On synthetic income panels

François Bourguignon, A. Hector Moreno M.

To cite this version:

François Bourguignon, A. Hector Moreno M.. On synthetic income panels. 2018. halshs-01988068v1

HAL Id: halshs-01988068
https://shs.hal.science/halshs-01988068v1

Preprint submitted on 21 Jan 2019 (v1), last revised 6 Jul 2020 (v2)

WORKING PAPER N° 2018 – 63

On synthetic income panels

François Bourguignon
A. Hector Moreno M.

JEL Codes: D31 ; I32
Keywords : Synthetic panel, income mobility, Mexico

On synthetic income panels *

François Bourguignon[1] and A. Héctor Moreno M.[2]

[1]Paris School of Economics

[2]Paris School of Economics – Université Paris 1 Panthéon-Sorbonne

November 15, 2018

Abstract

In many developing countries, the increasing public interest for economic inequality and mobility runs into the scarce availability of longitudinal data. Synthetic panels based on matching individuals with the same time-invariant characteristics in consecutive cross-sections have been proposed as a substitute to such data - see Dang and Lanjouw (2014). The present paper improves on the calibration methodology of such synthetic panels in several directions: a) it abstracts from (log) normality assumptions; b) it improves on the estimation of auto-correlation of unobserved determinants of (log) earnings; c) it considers the whole 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. The income mobility matrix in the synthetic panel calibrated on the former turns out to be very close to the observed matrix in the latter.

JEL Classification: D31, I32

Key words: Synthetic panel, income mobility, Mexico

[*] We gratefully acknowledge comments on previous versions from Andrew Clark, Hai-Anh Dang, Dean Jolliffe, Marc Gurgand, Jaime Montaña, Christophe Muller, Umar Serajuddin, Elena Stancanelli and attendants to internal seminars at the Paris School of Economics and NEUDC 2015 at Brown University. Hector Moreno acknowledges financial support from CONACYT and CEEY.

1 Introduction

The issue of income mobility is inextricably linked to the measurement of inequality and poverty. Incomes of persons A and B may be very different at both times t and t'. But can this difference be truly considered as inequality if persons A and B switch income level between t and t'? Likewise, should a person above the poverty line in period 1 be considered as non-poor if it is below the line in period 2? Clearly, this depends on how much above the line she was in the first period and how much below in the second. Measuring inequality and poverty in a society may thus be misleading if one uses only a snapshot of income disparities at a point of time instead of individual income sequences.

Longitudinal or panel data that would permit analysing the dynamics of individual incomes are seldom available in developing countries. Yet, snapshots of the distribution of income are increasingly available under the form of repeated cross-sectional household surveys. The idea thus came out 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 ones. They may cover a larger number of periods and they suffer much less from typical panel data problems like attrition, non-response and measurement errors (Verbeek, 2007). But, of course, their reliability depends on the quality of the matching method.

This type of approach has received much attention recently (Dang et.al. , 2014; Cruces et al., 2011, Ferreira et al., 2011,). These papers are based on the methodology designed in Dang et al. (2014) - which was circulated as a working paper in 2011.[1] This methodology permits to obtain an upper and a lower bound of mobility, in and out of poverty, by matching individuals with identical time invariant characteristics and assuming that part of their (log) income that is independent of these characteristics is normally distributed across the two periods with a correlation coefficient equal respectively to 0 or 1. Dang and Lanjouw (2013) refined this method by providing a point estimate of income mobility based on a correlation coefficient estimated through pseudo-panel techniques applied to the two cross-sections.

Unsurprisingly, the properties of such synthetic panels are strongly dependent on the assumptions being made and the way key parameters are estimated. In the methodology designed by Dang and Lanjouw (op. cit.), for instance, the bi-normality assumption made on the joint distribution of initial and final incomes – conditionally on time invariant characteristics

[1] Bourguignon et al. (2004) was an earlier attempt in the same direction.

- and the way the associated coefficient of correlation is estimated strongly influence the synthetic income 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 purpose of the present paper is to analyze in some depth the properties of synthetic panels and their precision in reproducing income dynamics. This is done first by generalizing the original estimation and simulation methodology in Dang et al. (2014) and Dang and Lanjouw (2013) so as to avoid the most arbitrary assumptions found there and then by exploring the 'confidence set' of mobility matrices generated by the confidence intervals on key parameters as the correlation coefficient mentioned above. Departure from previous work includes explicitly involving the calibration of synthetic panels within the realm of AR(1) processes, conditional on time invariant, a more robust estimation of the associated auto-regressive coefficient, and going beyond the normality assumption. Also, the focus of the exercise is the whole income mobility matrix, rather than the share of population moving in and out of poverty.

The validity and the precision of the synthetic panels constructed with that method are tested by comparing the synthetic mobility matrix obtained on the basis of 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 authentic panel. However, simulations performed by allowing the AR(1) coefficient to vary within its estimation confidence interval show a rather high variability of the synthetic mobility matrix and associated income mobility measures. This should plead in favour of extreme caution in analyzing income mobility based on synthetic panel techniques.

The paper is structured as follows. Section two and three describe methodology used in this paper to construct synthetic panels based on AR(1) conditional income processes, comparing it to previous work in this area. Section four present the data used to test this methodology. Section five presents the central results of the whole procedure and compare the central estimate of the synthetic income mobility matrix and various mobility measures to those obtained from the authentic panel. In section six, some sensitivity analysis is performed on various aspects of the methodology so as to test its robustness. The last section concludes.

2 The construction of a synthetic panel

2.1 Matching techniques and the synthetic panel approach

Consider two rounds of independent cross-section data at time t and t'. If $y_{i(\tau)\tau'}$ denotes the (log) income in period τ' of an individual i observed in period τ, what is actually observed is $y_{i(t)t}$ and $y_{i(t')t'}$.[2] Constructing a synthetic panel is somehow 'inventing' a plausible value for $y_{i(t)t'}$.

A 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 time invariant attributes observed in the two cross-sections is estimated with OLS:

$$y_{i(\tau)\tau} = z_{i(\tau)}\beta_\tau + \epsilon_{i(\tau)\tau} \text{ for } \tau = t, t' \tag{1}$$

where β_τ represents the vector of 'returns' to fixed individual attributes, z, and $\epsilon_{i(\tau)}$ denotes a 'residual' that stands for the effect of time variant individual characteristics and other unobserved time invariant attributes. Fixed attributes may include year of birth, region of birth, education, parent's education, etc. More on this in a subsequent section. For now it is just enough to stress that it would not make sense to introduce time-varying characteristics in the income model (1), even though some of them may be observed as their value in the terminal (or initial) year are essentially unknown.

Denote $\hat{\beta}_\tau$ and $\hat{\epsilon}_{i(\tau)\tau}$ and $\hat{\sigma}^2_\tau$ at time τ=t,t' respectively the vector of estimated returns, the corresponding residuals and their variance as obtained from OLS:

$$y_{i(\tau)\tau} = z_{i(\tau)}\hat{\beta}_\tau + \hat{\epsilon}_{i(\tau)\tau} \text{ for } \tau = t, t' \tag{2}$$

Consider now an individual i observed in the first period, t. 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 OLS estimates. The remaining is the change in the residual term: $\hat{\epsilon}_{i(t)t'} - \hat{\epsilon}_{i(t)t}$. The problem is that the first term in this difference is not observed. The

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

issue in constructing a synthetic panel thus is the way of finding a plausible value for it. Let $\tilde{\epsilon}_{i(t)t'}$ be that 'virtual' residual. At this stage, the only information available about it is that it has zero mean and variance $\hat{\sigma}^2_{t'}$ across the whole population.

2.2 Previous approaches

In their first attempt at constructing synthetic panels, Dang et al. (2011, 2014) simply assumes the virtual residual at time t' to be normally distributed conditional on the residual $\hat{\epsilon}_{i(t)t}$ at time t with an arbitrary correlation coefficient, ρ. Assuming that the initial residual is also normally distributed, then the synthetic income mobility process can be described by the joint CDF:

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

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

In their initial paper, Dang et al. (2011,2014) considered the two extreme cases of ρ=0 and ρ=1, so as to obtain an upper and a lower limit on mobility. Applying this approach to the probability of getting in or out of poverty in Peru and in Chile, the corresponding ranges proved to be rather broad. In other words, the change $(\hat{\beta}_{t'} - \hat{\beta}_t)$ in the returns to fixed attributes was playing a limited role in explaining income mobility.

In a later, unpublished paper, Dang & Lanjouw (2013) generalized 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 birth cohorts in the two samples, ρ^{y_c}, as in pseudo-panel analysis. Then, the covariance between (log) incomes is approximated by $cov_y = \rho^{y_c} \cdot \sigma^2_{yt}\sigma^2_{yt'}$ where $\sigma^2_{y\tau}$ is the variance of (log) income at time τ. Then it comes from the two equations in (2), if both applied to the same sample of individuals, that:

$$Cov_y = \beta_t' Var(z)\beta_{t'} + Cov_\epsilon \tag{3}$$

where Var(z) is the variance-covariance matrix of the fixed 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 then possible to get an approximation of the correlation coefficient between these residuals.

This appears as a handy way of getting an estimate of the correlation coefficients between initial and terminal cross-sections by relying on their pseudo-panel dimension. Yet, it will be seen below that this method is not fully correct.

2.3 Synthetic panels with AR(1) residuals

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

$$y_{i(t)t'} = z_{i(t)}\beta_{t'} + \epsilon_{i(t)t'} \text{ with } \epsilon_{i(t)t'} = \rho\epsilon_{i(t)t} + u_{i(t)t'} \tag{4}$$

where the 'innovation terms', $u_{i(t)t'}$, are assumed to be orthogonal to $\epsilon_{i(t)t}$ 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 such that: $0 < \rho < 1$.

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, income is regressed on time invariant attributes in the two periods as in (2). Equation (4) can then be used to figure out what the residual of the income model, $\tilde{\epsilon}_{i(t)t'}$ could be in time t' for observation i(t):

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

where $\tilde{u}_{i(t)t'}$ has to be drawn randomly within the distribution of the innovation term, of which CDF will de denoted $G^u_{t'}$. If estimations or approximations of ρ and the distribution $G^u_{t'}$ are available, the virtual income of individual i(t) in period t' can be simulated as:

$$\tilde{y}_{i(t)t'} = z_{i(t)}\hat{\beta}_{t'} + \rho\hat{\epsilon}_{i(t)t} + G^{u\,-1}_{t'}(p_{i(t)}) \tag{5}$$

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

$$\tilde{y}_{i(t)t'} = \rho y_{i(t)t} + z_{i(t)}(\hat{\beta}_{t'} - \rho\hat{\beta}_t) + G^{u\,-1}_{t'}(p_{i(t)}) \tag{6}$$

Thus the virtual income in period t' of individual i(t) observed in period t depends on his/her observed income in period t, $y_{i(t)t}$, his/her observed fixed attributes, $z_{i(t)}$, and a random term drawn in the distribution $G^u_{t'}$. Because those virtual incomes are drawn randomly for each individual observed in period t, the income mobility measures derived from this exercises necessarily depends on the set of drawings. Various simulations will have to be performed to compute the expected value of these measures - and, actually, their distribution.

The two unknowns, ρ and $G^u_{t'}(\cdot)$ 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 first focus on the estimation of the auto-regressive coefficient, ρ through pseudo-panel techniques.

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) - see in particular Moffit (1993), McKenzie (2004) and Verbeek (2007). We very much follow the methodology proposed by the latter when estimating dynamic linear models on repeated cross-sections. Note, however, that in comparison with this literature, a specificity of the present methodology is to rely on only two rather a substantial number 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: year of birth - as in Dang and Lanjouw - but possibly regions of birth, school achievement, gender, etc... The important assumption in defining these groups of observations is that the AR(1) coefficient should reasonably be identical for all of 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:

$$\bar{\hat{\epsilon}}_{gt'} = \rho \bar{\hat{\epsilon}}_{gt} + \eta_{gt'} \tag{7}$$

where $\bar{\hat{\epsilon}}_{g\tau}$ 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 $\bar{\hat{\epsilon}}_{gt}$ with variance σ_u^2/n_{gt} where n_{gt} is the number of observations in group g. The estimation of (7) raises a major difficulty, however. It is that the group means of residuals of OLS regressions are asymptotically equal to zero at both dates t and t' so that (6) 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:

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

for each group g leads to:

$$\sigma^2_{\epsilon gt'} = \rho^2 \cdot \sigma^2_{\epsilon gt} + \sigma^2_{ugt'}$$

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

$$\sigma^2_{\epsilon gt'} = \rho^2 \cdot \sigma^2_{\epsilon gt} + \sigma_u^2/n_{gt} + \omega_{ut'} \tag{8}$$

where $\omega_{ut'}$ stands for the deviation between the group variance of the innovation term and its expected value and can thus be assumed to be zero mean, 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:

$$\overline{y}_{gt'} = \rho\overline{y}_{gt} + \overline{z}_{gt}\gamma + \overline{u}_{gt'} \tag{9}$$

where it has been reasonably assumed that $\overline{z}_{gt}$ and $\overline{z}_{gt'}$ were close to each other 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 $\overline{u}_{gt'}$ is heteroskedastic with variance σ_u^2/n_{gt}.

Note that this approach departs from Dang and Lanjouw (2013). As seen above they derive the covariance of residuals from the covariance of (log) incomes through (3). The latter is estimated through OLS applied to:

$$\overline{y}_{gt'} = \delta\overline{y}_{gt} + a + \theta_{gt'} \tag{10}$$

and $cov_y = \hat{\delta}\sigma_{yt}\sigma_{yt'}$. As can be seen from (9), however, a term in $\overline{z}_{gt}$ is missing on the RHS of (10), which means that the residual term $\theta_{gt'}$ is not independent of the regressor $\overline{y}_{gt}$. It follows that $\hat{\delta}$ is biased, the same being true of the covariance of (log) incomes.

The two approaches proposed above to get an unbiased estimate of the auto-regressive coefficient ρ can be combined by estimating (8) and (9) simultaneously. 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-groupings. For instance, there may be good reasons to expect that ρ declines with age. Of course, this would require that individuals are described by enough fixed 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.

2.3.2 Calibrating the distribution of the innovation terms

It turns out that once an estimate of the autoregressive coefficient ρ is available, the distribution $G^U_{t'}(\cdot)$ the innovation terms, $u_{i(t)t'}$, be recovered from the data.

The AR(1) specification implies:

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

where ρ is the pseudo-panel estimator obtained in (8) or (9), the $\tilde{\epsilon}_{i(t)t'}$ are the virtual residuals and the $\tilde{u}_{i(t)t'}$ are the randomly generated innovation terms. The problem is to find the distribution $G^U_{t'}(\cdot)$ of the innovation terms such that the distribution of the virtual residuals be the same as the distribution, $F_{t'}$ of the observed OLS residuals $\hat{\epsilon}_{i(t')t'}$ obtained with the income regression (1). With a continuous time formulation, this distribution must satisfy the following functional equation:

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

where F_τ is the cdf of the observed residuals $\hat{\epsilon}_{i(\tau)\tau}$ and $g^u_{t'}$ the density of the innovation term. Hence, knowing the distribution of the residuals in the two periods and the autocorrelation coefficient it is logically 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.

Yet 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 chosen based on the assumption that the distribution $g^u_{t'}$ is a mixture of normal variables, whose parameters are to be determined so as to minimize the square of the difference between the two sides of (11). It turned out to give rather satisfactory results but this is only an approximation, which justifies describing that methodology as a 'calibration' rather than an 'estimation'. The detail of the calibration of the distribution $G^u_{t'}$ with a mixture of two normal distributions is given in the **Appendix A** to this paper.

2.4 Practical summary

Practically, the whole procedure leading to the construction of a synthetic 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 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, $\hat{\epsilon}_{i(t)t}$ and $\hat{\epsilon}_{i(t')t'}$, and the returns to time invariant attributes, β_t and $\beta_{t'}$.

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).

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

3. Distribution of innovation terms. Calibrate the set of parameters, $\theta|\hat{\rho}$, of the distribution of the innovation term supposed 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 (6). Evaluate income mobility matrices and mobility measures based on that drawing.

5. Repeat 4 to obtain the expected value and distribution of the mobility matrices and measures.

3 Construction and validation: Mexico 2002-2005

The procedure detailed above allowed us to estimate the synthetic income in 2005 for the households sampled in 2002. Synthetic panel can be obtained either at household or individual level, although the former bring about access to a larger set of time-invariant attributes. We focused on households as observational units, as these tend to offer a wider perspective of family wellbeing and gave access to a larger set of time-invariant attributes. The results in terms of income mobility were then compared to the actual income mobility observed in the panel.

3.1 Data

The data corresponds to the Mexican Family Life Survey (MxFLS onwards). It is based on a sample of households that is representative at national, regional and urban-rural level. It was fielded by the National Institute of Statistics (INEGI by its acronym in Spanish) but was coded and critically assessed by its study directors. The MxFLS is a multi-thematic and longitudinal database, which gathers information on socioeconomic indicators, migration, demographics and health indicators on the Mexican population. This panel survey 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 collected 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 in the second wave. Due confidentiality, data on the simple design (primary/secondary sampling units) are not public (see MXFLS website).

3.2 Income estimates

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.[3] Total income is then divided by the household size in order to obtain per capita income and is deflated by the Consumer Price Index (August 2005=100) to make 2002 and 2005 data comparable.

In order to focus on the steadiest set of households and to facilitate the use pseudo-panel instruments, the sample was restricted to households whose head was aged between 25 and 62 years in 2002 which is the baseline (28-65 years old in 2005). Finally, to overcome possible adverse effects due to atypical observations two percent of the sample in the two ends of the income distribution and households with missing income were discarded.

3.3 Time invariant attributes and the income models

Time-invariant attributes could stem from multiple criteria and sources. Individual deterministic attributes like the year of birth, sex, educational achievement and ethnicity are the most natural set of characteristics. Depending on the issue of interest, the time horizon and country studied other variables can be obtained from the household characteristics like the household size which could be introduced in terms of its demographic composition. Consider also the location, the population density in the area of residence (urban or rural localities), and the state or regional fixed effects depending on the territorial representativeness of the survey. Needless to say, all variables ought to strictly follow the same definition and construction in all periods.

It is reasonably questioning how realistic is the assumption of 'time invariance' of these attributes. In this respect, it helps to bear in mind that the longer the period between the cross-sections, the more severe ought to be the time invariability criterion. The long-standing feature of these attributes is perhaps more important than the number of variables when conceiving the specification of the income model. Many variables are not strictly time-invariant and should easily be discarded like current employment status and occupation but this has to be considered on the particular case of the country under analysis. Other variables could be considered time-invariant under reasonable circumstances, like marital status and highly-valuable wealth possessions (dwelling or physical assets) during periods of economic stability.

We followed a grading approach by the use of alternative model specifications to assess the

[3] This definition changed to introduce a multidimensional poverty approach. See CONEVAL (2013).

sensibility of variables selection. All of them stuck to a strict degree of time invariability. The first specification uses the head's individual characteristics like gender, formal years of schooling, birth year and the household composition by age groups. It also includes variables for the size of the locality (urban/rural), marital status and regions.[4] An alternative specification includes long-lasting productive assets such as real estate and farming assets (land for agricultural production and cattle), and household dwelling as well as the possession of other dwellings other than the one in use. This is our most preferred model. See **Appendix B** for descriptive statistics and OLS estimates.

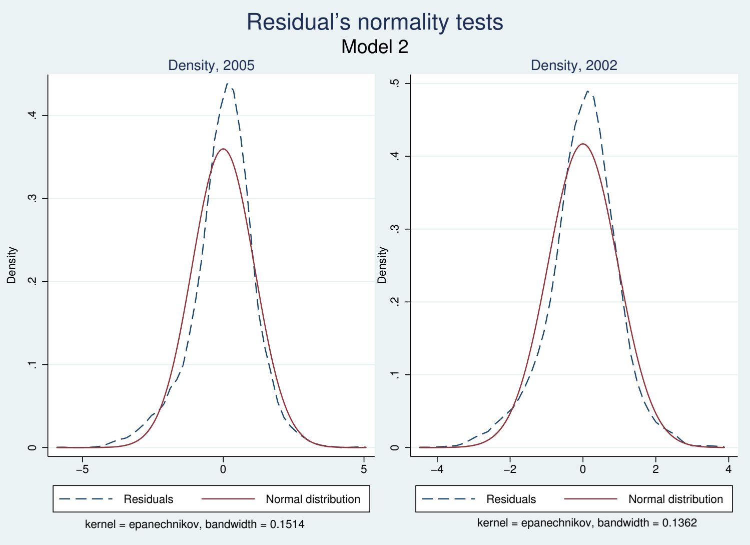
It is important to mention some restrictions encountered to enrich the income model. The survey collected data on ethnicity, religious conviction and household head literacy. Also contains data on historic or retrospective data like birth city size; the year of marriage; household's head's parents' education, place of birth and migration records. Those attributes, like many others, were gathered by the survey but finally not included in the income model due to: 1) high prevalence of missing data, 2) lack of statistical significance, or 3) extremely low frequency.

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 the **Graph 1** shows the kernel distribution of (log) income residuals for the last model specification in both years, and compares it with the normal distribution. This and the Skewness and Kurtosis tests along with the Shapiro-Wilk normality tests confirm that the normality assumption in the distribution of residuals is strongly rejected.[5]

[4] Of course, the problem here is how frequent migration may be but its effect is expected to be low during this short period analysis. According to census data the internal migration rate in Mexico, from 2000 to 2005, was around 2% (Chavez & Wanner, 2012).

[5] Skewness & Kurtosis tests rejects the null hypothesis of normality [Chi2=(172.75 & 301.58) with Prob>Chi2=(0.00 & 0.00) for 2005 and 2002 respectively. The Shapiro-Wilk W test reject the hypothesis that both residuals are normally distributed [W=(0.98 & 0.98), V=(36.1 & 57.6) with Prob>Z=(0.00 & 0.00) for 2005 and 2002 respectively. Large values of V (larger than 1=median) indicate non normality (95% critical values are between 1.2-2.4).

Figure 1: **Residual's normality test for 2002 and 2005**



3.4 The autocorrelation coefficient and the calibration parameters

Estimating the autocorrelation coefficient is a central, and a sensitive task in this procedure. This section aims at obtaining an autocorrelation coefficient with the two waves of cross-sectional data at hand. Firstly, observations were grouped by some common characteristics. In our case, thirty-five clusters were obtained by the interaction of seven birth-year cohorts, of 6 years interval each, and five 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.[6]

Table 1 shows separate estimates from equations 8 and 9. Results have the expected signs and order of magnitude. The genuine coefficient here served as a benchmark and appears to be around 0.25. Though the genuine parameter is close to these set of estimates it is reassuring that the combined use of these two approaches, through a non-linear equation system, delivers a more accurate estimate whose confidence intervals are fully consistent with those from the actual panel.[7]

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

[7] Similar results are obtained with twenty-eight clusters by the use of four, instead of five, educational groups.

Table 1: **Rho estimates by model and method, 2002-2005**

	Pseudo panel			**Genuine panel**
Models	Equation 8 Non linear (1)	Equation 9 Linear (2)	Eq. system (8, 9) Non linear (3)	With microdata (residuals) (4)
Model 1	0.292*	0.132	0.254**	0.257***
	(-0.054 - 0.638)	(-0.139 - 0.404)	(0.042 - 0.466)	(0.235 - 0.280)
Model 2	0.176	0.158	0.299***	0.226***
	(-0.823 - 1.174)	(-0.100 - 0.416)	(0.145 - 0.452)	(0.203 - 0.249)

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

The rho estimate and its corresponding 95% confidence interval now enabled us to determine the set of calibration parameters, $(\theta|\hat{\rho})$, from the empirical basis of two normal variables. We followed two regimes. The first regime employs the point estimate of rho reported in Table 1. The second is based on $i = 100$ different rhos, with their corresponding set of calibration parameters $(\theta_i|\hat{\rho}_i)$. In this case, rho is randomly obtained from a normal distribution within its 95% confidence interval. This means that the mean of this random drawings corresponds to the point estimate, but some of them might deviate. **Table 2** shows the descriptive statistics of the resulting parameters for each regime and model. These parameters characterize the distribution of the innovation terms which was the last input to compute the expected value of mobility measures and their distribution as described in the following section.

Table 2: **Innovation terms' calibration parameters by regime and model**

	Regime 1	**Regime 2**			
Parameters	Point estimate (1)	Mean (2)	Std. Dev. (3)	Min (4)	Max (5)
Model 1					
μ_1	-0.03	0.00	0.02	-0.10	0.04
σ_1	1.23	1.13	0.11	0.78	1.78
p_1	0.54	0.51	0.03	0.35	0.61
μ_2	-0.04	0.00	0.01	-0.09	0.04
σ_2	1.02	1.10	0.10	0.80	1.26
Model 2					
μ_1	-0.04	0.00	0.01	-0.06	0.03
σ_1	1.36	1.10	0.10	0.80	1.26
p_1	0.55	0.51	0.03	0.40	0.56
μ_2	-0.04	0.00	0.01	-0.06	0.04
σ_2	0.46	1.07	0.09	0.89	1.30

Note: Parameters in regime 2 obtained from 100 optimization processes.

3.5 Estimation results

This section provides empirical estimates from a household level synthetic panel over a period characterized by positive economic growth in Mexico.[8] We first examined the shape of a synthetic distribution for 2005 compared with the genuine income distribution. **Graph 2** shows the kernel density for both regimes to provide a first visual element to assess the shape of the distribution at every income level. This preliminary inspection shows that even a basic model specification is capable of reproducing the shape of the actual income distribution.

[8] According to World Bank's World Development Indicators, the Mexican economy grew by 0.8%, 4.0% and 3.2% from 2002 - 2005 in annual basis.

Figure 2: **Genuine and synthetic income by model and regime**

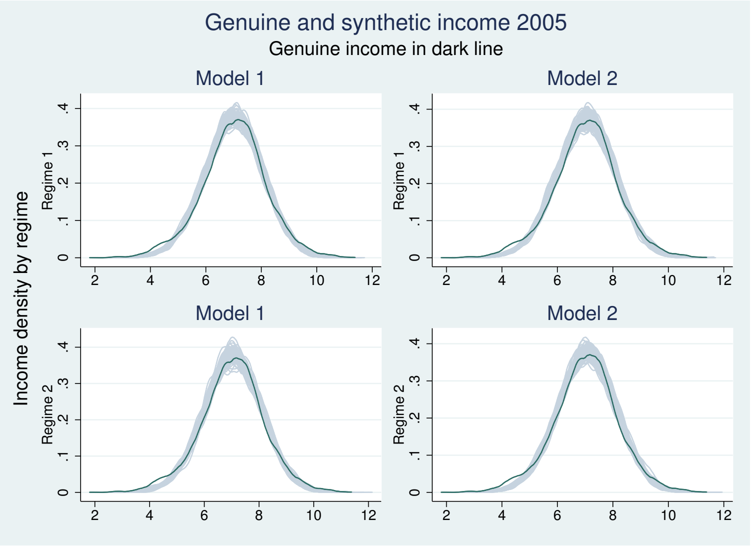


Table 3 shows the resulting synthetic panel 2002-2005 through a transition matrix defined on the income brackets from the quintiles of 2002. This means that the marginal income distribution in the base year is the same in the genuine and the synthetic panels by construction. Each line shows the movement of individuals that belonged to a specific income quantile in the baseline over the same, real income, references in the final year. The table contains three sections. Section A and C correspond to two different regimes respectively: regime 1 computes the virtual income using the point estimate of rho and 500 repetitions, whereas regime 2 performs it with 100 random draws of rho and the corresponding calibration parameters for each. Section B contains the genuine estimates.

The synthetic figures appear close to the genuine ones and fall within their 95% confidence intervals (reported in parentheses). As expected, working with various values of rho, i.e. regime 2, deliver slightly larger confidence intervals so we stick to this regime on the remaining part of the document. In general, both the genuine and the synthetic panels suggest a process of upward mobility implied by a reduction in the share of households below the income limits of the first quintile from 2002 to 2005.

We also used the Mann-Whitney test to assess the synthetic rank distribution of 2005 conditioned on its rank at the origin. The test delivers a statistic based on the difference between the sum of the ranks of both distributions: the genuine and the synthetic one. To increase the sensitivity of the test we use twenty equally sized groups.[9] Results in **Table 4** shows

[9] This test utilizes information regarding the rank order and constitutes an alternative for the two-sample

Table 3: **Transition matrix, 2002-2005**

Confidence intervals in parentheses
Percentage of population

		2005 groups (Destination)					
		1	2	3	4	5	Total
A. Synthetic panel (Regime 1)							
	1	7.5	6.3	3.6	2.1	0.6	20
		(6.9-8.1)	(5.7-7)	(3-4.1)	(1.7-2.5)	(0.4-0.8)	
	2	3.5	6.1	5.0	3.9	1.5	20
2002		(3-4)	(5.4-6.7)	(4.4-5.7)	(3.3-4.6)	(1.1-1.9)	
Quintiles	3	1.7	4.7	5.2	5.5	3.0	20
(Origin)		(1.4-2.2)	(4.1-5.3)	(4.5-5.9)	(4.8-6.1)	(2.5-3.5)	
	4	0.8	3.2	4.7	6.4	5.0	20
		(0.6-1.1)	(2.6-3.7)	(4.1-5.3)	(5.7-7)	(4.4-5.6)	
	5	0.2	1.4	2.9	5.8	9.6	20
		(0.1-0.4)	(1-1.8)	(2.4-3.6)	(5.1-6.5)	(8.9-10.3)	
Marginal Dist.		13.7	21.6	21.4	23.7	19.6	100
B. Authentic panel							
	1	6.6	6.0	3.5	2.9	1.1	20
2002	2	3.9	5.7	5.0	4.0	1.4	20
Quintiles	3	2.7	4.0	5.8	5.5	2.0	20
(Origin)	4	1.8	2.5	3.5	7.4	4.8	20
	5	0.6	2.0	2.5	4.7	10.1	20
Marginal Dist.		15.5	20.2	20.4	24.5	19.4	100
C. Synthetic panel (Regime 2)							
	1	7.4	6.3	3.6	2.1	0.6	20
		(6.2-8.6)	(5.7-6.9)	(3-4.2)	(1.4-2.9)	(0.2-1)	
2002	2	3.3	6.1	5.1	4.0	1.5	20
Quintiles		(2.9-3.8)	(5.3-6.9)	(4.4-5.8)	(3.4-4.5)	(1-2.1)	
(Origin)	3	1.6	4.6	5.3	5.6	3.0	20
		(1.1-2.2)	(3.9-5.2)	(4.5-6)	(4.9-6.3)	(2.5-3.6)	
	4	0.8	3.0	4.7	6.4	5.1	20
		(0.3-1.1)	(2.4-3.6)	(4.1-5.3)	(5.4-7.3)	(4.6-5.8)	
	5	0.2	1.4	2.8	5.7	9.9	20
		(0-0.5)	(0.7-2)	(2-3.3)	(4.8-6.6)	(8.6-11)	
Marginal Dist.		13.3	21.4	21.4	23.8	20.1	100
D. Ratio of marginal distributions (synthetic/genuine)							
Regime 1		0.89	1.07	1.05	0.97	1.01	1.0
Regime 2		0.86	1.06	1.05	0.97	1.04	1.0

Notes: 95% confidence interval (C.I.) in parentheses. Groups in 2005 from real income quintile limits in the baseline. Regime 1 refers to 500 random repetitions from one set of calibration parameters. Regime 2 refers to 100 optimization processes where each rho is randomly drawn within its C.I.

that our synthetic estimates satisfactorily reproduce the dynamic described by the genuine panel in almost all points of the distribution, the exception being the ventile at the bottom of the 2002 income distribution. A high share of samples passed this test in most of remaining groups.

Table 4: **Rank test: Synthetic Vs. Genuine**

Mann-Whitney Test conditional on the baseline rank

2002	z	Prob.> $\|z\|$	Share of samples	2002	z	Prob.> $\|z\|$	Share of samples
ventil			that pass the test	ventil			that pass the test
1	3.92	0.00	0.10	11	0.88	0.38	0.93
2	1.86	0.06	0.60	12	0.90	0.37	0.91
3	1.04	0.30	0.87	13	0.97	0.33	0.92
4	0.74	0.46	0.97	14	1.10	0.27	0.89
5	0.80	0.42	0.98	15	0.99	0.32	0.92
6	0.52	0.60	1.00	16	1.06	0.29	0.92
7	0.55	0.58	0.99	17	1.86	0.06	0.63
8	0.94	0.35	0.96	18	0.96	0.34	0.92
9	0.54	0.59	1.00	19	1.23	0.22	0.82
10	0.72	0.47	0.99	20	1.10	0.27	0.83

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). H_0: 2005 synthetic rank = 2005 genuine rank. The share refers to samples with $z < z_{95\%}$. Weighted sample restricted to household's heads aged 25-62 in 2002. Repetitions from 100 optimizations.

Poverty dynamics is the most popular empirical application of this type of procedures. To illustrate the performance of this approach on this issue we computed two sets of poverty transitions, in-and-out of poverty, using the upper limits from the first two income quintiles as poverty thresholds. These thresholds constitute a direct reference to the 'shared prosperity' goal adopted by the World Bank recently.

Table 5 shows that the proposed approach delivers an encouraging approximation to actual figures in all poverty transitions. For instance, our estimate for persistent poverty for =0.20 (and 0.30), being poor in both periods, using the first poverty line is 6.5% (7.3% respectively)

t-test of independent samples (Kirk, 2008).

whereas the actual figure is 6.6%. The largest difference is found in the downward mobility group. Larger values of rho illustrate the sensitivity of this type of methodologies to this parameter. Interestingly substantial differences emerge when using a correlation coefficient that separates from the actual parameter.[10] Note that this occurs on top of the calibration procedure here implemented. These results reinforce the utterly importance of this parameter with this and similar methodological approximations.

Table 5: **Poverty dynamics with alternative poverty lines and rho values**

Percentage of households

	Genuine ρ (1)	$\rho = 0.20$ (2)	$\rho = 0.30$ (3)	$\rho = 0.40$ (4)	$\rho = 0.50$ (5)	$\rho = 0.60$ (6)
A. Using income limits from quintile 1 as poverty line						
Poor 02, Poor 05	6.6	6.5	7.4	8.4	9.6	10.9
Poor 02, Non poor 05	13.5	13.5	12.7	11.6	10.4	9.1
Non poor 02, Poor 05	8.9	6.6	6.0	5.1	4.2	3.4
Non poor 02, Non poor 05	71.1	73.3	74.0	74.9	75.8	76.6
Total	100.0	100.0	100.0	100.0	100.0	100.0
B. Using income limits from quintile 2 as poverty line						
Poor 02, Poor 05	22.1	21.8	23.1	24.8	26.4	28.1
Poor 02, Non poor 05	17.9	18.2	16.9	15.2	13.6	11.9
Non poor 02, Poor 05	13.5	12.7	11.5	10.1	8.6	7.0
Non poor 02, Non poor 05	46.5	47.3	48.5	49.9	51.4	53.0
Total	100.0	100.0	100.0	100.0	100.0	100.0

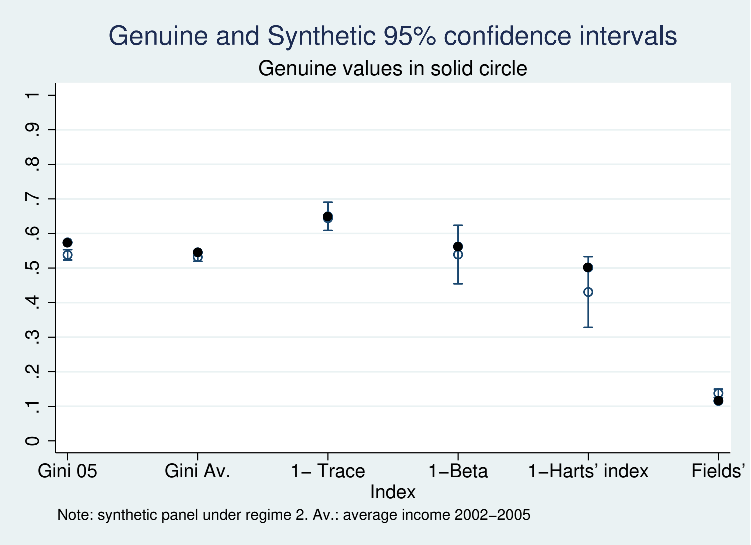
Notes: Using upper income quintile limits as observed in 2002 as poverty lines in both periods.

Lastly, once equipped with a synthetic estimate for each household we also computed some measures of income inequality and income mobility. Income mobility indicators constitute natural candidates for a final robustness check. Graph 4 plots a 95% confidence interval for these inequality measures against their true point-estimate. Results suggest no statistically significant differences between the synthetic and the authentic estimates. Similarly, the four

[10] In each case, one-hundred optimization process were obtained from random values of rho being drawn within a 95% confidence interval following a normal distribution $\mathcal{N}(\rho_i, SE_\rho)$. $SE_{\rho=0.29}$ refers to the standard errors for rho in Table 1. The hypothetical values for rho are i = {0.20, 0.30, 0.40, 0.50, 0.60}.

most popular mobility indicators (the Hart's index, the 1-Beta index, the 1-Trace index, and the Fields' index of 'Mobility as an Equalizer of Longer-term Incomes') appear in line with the actual estimates. Together our synthetic results confirm some process of upward mobility in Mexico during 2002-2005.

Figure 3: **Income inequality and income mobility indicators**



3.6 Concluding remarks

This document proposes an alternative approach to improve the construction of synthetic panels using micro data from repeated cross-sections. We performed an empirical validation through the use of two consecutive waves of a genuine panel survey in Mexico. The procedure delivered very satisfactory results as the marginal distribution of the synthetic income accurately reproduced the genuine one and the resulting panel was consistent with its genuine panel dimension. This was confirmed by multiple tests in several potential applications. The proposed approach allowed examining the importance of the autocorrelation coefficient used in this and other synthetic panel methodologies. Indeed, this parameter constitutes a central and sensitive component in the construction of synthetic panels. Our results seem of sufficient quality to envisage a systematic application of this methodology with sequences of two cross-sectional household surveys over a longer time span so as to study a possible evolution in the income mobility of the population on top of that of instantaneous inequality.

4 Bibliography

Arellano and Bond. 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 and McKenzie. 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 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.

Bourguignon, Goh and Kim. 2004. “Estimating individual vulnerability to poverty with pseudo panel-data”. World Bank Policy research paper 3375. August 2004.

Browning, Deaton and Irish. 1985. A Profitable approach to Labour supply and commodity Demands over Life Cycle, Econometrica, 50. Number 3. May 1985.

Chávez-Juárez and Wanner. 2012. “Determinants of Internal Migration in Mexico at an Aggregated and a Disaggregated Level” (March 26, 2012).

Cruces, Lanjow, Luccetti, Perova, Vakis and Viollaz. 2011. “Intra-generational Mobility and Repeated Cross-Sections: A three country validation exercise”. The World Bank. Latin-American and the Caribbean Region. Poverty, Equity and Gender Unit. December 2011. Policy Research Working Paper 5196.

Dang and Lanjouw. 2015. Poverty Dynamics in India between 2004 and 2012: Insights from Longitudinal Analysis Using Synthetic Panel Data”. World Bank Policy Research paper 5916. Policy Research Working Paper 7270.

Dang, Lanjow, Luoto and McKenzie. 2014. “Using Repeated Cross-Sections to Explore Movements in and out of Poverty”, Journal of Development Economics 107 (2014). Elsevier.

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

Dang, Lanjow, Luoto and McKenzie. 2011. “Using Repeated Cross-Sections to Explore Movements in and out of Poverty”, January. The World Bank. Policy Research Paper, 5550.

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

Ferreira, Messina, Rigolini, Lopez, Lugo, and Vakis. 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.

Filmer, D. and Pritchett, L. 1994. “Estimating Wealth Effects without Expenditure Data - or Tears: An Application to Educational Enrolments in States of India. The World Bank Policy Research Working Paper. WPS 1994.

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

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

Rubalcava, Luis y Teruel, Graciela (2006). “Mexican Family Life Survey, First Wave”, Working Paper, www.ennvih-mxfls.org

Rubalcava, Luis y Teruel, Graciela (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.

Appendices

A Algorithm to calibrate the distribution of the innovation terms

Let $\hat{\epsilon}_{i(t)t}$ be the residuals of the income equation in period t and $\hat{\epsilon}_{i(t')t'}$ 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:

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

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:

$$H_{t'}(x) = \sum_{m=1}^{M} F_t\left[\frac{(x-\overline{u}_m)}{\hat{\rho}}\right]\cdot g^u_{t'}(\overline{u}_m,\theta) \tag{A2}$$

Where $\overline{u}_m = (U_m - U_{m-1})/2$ and,

$$g^u_{t'}(\overline{u}_m,\theta) = \left[\frac{G^u_{t'}(u_m;\theta) - G^u_{t'}(u_{m-1};\theta)}{u_m - u_{m-1}}\right] \tag{A3}$$

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

$$G^u_{t'}(U|\theta) = p_1\cdot\mathcal{N}\left(\frac{U-\mu_1}{\sigma_1}\right) + (1-p_1)\cdot\mathcal{N}\left(\frac{U-\mu_2}{\sigma_2}\right) \tag{A4}$$

where N( ) is the cumulative of a Gaussian. The set of parameters that characterize this mixture of normal variables is thus: $(\theta|\rho) = (p_1,\mu_1,\sigma_1,\mu_2,\sigma_2)$. These parameters must satisfy the zero mean constraint on the innovation term:

$$p_1\mu_1 + (1-p_1)\mu_2 = 0$$

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:

$$\min_{\theta} = \sum_{k=1}^{K} \Big[F_{t'}(x_k) - H_{t'}(x_k) \Big]^2 \tag{A5}$$

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

B Additional tables

Table 1 **Descriptive statistics, 2002-2005**

	2002				**2005**			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Ln real income	6.77	1.30	0.20	11.91	6.99	1.13	1.81	11.38
HH sex (female)	0.17	0.38	0.00	1.00	0.17	0.37	0.00	1.00
HH birth year	1959	9.9	1940	1977	1959	9.8	1940	1977
HH schooling (years)	6.72	4.38	0.00	18.00	6.73	4.38	0.00	18.00
HM aged<3 (dummy)	0.20	0.40	0.00	1.00	0.14	0.34	0.00	1.00
HM aged 3-24 (2002)	2.39	1.70	0.00	11.00	2.44	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	0.59	0.49	0.00	1.00	0.63	0.48	0.00	1.00
Region	1.50	1.09	0.00	3.00	1.51	1.09	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	0.03	0.18	0.00	1.00
Farming assets	0.10	0.30	0.00	1.00	0.09	0.28	0.00	1.00
Dwellings property	0.24	0.42	0.00	1.00	0.19	0.39	0.00	1.00

Notes: HH: household head, HM: Household members

Table 2: **Estimated coefficients of income model, 2002 & 2005**

	2002	2002	2005	2005
Time invariant variables	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 St. & Financial 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: Standard errors in parentheses *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Sample restricted to household heads (HH) aged 25-62 as observed in the baseline. HM stands for household member.