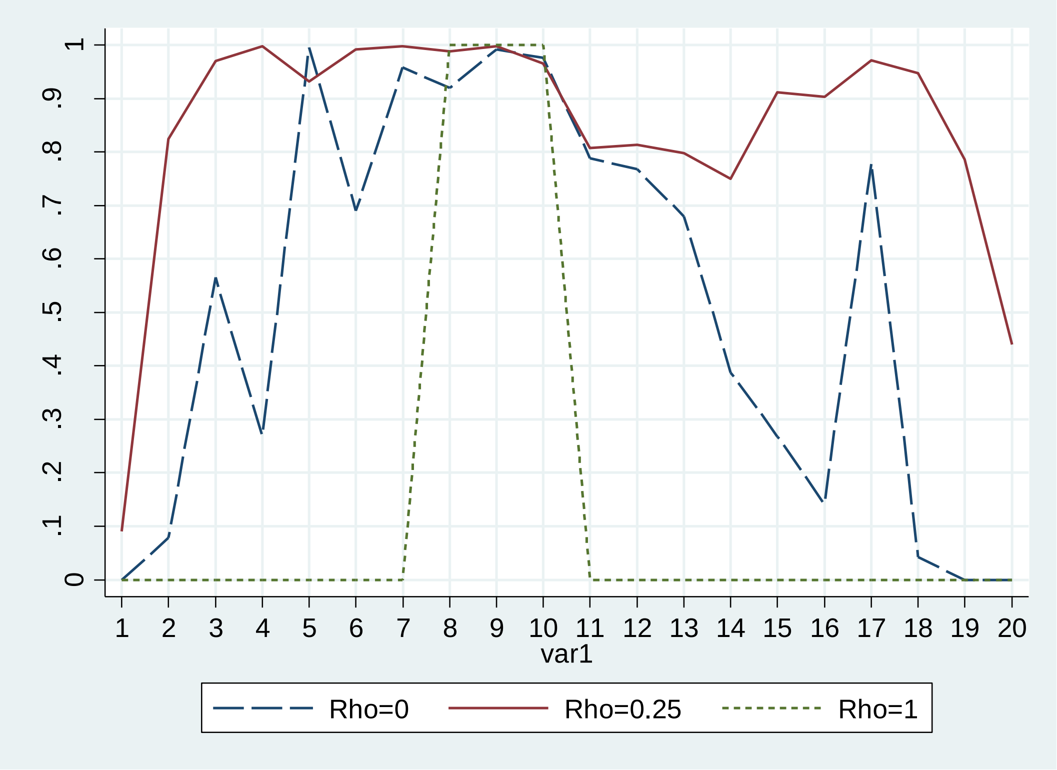
Graph 2. Distribution of the calibration parameters conditional on rho



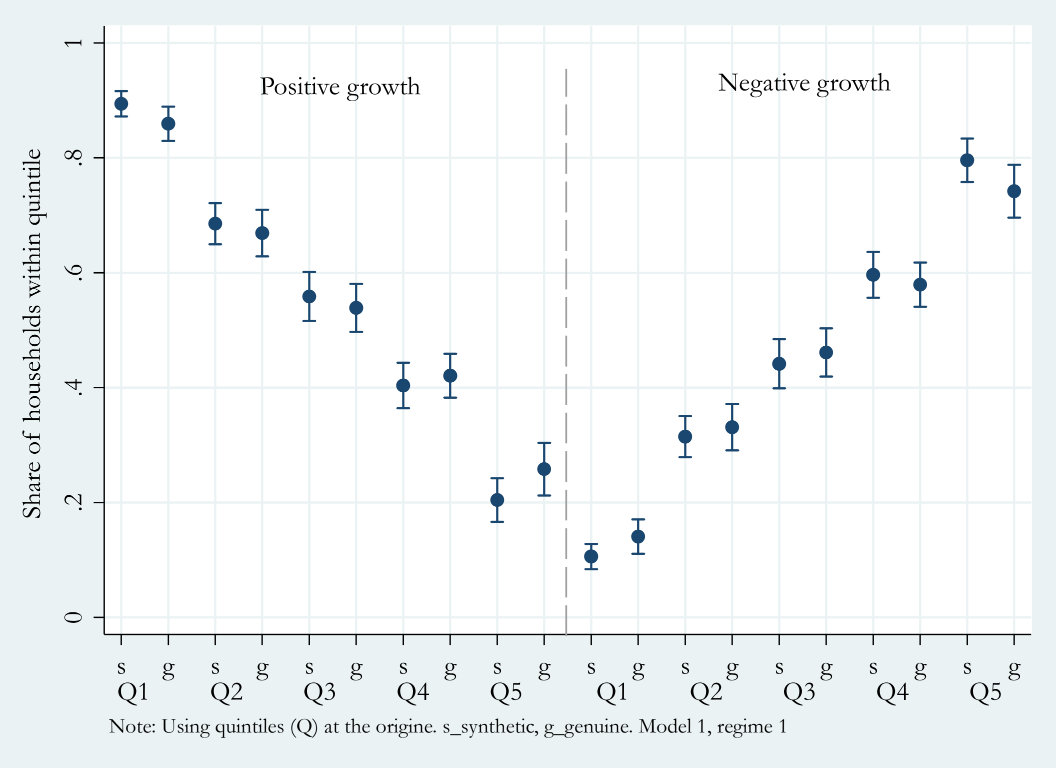
Graph 3. Genuine and mean synthetic income, 2005



Graph 4. Mann-Whitney test. Shares of samples (random drawings) that pass the test of identity of the synthetic and genuine panel final income distributions conditional on initial income ventile (Model 1, regime 1)



Graph 5. Absolute Mobility (Model 1, Regime 1)



Graph 6. Non-Anonymous Growth Incidence Curve (Model 1, Regime 1)

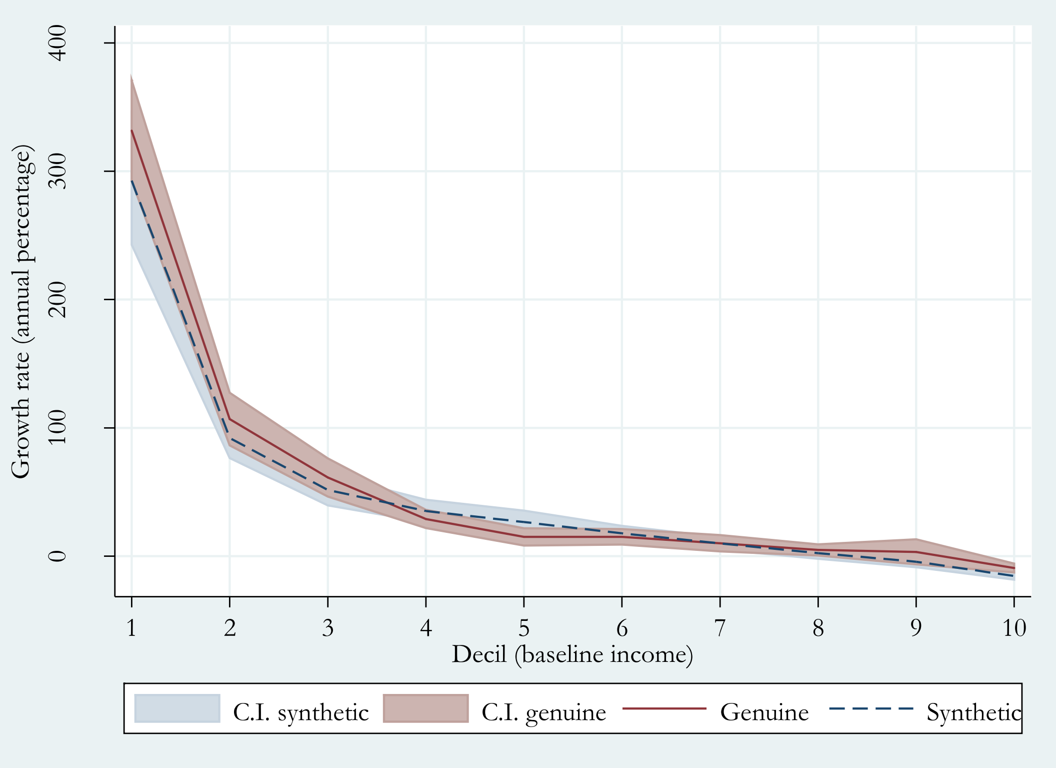


Table 1. Rho estimates by model and method

	Pseudo panel			Genuine panel
Models	**Equation 8**	**Equation 9**	**Eq. system (8, 9)**	**With microdata (residuals)**
	Non linear	**Linear**	**Non linear**	
	(1)	**(2)**	**(3)**	**(4)**
Model 1	0.292*	0.132	0.254**	0.257***
	(-0.05-0.64)	(-0.14-0.40)	(0.04-0.47)	(0.235 - 0.280)
Model 2	0.176	0.158	0.299***	0.226***
	(-0.82-1.17)	(-0.1-0.42)	(0.15-0.45)	(0.203 - 0.249)

Notes: *** $p<0.01$, ** $p<0.05$, * $p<0.1$. 95%. 95% Conf. Interval in parentheses. GLS estimates controlling for time invariant variables. Each estimate represents the coefficient from a different regression.

Table 2. Transition matrix (Model 1), 2002-2005

	Income Bracket	**1**	**2**	**3**	**4**	**5**	***Total***
Synthetic regime 1 2002 Quintiles (Origin)	1	6.7	6.4	3.7	2.4	0.8	*20*
		(6.2-7.3)*	(5.7-6.9)*	(3.2-4.3)*	(1.9-2.9)	(0.5-1.1)	
	2	3.3	6.0	4.9	4.1	1.7	*20*
		(2.9-3.8)*	(5.3-6.6)*	(4.4-5.6)*	(3.5-4.7)*	(1.4-2.1)*	
	3	1.7	4.7	5.0	5.4	3.1	*20*
		(1.3-2.2)	(4-5.3)*	(4.4-5.7)*	(4.8-6.2)*	(2.6-3.7)	
	4	0.9	3.3	4.6	6.2	5.0	*20*
		(0.6-1.2)	(2.9-3.8)	(4.1-5.2)	(5.5-6.8)	(4.4-5.5)*	
	5	0.3	1.6	3.1	5.9	9.0	*20*
		(0.1-0.5)*	(1.3-2.1)*	(2.6-3.7)*	(5.1-6.6)	(8.2-9.8)	
Marginal Dist.		13.0	22.0	21.4	24.1	19.5	*100*
		(11.9-14.1)	(20.5-23.5)	(19.8-23.1)*	(22.3-25.8)*	(18.1-20.9)*	
Genuine 2002 Quintiles (Origin)	1	6.3	6.0	3.4	3.2	1.1	*20*
		(5.5-7.1)	(5.2-6.8)	(2.7-4.2)	(2.5-3.8)	(0.8-1.4)	
	2	3.8	5.6	5.0	4.1	1.5	*20*
		(3.1-4.5)	(4.8-6.5)	(4.3-5.8)	(3.2-4.9)	(1.1-1.9)	
	3	2.6	4.1	5.6	5.7	2.0	*20*
		(1.9-3.3)	(3.4-4.8)	(4.6-6.6)	(4.8-6.5)	(1.5-2.5)	
	4	1.6	2.7	3.6	7.3	4.8	*20*
		(1.1-2.1)	(2.1-3.3)	(3-4.2)	(6.2-8.4)	(4-5.7)	
	5	0.5	1.9	2.6	4.8	10.0	*20*
		(0.3-0.7)	(1.2-2.7)	(2-3.2)	(3.9-5.8)	(8.6-11.4)	
Marginal Dist.		14.8	20.4	20.3	25.0	19.5	*100*
		(13.6-16)	(18.9-21.8)	(18.7-21.9)	(23.2-26.8)	(17.9-21.1)	
Synthetic regime 2 2002 Quintiles (Origin)	1	6.8	5.8	3.9	2.6	0.9	*20*
		(5.7-7.8)*	(5.3-6.5)*	(3.6-4.1)	(1.7-3.5)*	(0.4-1.4)*	
	2	3.7	6.4	4.4	3.8	1.7	*20*
		(3.5-3.9)*	(5.9-6.9)	(3.9-4.9)	(3.4-4.3)*	(1.4-2)*	
	3	1.6	5.0	5.0	5.2	3.3	*20*
		(1.2-2)	(4.6-5.3)	(4.7-5.6)*	(4.8-5.4)	(2.8-3.7)	
	4	1.0	3.4	4.0	5.9	5.7	*20*
		(0.7-1.3)	(2.9-3.8)	(3.7-4.2)	(5.7-6.4)	(5.6-6)	
	5	0.3	1.8	3.1	5.6	9.0	*20*
		(0.1-0.7)*	(1.3-2.6)*	(2.4-3.8)*	(5.2-6)	(7.8-10.2)*	

Marginal Dist.	13.5	22.3	20.4	23.1	20.7	*100*
	(13.2-13.8)	(21.9-22.8)	(19.9-20.9)*	(22.5-23.8)	(20.4-21)	

Notes: Percentages of population (weighted sample). * Indicates that the genuine estimate is in the 95% Conf. Interval (in parentheses). Groups in 2005 obtained from real income quintile limits observed in 2002. Each group contains 20% of the households in the baseline. The confidence intervals for the synthetic estimates refer to the 5%-95% quantiles among the distribution of 500 drawings for regime 1, and 5,000 drawings for regime 2.

Table 3. Poverty dynamics (2002-2005) with alternative rho specifications (Model 1, regimes 1 & 2 with $\rho=0.25$)

	Genuine	$\rho=0$	Regime 1	Regime 2	$\rho=1$
A. Using income limits from quintile 1 as poverty line					
Poor 02, Poor 05	6.3	4.7	6.7	6.8	14.5
	(5.5-7.1)*	(4.1-5.4)	(6.1-7.4)*	(5.4-8.2)*	(13.4-15.7)
Poor 02, Non poor 05	13.7	15.3	13.3	13.2	5.5
	(12.4-15)*	(14.6-15.9)	(12.6-13.9)*	(11.8-14.6)*	(4.8-6.1)
Non poor 02, Poor 05	8.5	9.5	6.2	6.6	0.0
	(7.5-9.6)*	(8.5-10.5)*	(5.4-7.1)	(5.1-8.1)	(0-0)
Non poor 02, Non poor 05	71.5	70.5	73.7	73.4	80.0
	(69.7-73.3)*	(69.5-71.5)*	(72.9-74.6)	(71.9-74.9)	(78.8-81.3)
Total	100.0	100.0	100.0	100.0	100.0
B. Using income limits from quintile 2 as poverty line					
Poor 02, Poor 05	21.7	18.2	22.4	22.7	33.5
	(20.2-23.3)*	(17.1-19.3)	(21.3-23.5)*	(20.2-25.2)*	(31.6-34.8)
Poor 02, Non poor 05	18.3	21.9	17.7	17.3	6.6
	(16.9-19.8)*	(20.8-22.9)	(16.6-18.7)*	(14.8-19.9)*	(5.9-7.4)
Non poor 02, Poor 05	13.5	15.7	12.6	13.1	0.1
	(12.3-14.7)*	(14.4-16.9)	(11.4-13.7)*	(10.7-15.4)*	(0-0.2)
Non poor 02, Non poor 05	46.5	44.3	47.4	46.9	59.9
	(44.5-48.4)*	(43-45.5)	(46.2-48.5)*	(44.5-49.2)*	(58.4-61.4)
Total	100.0	100.0	100.0	100.0	100.0

Notes: Percentages of households (weighted sample). Conf. Interval in parentheses. * Indicates that the genuine estimate is in the 95% Conf. Interval. Using upper income quintile limits, as observed in 2002, as poverty lines in both periods. The confidence intervals for the synthetic estimates refer to the 5%-95% quantiles among the distribution of 500 drawings for regime 1, and 5,000 drawings for regime 2.

Appendix 1. Algorithm to calibrate the distribution of the innovation terms

Let $\hat{\varepsilon}_{i(t)t}$ be the residuals of the income equation in period t and $\hat{\varepsilon}_{i(t')t'}$ be the same for the observations in period t'. We first obtain a continuous Gaussian Kernel approximation of the corresponding cumulative distribution functions F_t and $F_{t'}$ as follows:

$$F_\tau(x) = \frac{1}{N_\tau h}\sum_{i=1}^{N_\tau} exp\left[-\frac{\left(x-\hat{\varepsilon}_{i(\tau)\tau}\right)^2}{h^2}\right] \qquad \text{(A1)}$$

where N_τ is the number of observations in the cross-section τ and h is the bandwidth of the kernel approximation. Then define the following approximation of the integral term in (11) in the main text:

$$H_{t'}(x) = \sum_{m=1}^{M} F_t\left[\frac{x-\bar{u}_m}{\hat{\rho}}\right] \cdot g_{t'}^{u}(\bar{u}_m, \theta) \qquad \text{(A2)}$$

Where:

$$\bar{u}_m = (U_m + U_{m-1})/2 \ \text{ and } g_{t'}^{u}(\bar{u}_m, \theta) = \left[\frac{G_{t'}^{u}(U_m;\theta) - G_{t'}^{u}(U_{m-1};\theta)}{U_m - U_{m-1}}\right] \ \text{(A3)}$$

The U_m are M arbitrary real numbers spanning the range of variation of the innovation term and $G_{t'}^{u}(U;\theta)$ stands for the cdf of the innovation term. The calibration of the synthetic panel is based on the assumption that $G_{t'}^{u}(U;\theta)$ is the cdf of a mixture of two normal variables. It is formally given by:

$$G_{t'}^{u}(U|\theta) = p \cdot N\left(\frac{U-\mu_1}{\sigma_1}\right) + \ (1-p) \cdot N\left(\frac{U-\mu_2}{\sigma_2}\right) \qquad \text{(A4)}$$

where N(.) is the cumulative of a Gaussian. The set of parameters that characterize this mixture of normal variables is thus: $(\theta|\rho) = \{p, \mu_1, \sigma_1, \mu_2, \sigma_2\}$. These parameters must satisfy the zero mean constraint on the innovation term:

$$p\mu_1 + (1-p)\mu_2 = 0$$

Finally, (A3) shows how the density is approximated in intervals generated by the grid of real numbers U_m .

The set of parameters θ defining the distribution of the innovation term is obtained by minimizing the following distance between the actual distribution of the residual term in the cross-section t' and the theoretical distribution generated by the AR(1) defined on the residuals of the cross-section *t* and the distribution of the innovation term:

$$Min_{\theta} = \sum_{k=1}^{K} [F_{t'}(x_k) - H_{t'}(x_k)]^2 \qquad \text{(A5)}$$

Where the $x_k{}'s$ are a set of arbitrary values spanning the range of variation of $\hat{\varepsilon}_{i(t')t'}$.

Appendix 2. Additional Tables

Table A1. Descriptive statistics, 2002-2005

Variables	2002				2005			
	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
Ln real income *	6.87	1.32	0.20	11.91	7.04	1.17	1.81	11.38
HH sex (female)	0.18	0.38	0.00	1.00	0.17	0.38	0.00	1.00
HH birth year	1960	9.98	1940	1977	1959	9.83	1940	1977
HH schooling (years)	7.15	4.52	0.00	18.00	7.17	4.55	0.00	18.00
HM aged<3 (dummy)*	0.21	0.41	0.00	1.00	0.14	0.34	0.00	1.00
HM aged 3-24 (2002)	2.34	1.69	0.00	11.00	2.36	1.73	0.00	12.00
HM aged>65 (2002)	0.05	0.23	0.00	2.00	0.05	0.23	0.00	2.00
Urban area	0.76	0.43	0.00	1.00	0.78	0.42	0.00	1.00
Region	1.40	1.03	0.00	3.00	1.41	1.03	0.00	3.00
HH married	0.71	0.45	0.00	1.00	0.72	0.45	0.00	1.00
Real estate & Fin assets	0.04	0.20	0.00	1.00	0.04	0.19	0.00	1.00
Farming assets*	0.09	0.29	0.00	1.00	0.08	0.27	0.00	1.00
Dwellings property*	0.24	0.43	0.00	1.00	0.19	0.39	0.00	1.00

Notes: HH_ household head, HM_ Household members. Using sample weights. * Indicates a statistical difference across the survey rounds.

Table A2. Estimated coefficients of income model, 2002 & 2005

Time invariant variables	**2002**		**2005**	
	lnincome	lnincome	lnincome	lnincome
	(1)	**(2)**	**(1)**	**(2)**
HH Sex (female)	-0.213***	-0.202***	-0.128***	-0.115***
	(0.0492)	(0.0488)	(0.0435)	(0.0432)
HH birthyear	-0.0172***	-0.0156***	-0.0177***	-0.0174***
	(0.00189)	(0.00189)	(0.00166)	(0.00166)
HH Schooling (years)	0.0744***	0.0731***	0.0755***	0.0759***
	(0.00425)	(0.00423)	(0.00372)	(0.00373)
HM aged<3 (dummy)	-0.285***	-0.293***	-0.354***	-0.353***
	(0.0443)	(0.0438)	(0.0451)	(0.0447)
HM aged 3-24 in 2002	-0.136***	-0.136***	-0.127***	-0.126***
	(0.00987)	(0.00977)	(0.00847)	(0.00840)
HM aged>65 in 2002	-0.164**	-0.194***	-0.198***	-0.220***
	(0.0703)	(0.0692)	(0.0625)	(0.0626)
Urban	0.607***	0.665***	0.504***	0.541***
	(0.0352)	(0.0357)	(0.0313)	(0.0317)
Regions	0.118***	0.132***	0.0588***	0.0721***
	(0.0149)	(0.0149)	(0.0130)	(0.0131)
HH Married	-0.0110	-0.0317	0.0617*	0.0559
	(0.0411)	(0.0407)	(0.0364)	(0.0362)
Real Sate & Fin assets		0.383***		0.403***
		(0.0804)		(0.0799)
Farming assets		0.197***		0.139***
		(0.0568)		(0.0538)
Dwellings property		0.143***		0.0778**
		(0.0399)		(0.0385)
Constant	39.92***	36.55***	41.11***	40.39***
	(3.694)	(3.693)	(3.251)	(3.245)
Observations	4,926	4,838	4,748	4,671
Adjusted R-squared	0.246	0.268	0.265	0.283

Note: *p>0.1, **p>0.05, ***p>0.01. Standard errors in parentheses. Sample restricted to heads aged 25-62 as observed in the baseline. HH_ household head, HM_ household member.

Table A3. Transition matrix (Model 2), 2002-2005

	Income Bracket	1	2	3	4	5	*Total*
Synthetic regime 1 2002 Quintiles (Origin)	1	7.4	6.3	3.6	2.1	0.6	*20*
		(6.8-8)	(5.7-6.9)*	(3-4.1)*	(1.7-2.6)	(0.4-0.9)	
	2	3.4	6.0	5.0	4.0	1.6	*20*
		(3-3.9)	(5.3-6.6)*	(4.4-5.6)*	(3.5-4.6)*	(1.2-2)*	
	3	1.7	4.6	5.1	5.5	3.1	*20*
		(1.3-2.1)	(4-5.2)*	(4.5-5.8)*	(4.8-6.2)*	(2.6-3.7)	
	4	0.8	3.1	4.6	6.4	5.2	*20*
		(0.6-1)	(2.6-3.5)	(4-5.2)	(5.7-7)	(4.6-5.9)*	
	5	0.2	1.3	2.8	5.7	9.9	*20*
		(0.1-0.4)	(1-1.8)	(2.2-3.3)*	(5-6.4)	(9.1-10.6)*	
Marginal Dist.		13.5	21.2	21.1	23.7	20.4	*100*
		(12.4-14.6)	(19.7-22.7)*	(19.4-22.8)*	(22-25.4)*	(18.9-22)*	
Genuine 2002 Quintiles (Origin)	1	6.6	6.0	3.5	2.9	1.1	*20*
		(5.7-7.4)	(5.2-6.7)	(2.8-4.2)	(2.2-3.7)	(0.7-1.5)	
	2	3.9	5.7	5.0	4.0	1.4	*20*
		(3.2-4.6)	(4.8-6.6)	(4.3-5.8)	(3.1-4.8)	(1-1.8)	
	3	2.7	4.0	5.8	5.5	2.0	*20*
		(1.9-3.5)	(3.3-4.7)	(4.9-6.7)	(4.7-6.4)	(1.5-2.6)	
	4	1.8	2.5	3.5	7.4	4.8	*20*
		(1.2-2.3)	(1.9-3.1)	(2.8-4.3)	(6.4-8.4)	(4.1-5.5)	
	5	0.6	2.0	2.5	4.7	10.1	*20*
		(0.3-0.9)	(1.3-2.7)	(1.9-3.2)	(3.8-5.6)	(8.8-11.4)	
Marginal Dist.		15.5	20.2	20.4	24.5	19.4	*100*
		(14.2-16.8)	(18.8-21.5)	(18.8-22)	(22.7-26.4)	(17.8-21.1)	
Synthetic regime 2 2002 Quintiles (Origin)	1	7.8	6.0	3.5	2.4	0.4	*20*
		(6.6-8.8)	(5.8-6.2)*	(3.3-3.7)*	(1.6-2.8)	(0.2-0.7)	
	2	3.4	5.8	5.1	3.9	1.8	*20*
		(3.3-3.5)	(5.6-6.2)*	(4.8-5.5)*	(3.8-4.1)*	(1.4-2.1)	
	3	1.7	5.2	5.2	4.8	3.0	*20*
		(1.2-2)	(5-5.5)	(4.8-5.7)	(4.6-5.1)	(2.8-3.2)	
	4	0.9	3.1	5.0	6.0	5.0	*20*
		(0.6-1.2)	(2.9-3.4)	(4.7-5.3)	(5.2-6.7)	(4.9-5.2)	
	5	0.1	1.4	3.2	5.7	9.5	*20*
		(0.1-0.2)	(0.8-1.8)	(2.5-3.6)*	(5.5-6.1)	(8.5-10.3)*	

Marginal Dist.	13.9	21.5	22.0	22.9	19.8	*100*
	(13.5-14.3)	(21-22)	(21.4-22.5)	(22.4-23.3)	(19.5-20)	

Notes: Percentages of population (weighted sample). * Indicates that the genuine estimate is in the 95% Conf. Interval (in parentheses). Groups in 2005 obtained from real income quintile limits in 2002 -when each group contains 20% of the households. Genuine estimates differ between model 1 and 2 due to missing values. The confidence intervals for the synthetic estimates refer to the 5%-95% quantiles among the distribution of 500 drawings for regime 1, and 5,000 drawings for regime 2.

Table A4. 2005 rank test: Synthetic Vs. Genuine conditional on the baseline rank (Model 1, Regime 1)

Mann-Whitney test [H_0: 2005 ranking (synthetic=genuine)]

2002 ventil	Mean of z_i					Share of samples that pass the test				
	ρ=0	ρ=0.15	ρ=0.25	ρ=0.45	ρ=1	ρ=0	ρ=0.15	ρ=0.25	ρ=0.45	ρ=1
1	4.40	0.50	2.75	8.23	16.21	0.00	0.98	0.09	0.00	0.00
2	2.82	0.53	1.43	5.22	14.10	0.08	0.99	0.82	0.00	0.00
3	1.82	0.50	0.88	3.17	9.15	0.57	1.00	0.97	0.00	0.00
4	2.34	0.96	0.50	1.82	8.20	0.27	0.96	1.00	0.61	0.00
5	0.62	0.66	1.08	2.27	5.82	1.00	1.00	0.93	0.27	0.00
6	1.60	0.81	0.57	0.57	3.33	0.69	0.97	0.99	0.99	0.00
7	0.90	0.55	0.50	0.64	2.47	0.96	1.00	1.00	0.99	0.00
8	0.98	0.66	0.62	0.52	0.48	0.92	0.99	0.99	1.00	1.00
9	0.55	0.50	0.49	0.46	0.63	0.99	1.00	1.00	1.00	1.00
10	0.79	0.69	0.73	0.86	1.49	0.98	0.98	0.97	0.95	1.00
11	1.34	0.59	1.09	2.39	5.39	0.79	0.99	0.81	0.77	0.00
12	1.61	0.73	1.05	2.08	5.08	0.77	0.98	0.81	0.80	0.00
13	1.89	1.04	1.23	2.06	5.50	0.68	0.90	0.80	0.80	0.00
14	2.52	1.35	1.37	2.04	6.01	0.39	0.77	0.75	0.80	0.00
15	2.46	1.64	1.06	0.78	5.74	0.27	0.67	0.91	0.97	0.00
16	2.71	1.83	1.09	1.03	6.99	0.14	0.57	0.90	0.95	0.00
17	1.46	0.58	0.75	2.76	9.60	0.78	0.99	0.97	0.12	0.00
18	3.19	1.90	0.92	1.69	8.12	0.04	0.52	0.95	0.74	0.00
19	4.32	2.68	1.54	1.26	7.81	0.00	0.10	0.79	0.93	0.00
20	6.30	3.94	2.05	2.06	12.22	0.00	0.00	0.44	0.42	0.00

Notes: The table shows the test of destination ranks (the rank in 2005 synthetic Vs genuine) conditioned on the real income ventile limits in the baseline (2002). Using Regime 1 (500 reps). H_0: synthetic rank 2005 = genuine rank in 2005. The share of samples, or drawings of residuals, that pass the test refers to those with $z<z_{95\%}$.