On synthetic income panels

François Bourguignon, A. Hector Moreno M.

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

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

Funded by a French government subsidy managed by the ANR under the framework of the Investissements d'avenir programme reference ANR-17-EURE-001

On synthetic income panels *

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

[1]Paris School of Economics (PSE)
[2]The University of Oxford (OPHI)

July 6, 2020

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 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 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 turns out to be close to the genuine matrix observed in the panel. However, this need not be true for extreme values of the AR coefficient in the confidence interval of its estimation.

JEL Classification: D31, I32

Key words: Synthetic panel, income mobility, Mexico

*We gratefully acknowledge comments on previous versions from Andrew Clark, Hai Ann Dang, Dean Jolliffe, Marc Gurgand, Jaime Montaña, Christophe Muller, Umar Serajuddin, Elena Stancanelli and attendants to internal seminars at the Paris School of Economics in 2015 and the University of Oxford in March 2020. 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

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 may depend 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, to some extent, measurement error (Verbeek, 2007). Of course, 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 correlation coefficient equal to 0 or 1 for the upper and lower bounds respectively. A more recent strand of the literature follows a methodological refinement in Dang and Lanjouw (2013, 2016) that collapses these bounds of mobility into a point estimate based on a correlation coefficient estimated through pseudo-panel techniques.[2]

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

[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 (2018) use the first two moments of aggregate data to provide bounds of individual level income mobility over long periods

[2] See for instance Cruces et al. (2015), Ferreira et al. (2013)

by Dang and Lanjouw (2014), 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 influence the synthetic income 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 work thanks to a more rigorous treatment of the estimation of the correlation coefficient that explicitly relies on a plausible AR(1) specification. It also departs from previous methodologies by departing from normality assumption, which allows to for a quasi-perfect fit to both the initial and final cross-sectional distributions. Finally, 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 describes methodology used in this paper to construct synthetic panels based on AR(1) stochastic income processes, comparing it to previous work in this area. Section three present the data used to test this methodology. Section four 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 five, 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 $y_{i(t')t'}$.[3] 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). Some of them may be observed in the initial year, but their value in the terminal year is 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.

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

The 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 is its distribution in the population.

2.2 Previous approaches

In their first attempt at constructing synthetic panels, Dang et al. (2014) assume 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, not surprisingly, 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. The covariance between (log) incomes is approximated by $cov_y = \rho^{y_c} \cdot \sigma_{yt}\sigma_{yt'}$ 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:

$$Cov_y = \beta_t' Var(z)\beta_{t'} + \rho \cdot \sigma_t\sigma_{t'} \tag{3}$$

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 the residuals.

This appears as 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.

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 a 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 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, 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 be 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 exercise necessarily depends on the set of drawings. Various draws will have to be performed to compute the expected value of these measures - and, most importantly, 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 series 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 as well as the variance of the innovation term, σ_u^2, should reasonably be assumed to be identical among them.

If 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 (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:

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

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, which is only asymptotically correct[4], 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 $\overline{y}_{gt}$. As the missing variable is positively correlated with the regressor, it follows that $\hat{\delta}$ is biased upward, 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.[5] 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.[6]

[4] Assuming no difference in the sampling procedure.

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

[6] This would be less a problem with cross-sectional survey data samples typically much larger than the

2.3.2 Calibrating the distribution of the innovation terms

In theory, once an estimate of the autoregressive coefficient ρ is available, the distribution $G_{t'}^{U}(\cdot)$ 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 $\hat{\rho}$ 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 $G_{t'}^{U}(\cdot)$ of the innovation terms such that the distribution of the virtual residuals be the same as the distribution of the observed OLS residuals $\hat{\epsilon}_{i(t')t'}$ obtained with the income regression (1). Using a continuous notation, $G_{t'}^{U}$ must thus satisfy the following functional equation:

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

where $F_\tau(\cdot)$ is the cdf of the observed residuals $\hat{\epsilon}_{i(\tau)\tau}$ 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 chosen instead, based on the approximation that the distribution $G_{t'}^{U}$ is a mixture of two normal variables with parameters $\theta = (p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ where p_1 refers to the probability that the distribution is $\mathcal{N}(\mu_1, \sigma_1)$ and (1-p_1) for $\mathcal{N}(\mu_2, \sigma_2)$. These parameters may be calibrated by minimizing the square of the difference between the two sides of (11). This method turned out to give rather satisfactory results, but, of course, it is only an approximation, even though less restrictive than the normality assumption in Dang et al. (2014).[7] The

panel data sample used in the paper to test the whole synthetic panel construction procedure.

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

detail of the calibration of the distribution $G_{t'}^U$ with a mixture of two normal distributions is given in **Appendix A** of 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 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, $\hat{\epsilon}_{i(t)t}$ and $\hat{\epsilon}_{i(t')t'}$, 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.

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, θ, of the distribution of the innovation term supposed to be a mixture of two normal variables, as described in the Appendix.

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 several times to obtain the expected value and distribution of the mobility matrices and measures.

3 Construction and validation of a synthetic panel in Mexico: 2002-2005

The procedure detailed above is now applied to the construction of synthetic incomes in 2005 of households sampled in 2002. The two cross-sections are drawn from a panel survey taken in 2002 and 2005 in Mexico. But of course, 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. The procedure can be conducted either at household or individual level. We focused on households as observational units, as these tend to offer a wider perspective on wellbeing.

3.1 Data

We use the Mexican Family Life Survey (MxFLS onwards). This survey is based on a sample of households that is representative at national, regional and urban-rural level. This longitudinal database gathers information on socioeconomic indicators, migration, demographics and health indicators on the Mexican population. It is expected to track the Mexican population throughout a period of at least ten years. Due to confidentiality, information on the sample design (sampling units) is not public (see MXFLS website). 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. We used the common sample that did not attrite; this is, the set of households observed both in 2002 and 2005.

3.2 Income definition

Household income data follow the 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.[8] Total income is 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 use pseudo-panel methods, the sample was restricted to households whose head was aged between 25 and 62 years in 2002 and 28-65 years old in 2005 and with non-missing income in both years. In addition, to overcome possible adverse effects due to atypical observations four percent of outliers were discarded (two per cent in each end of the income distribution).

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 household characteristics can be used: household size, taking into account the probability of a new-born during the three-year time interval, the area of residence as summarized by the State and urban/rural. [9]

Presumably, the more time invariant attributes the better the synthetic panel approximation. However, 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 use two model specifications, each with different degrees of time invariability, to assess the sensibility of variables selection. The first specification (Model 1) uses the household head's characteristics like 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-

[8]This definition changed to introduce a multidimensional poverty approach in 2008 (CONEVAL, 2013).

[9]The problem here is how frequent migration may be, but its implication is expected to be minor during a short period. According to census data, the internal migration rate in Mexico, from 2000 to 2005, was around 2% (Chavez & Wanner, 2012).

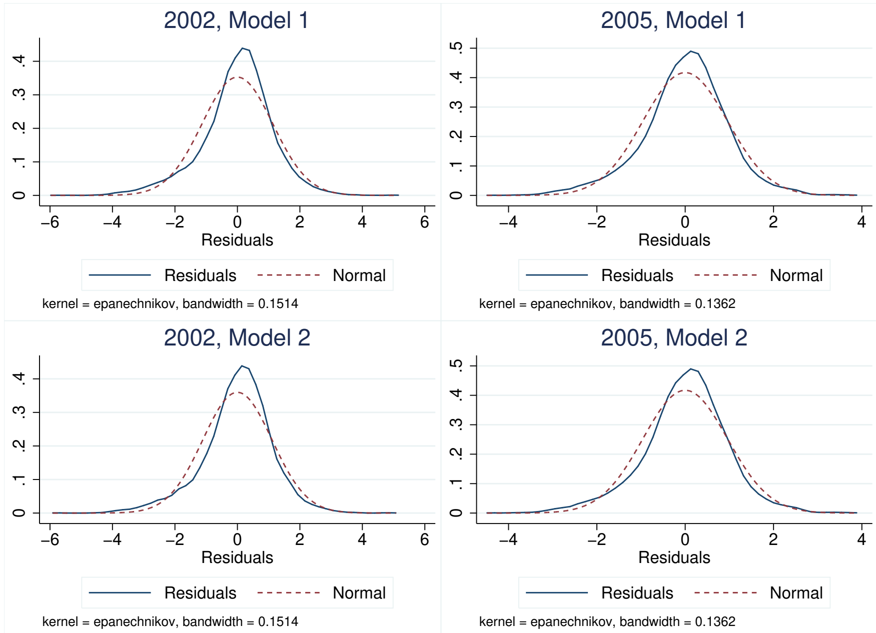
born in this three-year period. Although a bit less invariant, it also includes the area of residence (urban/rural), marital status and regions (northeast, west, centre, northwest, and south-southeast). An alternative specification (Model 2) includes long-lasting productive assets such as real estate and farming assets (land for agricultural production and cattle), and household dwelling as well as the possession of other dwellings other than the one in use. **Appendix B** show 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. It 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 these income model specifications due to high prevalence of missing data or extremely low frequency of non-modal values. We did not observe statistical differences across these two years on most of the variables actually used except for the dummy variables indicating the presence of children below three years old and long lasting assets (farming assets and dwellings property). Because of this, Model 1 is our most 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 the **Figure 1** shows the kernel distribution of (log) income residuals in both years, and compares it with the normal distribution. These figures and the Skewness and Kurtosis tests, along with the Shapiro-Wilk normality test, confirm that the normality assumption in the distribution of residuals is strongly rejected.[10]

[10]The Skewness & 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.

Figure 1: **Income models' residuals: kernel density by year and model**



3.4 The autocorrelation coefficient and calibration parameters

Estimating the autocorrelation coefficient is a central, and a crucial 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.[11] For instance a typical group would comprise households whose head was born between 1974 and 1978 with incomplete primary were assigned to one of these groups.

We then computed equations 8 and 9 with the resulting pseudo panel. The estimated genuine AR (1) coefficient, around 0.25 serves here as the benchmark. Regardless of the equation being used, the estimates in **Table 1** have the expected signs and order of magnitude. However, the combined use of these two approaches, through a non-linear equation system,

[11] Other studies working with pseudo panel methods use age interactions with other characteristics like manual or non-manual worker 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 one thousand observations. Antman & Mackenzie (2007b) and Antman & Mackenzie (2007) used 100 observations as a reference. In our case the vast majority of the groups possess no less than one hundred observations.

delivers a more accurate estimate, of which confidence interval is, not surprisingly, substantially broader than, but also fully consistent with that estimated for 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.

If the estimation of the correlation coefficient through pseudo-panel techniques is rather imprecise, 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**. As a matter of fact, it is well-known that measurement errors imply a downward bias on it. Measurement errors are not a problem in the pseudo-panel approach since they are averaged out when considering groups of households. The price to pay, however, is less precision.

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***
	(-0.05 - 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)	(-0.10 - 0.42)	(0.15 - 0.45)	(0.203 - 0.249)

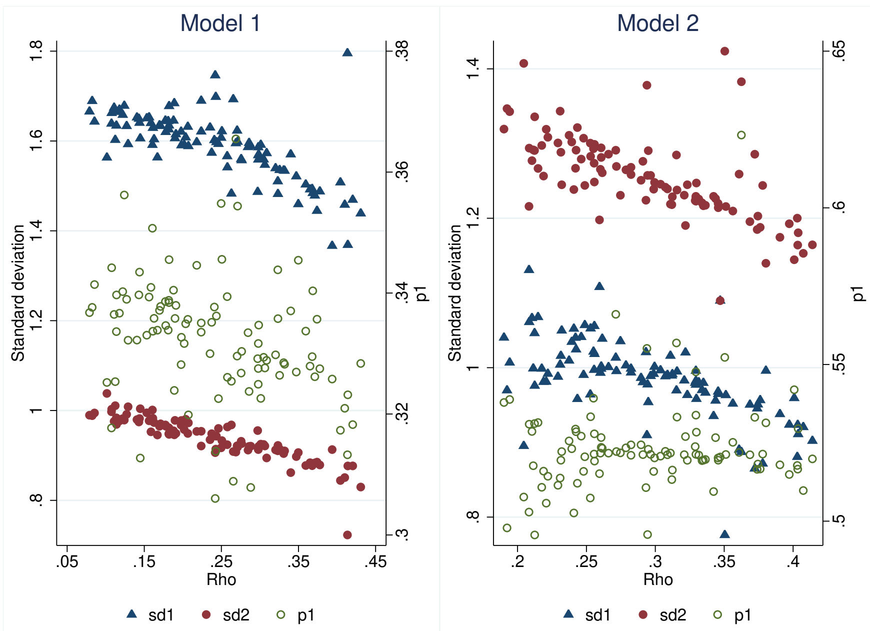
Note: *** p<0.01, ** p<0.05, * p<0.1. Conf. Interval in parentheses. GLS estimates controlling for time invariant variables. Each estimate represents the coefficient from a different regression.

The 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 done according to two different frameworks or 'regimes'. Regime 1 uses the point estimate of $\hat{\rho}$, in Table 1, to obtain a unique set of parameters $\theta(\hat{\rho}) = (p_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ of the distribution of the innovation term $G_{t'}^{U}$ according to the procedure described in Appendix. Then a value of this innovation terms is drawn from that distribution for every observation in the initial year to get a synthetic panel. However, because the random drawing is introducing noise, the procedure is repeated 500 times and mean values of mobility measures are shown, together with their 95% confidence intervals. The calibration parameters for model 1 are $\theta(\hat{\rho} = 0.25) = (p_1 = 0.33, \mu_1 = 0.007, \sigma_1 = 1.5, \mu_2 = -0.003, \sigma_2 = 0.94)$. [12]

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

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 system of equations 8 and 9 in table 1. We then use these coefficients as in regime one except that we repeat the procedure 50 times, rather than 500, for each one of the correlation coefficients. The mean value of mobility measures with regime 2 are therefore obtained from 5,000 repetitions (50X100). **Figure 2** shows a graphic description of the resulting parameters for each model.

Figure 2: **Distribution of the calibration parameters conditional on rho**

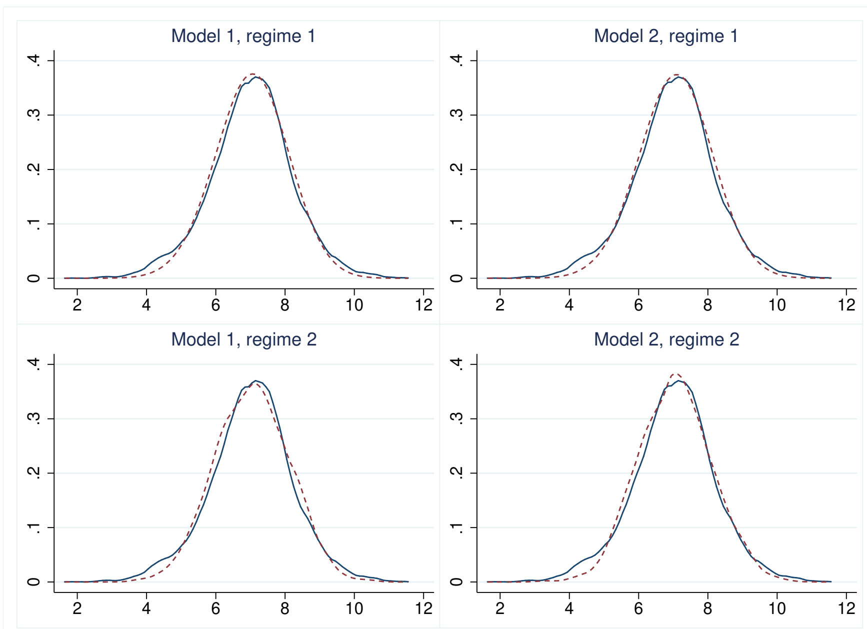


The mean of mobility indicators is not expected to be very different between these regimes but we do expect some difference in their precision particularly in regime 2 that reflects the imprecision of the estimate of rho. This is to be observed in the 5%-95% confidence intervals reported in the tables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

3.5 Synthetic panel results

We now examine income mobility with a synthetic panel and compare the results with the genuine panel. In doing so, it must be borne in mind that some growth took place in Mexico between 2002 and 2005, which affects some of the mobility measures to be analysed.[13] We first compare the shape of the (log) synthetic distribution, for each model and regime, with the genuine (log) income distribution in 2005. **Figure 3** shows both the kernel density of the genuine and virtual income. The synthetic income refers to the mean density from all repetitions of the calibration procedure. The figure provides a first visual assessment of the fit of the synthetic estimates and shows that even a very basic model specification sensibly reproduces the shape of the actual income distribution, except for a small discrepancy in the very 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.

Figure 3: **Genuine and synthetic income in the terminal year by model and regime (genuine in solid line)**



We then examine the transition matrix associated with the synthetic panel when using model 1 with both regimes. This transition matrix is defined in absolute (real income brackets) rather than relative (quintile) terms. The income brackets are defined by the limits of the 2002 income quintiles. The marginal distribution for 2002 thus exhibits 20% of the population in each bracket, but there are less people in the bottom bracket in 2005 and more in the upper brackets, reflecting the effects of economic growth. We plot three transition matrixes in a single table to facilitate the comparison of both regimes with the actual panel.

[13]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.

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 (results for model 2 in **Appendix**).

The synthetic transition probabilities for regime 1 appear close to the genuine ones in the sense that confidence intervals most often contain the observed frequency - i.e. 15 cases out of 25- (indicated by an asterisk), and do substantially overlap - i.e. 24 cases out of 25. As expected, working with a wider set of rho values, i.e. regime 2, tend to deliver slightly larger, although not always, confidence intervals.

We also used the Mann-Whitney test to evaluate the goodness of fit between the synthetic and the genuine 2005 distribution of income conditional on the ventile of origin. Actually, this test is equivalent to comparing the confidence intervals in the synthetic and genuine transition matrix in Table 2 on each cell but using ventiles, rather than quintiles, for the 2002 income to increase the sensitivity of this test. [14]

Figure 4 summarizes these results for model 1 and regime 1. The graph displays the share of drawings, among the 500 performed, passing the test for ρ=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 practically all ventiles of the baseline distribution, the exception being 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 ρ=0 and ρ=1, the arbitrary values used by Dang and Lanjouw (2014). Not surprisingly, results are much poorer with these extreme values. Assuming a perfect correlation of residuals in both the initial and terminal years delivers the poorest performance.

These results highlight the utmost importance of the value of ρ for the construction of synthetic estimates. This is not surprising. At the same time, however, it is noticeable that the difference of fit between the end values of the confidence interval of the pseudo-panel estimate, i.e. ρ=0.15 and ρ=0.45, in the **Appendix**, is quite limited, except at the extreme deciles of the initial distribution.

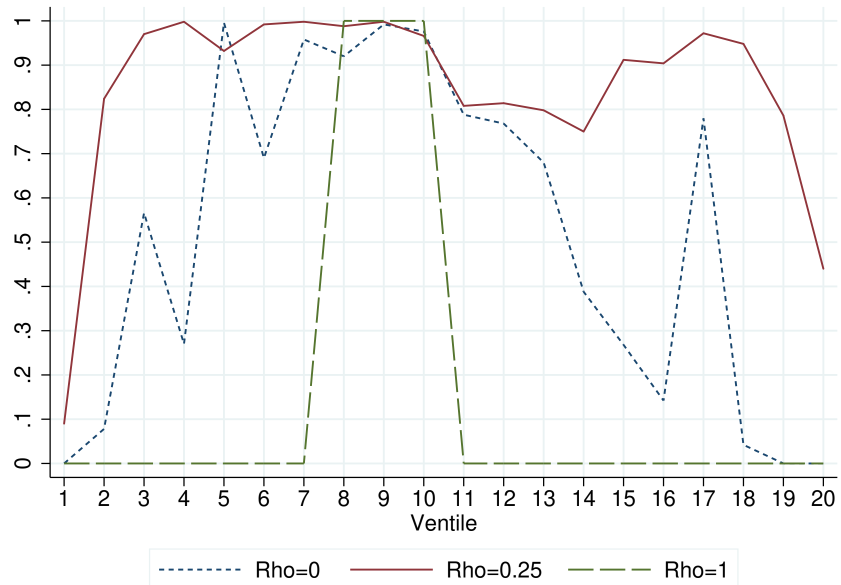
[14] This test utilizes information regarding the rank order and constitutes an alternative for the two-sample t-test of independent samples (Kirk, 2008).

Table 2: **Transition matrix, 2002-2005**

		2005 groups (Destination)					
		1	2	3	4	5	Total
A. Synthetic panel (Regime 1)							
	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
2002		(2.9-3.8)*	(5.3-6.6)*	(4.4-5.6)*	(3.5-4.7)*	(1.4-2.1)*	
Quintiles	3	1.7	4.7	5.0	5.4	3.1	20
(Origin)		(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)*	
B. Authentic panel							
	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
2002		(3.1-4.5)	(4.8-6.5)	(4.3-5.8)	(3.2-4.9)	(1.1-1.9)	
Quintiles	3	2.6	4.1	5.6	5.7	2.0	20
(Origin)		(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)	
C. Synthetic panel (Regime 2)							
	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
2002		(3.5-3.9)*	(5.9-6.9)	(3.9-4.9)	(3.4-4.3)*	(1.4-2)*	
Quintiles	3	1.6	5.0	5.0	5.2	3.3	20
(Origin)		(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: Using Model 1. 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 (in parentheses) for the synthetic estimates were obtained from 500 drawings for regime 1, and 5,000 drawings for regime 2.

Figure 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)**



Poverty dynamics is perhaps the main empirical application of synthetic panels so far. We computed two sets of poverty transitions, in-and-out of poverty, based on the upper limits of the first two income quintiles in a direct reference to the 'shared prosperity' goal adopted by the World Bank. As before, the calculation is made for alternative values of ρ, using model 1 with both regimes. **Table 3** shows that the estimation of persistent poverty, i.e. being poor in both periods, using the first poverty line is very close to the actual figure: 6.7% in the synthetic panel rather than 6.3% (regime 1) in the genuine panel -for the central value of ρ. It is also very close - three percent larger - with the second poverty line. In both cases, there is a significant overlap between the genuine and the synthetic confidence intervals and differences are not statistically significant in most cases (indicated by an asterisk). Larger differences are found for downward mobility, from non-poor to poor with the first poverty line only. On the other hand, the table confirms the extreme sensitivity of poverty mobility estimates based on synthetic panels to the value of the autoregressive coefficient -as before, the discrepancy between the synthetic estimates and the genuine figure increases with the value of ρ.

Table 3: **Poverty dynamics with alternative poverty lines and ρ values**

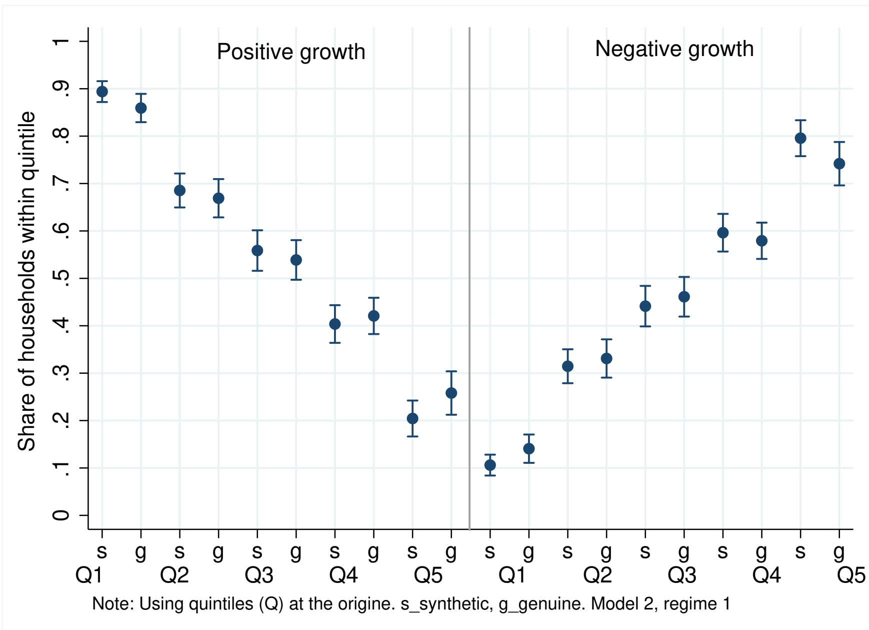
Model 1, regimes 1 and 2 with ρ=0.25

	Genuine ρ (1)	ρ=0 (2)	Regime 1 (3)	Regime 2 (4)	ρ=1 (5)
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.

We also use a simple measure of absolute income mobility -the fractions of households with higher and lower income in the terminal year- as a final robustness check. **Figure 5** shows the share of households with positive and negative income growth for model 1 with regime 1. The total share of households across these categories adds up 100%. Both the actual and synthetic panels show a clear pattern of progressive growth incidence where the poorest groups concentrated the largest growth gains while the richest groups assembled the largest losses. Differences in these absolute mobility measures are essentially non-significant. For instance, both the synthetic and the genuine figures show that around 90% of households in the poorest quintile experienced a positive rate, while the remaining 10% experienced a negative income growth.

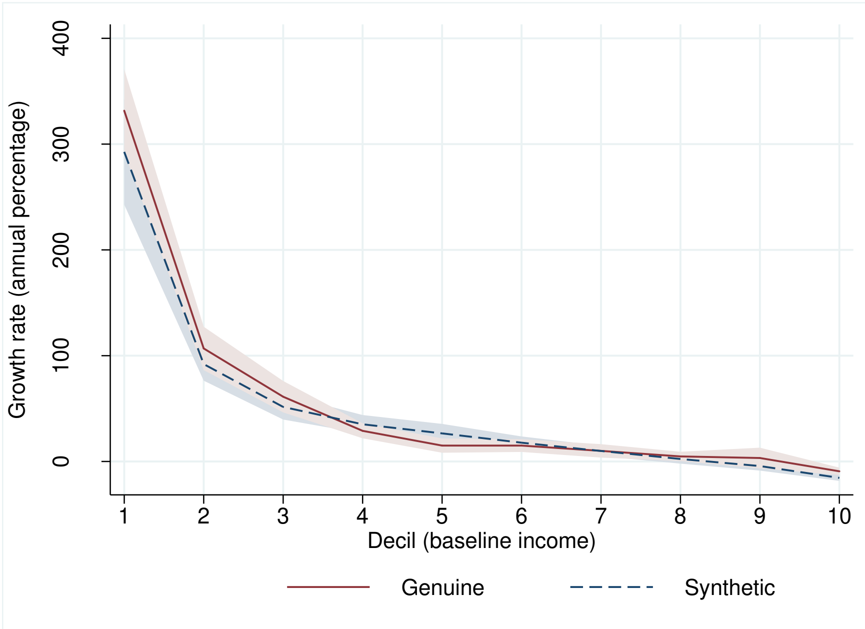
Figure 5: **Absolute Mobility (Model 1, Regime 1)**



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. **Figure 6** employs deciles of the genuine and synthetic income with their corresponding 95% confidence intervals using model 1 and regime 1. [15] These downward sloping NAGIC charts are remarkably similar in terms of their level and shape confirming a pattern of progressive growth. Once more, differences in these estimates are non-significant given that in most cases the genuine estimates fall within the synthetic 95% confidence intervals with an ample overlap between confidence intervals everywhere.

[15] 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 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.

Figure 6: **Non-Anonymous Growth Incidence Curve (Model 1, Regime 1)**



The interpretation of the latter results is not completely clear. Do they mean that income inequality was very much reduced in Mexico during the 2002-2005 period (which is actually the case), or that measurement errors introduce a mean reverting bias in the synthetic panel? As seen before, measurement errors should play a lesser role in the synthetic panel, although there are still present in the distribution of observed residuals of the income models. At the same time, given that the point estimate of the correlation coefficient of residuals is the same as that in the actual panel, not much difference is to be expected between the synthetic and actual income panel.

3.6 Concluding remarks

This paper proposed a methodological improvement in the construction of synthetic income panels based on repeated cross-sections. We performed an empirical validation by using two consecutive cross-sections of income based on a genuine panel survey in Mexico. Income mobility measures proved roughly similar in the synthetic and genuine panels and most often not statistically different. Yet, the validation also showed the extreme sensitivity of particular income mobility measures to the value of the auto-regressive coefficient used in modelling the effect the non-observed determinants of (log) income. An original pseudo-panel method developed in this paper yielded an estimate of that coefficient which is theoretically unbiased but with a rather broad interval of confidence. Under these conditions, 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. Yet, the size of the resulting confidence intervals for specific income mobility measures, including mobility in and out of poverty, prove to be substantial in the case of the Mexican income panel. It remains to be seen whether this would still be the case with other panels and in other countries.

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 F. 2010. "Non-anonymous growth incidence curves, income mobility and social welfare dominance: a theoretical framework with an application to the global economy". March 2010. Paris School of Economics. Working Paper no. 14.

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, Hai-Anh, Lanjouw, Peter. 2013. "Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections", June. The World Bank. Policy Research Paper, 6504.

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, Hai-Anh, Lanjouw, Peter. 2016. "Measuring Poverty Dynamics with Synthetic Panels

Based on Repeated Cross-Sections". Papers LACEA 2016.

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.

Deaton, Angus. 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.

Kraay, Aart C.; Van Der Weide, Roy. 2017. Approximating income distribution dynamics using aggregate data (English). Policy Research working paper; no. WPS 8123. Washington, D.C. : World Bank Group.

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

Moreno, A. H. 2018. "Long run economic mobility". Doctoral dissertation (Paris School of Economics). Economics and Finance. Université Panthéon-Sorbonne - Paris I, 2018. English.

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 expression 11:

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

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

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

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**

Time invariant variables	2002 lnincome (1)	2002 lnincome (2)	2005 lnincome (1′)	2005 lnincome (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.

Table 3: **Transition matrix, Model 2, 2002-2005**

		2005 groups (Destination)					
		1	2	3	4	5	Total
A. Synthetic panel (Regime 1)							
2002 Quintiles (Origin)	1	7.4	6.4	3.6	2.1	0.6	20
		(6.8-8)	(5.8-7)*	(3.1-4.1)*	(1.7-2.5)	(0.4-0.8)	
	2	3.3	6.1	5.1	4.0	1.5	20
		(2.9-3.8)	(5.4-6.7)*	(4.5-5.7)*	(3.5-4.6)*	(1.1-1.9)*	
	3	1.6	4.6	5.2	5.6	3.0	20
		(1.2-2)	(4-5.2)*	(4.6-5.9)*	(4.9-6.3)*	(2.5-3.6)	
	4	0.7	3.0	4.7	6.5	5.1	20
		(0.5-1)	(2.5-3.5)*	(4.1-5.3)	(5.8-7.1)	(4.6-5.7)*	
	5	0.2	1.3	2.8	5.8	9.9	20
		(0.1-0.4)	(0.9-1.7)	(2.2-3.4)*	(5-6.6)	(9-10.7)*	
Marginal Dist.		13.2	21.3	21.4	24.0	20.1	100
		(12.1-14.3)	(19.9-22.8)*	(19.7-23)*	(22.3-25.7)*	(18.6-21.6)*	
B. Authentic panel							
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)	
C. Synthetic panel (Regime 2)							
2002 Quintiles (Origin)	1	7.7	6.0	3.5	2.4	0.4	20
		(6.5-8.5)*	(5.8-6.3)*	(3.2-3.8)*	(1.8-2.9)*	(0.2-0.9)	
	2	3.4	5.8	5.0	3.9	1.8	20
		(3.3-3.5)	(5.5-6.1)*	(4.8-5.4)*	(3.8-4.1)*	(1.4-2.2)*	
	3	1.7	5.2	5.2	4.8	3.1	20
		(1.3-2.1)	(4.9-5.5)	(4.8-5.7)	(4.6-5)	(2.8-3.3)	
	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	5.8	9.4	20
		(0.1-0.2)	(0.8-1.9)	(2.5-3.6)*	(5.5-6.2)	(8.5-10.3)*	
Marginal Dist.		13.9	21.6	21.9	22.9	19.8	100
		(13.3-14.4)	(20.8-22.3)	(21.4-22.5)	(22.2-23.5)	(19.5-20)	

Notes: Using Model 2. 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 (in parentheses) for the synthetic estimates were obtained from 500 drawings for regime 1, and 5,000 drawings for regime 2.

Table 4: **2005 rank test: Synthetic Vs. Genuine conditional on the baseline rank (Model 1, Regime 1)**

2002	Mean of z_i					Share of samples that pass the test				
ventil	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 $z < z_{95\%}$.