Constructing Synthetic Panels for the Purpose of Studying Poverty Dynamics: A Primer

David Garcés-Urzainqui,
Vrije Universiteit Amsterdam (VU), The Netherlands
Peter Lanjouw,
Corresponding author, Vrije Universiteit, Amsterdam
Gerton Rongen
Vrije Universiteit Amsterdam (VU), The Netherlands

October 2021
About DEEP

Our mission is to build evidence, insights, and solutions that help end extreme poverty globally.

We aim to contribute to new global and national data and evidence that governments, decision makers, citizens and researchers can use to improve people’s lives and support the world’s poorest people in their efforts to escape extreme poverty.

We are a consortium of the Universities of Cornell, Copenhagen, and Southampton led by Oxford Policy Management, in partnership with the World Bank’s Development Data Group and funded by the UK Foreign, Commonwealth & Development Office.

Acknowledgments

This paper is part of the Oxford Policy Management (OPM) DEEP project undertaken on behalf of the Foreign, Commonwealth and Development Office of the UK. We are grateful to Simon Hunt, Stevan Lee, Umar Salam and Lucy Scott of OPM for their guidance and advice. We are further grateful to Hai-anh Dang, Chris Elbers, Peter Fisker and Finn Tarp for helpful discussions. All errors are our own.


Available at: https://doi.org/10.55158/DEEPWP3

Corresponding author: Peter Lanjouw, Vrije Universiteit, Amsterdam. Email: p.f.lanjouw@vu.nl

Follow the programme here:
Abstract

We describe a recently developed approach for constructing synthetic panels from cross-section data and we consider how it can be employed to study poverty dynamics. Initially introduced as a means of estimating upper and lower bounds on poverty transitions in the absence of panel data, further refinements to the method aim to permit also the calculation of point estimates. We describe the assumptions that underpin the basic approach and its extensions, and we discuss their plausibility. We chart applications of the method in various contexts, with a view to gauging its overall validity and robustness. While proper panel-based analysis of welfare dynamics is clearly preferable, we suggest that the method described here can be useful in the all-too-common situation where panel data are unavailable or suffer from particularly pressing quality concerns.
Table of contents

Abstract ...................................................................................................................................................... i
1 Introduction ............................................................................................................................................... 1
2 Synthetic panel method: the basics and bound estimates ................................................................. 4
3 Synthetic panel method: point estimates ......................................................................................... 8
4 Validation evidence ............................................................................................................................. 11
5 Concluding remarks .......................................................................................................................... 13
References ............................................................................................................................................... 15
1 Introduction

Success in the fight against poverty requires not just knowledge of the extent of poverty, and its possible evolution, but also an appreciation of the nuanced dynamics underlying poverty and income mobility. The ability to trace movements by individuals and households into and out of poverty is recognised to be essential for the design of policies and interventions that aim to tackle poverty. For example, when poverty is largely chronic, in the sense of the poor being stuck in poverty over long periods, this calls for policy measures that are aimed at the structural causes of poverty, which are likely to be quite different from the safety net-oriented policies that are often warranted where there is frequent movement into and out of poverty.

Panel data are traditionally employed to investigate dynamics of this kind. However, such data are rare, particularly in developing countries, because their collection can be complex and costly. As a result, there are important gaps in our knowledge and understanding of welfare dynamics in many developing country settings.

Non-availability of panel data has prompted a variety of efforts to develop pseudo-panels (or synthetic panels) from repeated cross-sectional data (see, for example, Deaton, 1985). Recent decades have seen notable improvements in the availability, quality and comparability of such cross-sectional ‘snap-shot’ surveys throughout the developing world. However, the original focus of these pseudo-panel efforts typically centred around scrutiny of cohort-level aggregates, and as such they were not particularly well-suited to the study of poverty dynamics.

A study by Dang, Lanjouw, Luoto and McKenzie (2014) – hereafter DLLM – introduced an approach for constructing synthetic panels from cross-sectional data that can provide lower-bound and upper-bound estimates of poverty transitions. Their approach involves estimating an income (or consumption) model for a given time period, including only time-invariant characteristics as regressors, and then imputing incomes at the household level into a comparable cross-sectional survey that pertains to a different time period. In their paper, the authors compare findings based on this method with those from actual panel data in Vietnam and Indonesia, and find that the bounds produced with the synthetic panels do indeed typically sandwich the ‘true’ estimates of poverty mobility. The paper also shows, however, that unless the cross-section data are particularly rich in regard to time-invariant household and individual characteristics, the bounds may be rather wide – thereby limiting inferences regarding the extent and degree of poverty mobility.

Subsequent research has explored the feasibility of producing point estimates of poverty transitions, rather than simply bounds (DLLM; Dang and Lanjouw, 2013, 2021; Elbers, 2021; Bourguignon, Dang and Moreno, 2021). The added precision comes at the cost of additional assumptions, not all of which can be easily tested. First of all, DLLM produce a parametric variant of their basic method, in which they assume bivariate normality of the error terms in the underlying income models. Within this framework, one can generate point estimates of
poverty transitions contingent on specifying a particular intertemporal correlation of the error terms. DLLM suggest that separate panel data evidence might be scrutinised to help propose a plausible correlation of errors. Dang and Lanjouw (2013, 2021) show that it is sufficient to merely observe the intertemporal correlation of incomes or consumption. However, because it remains the case that specifying the intertemporal correlation of incomes, or of the error terms from income models, requires additional, separate, panel data evidence, progressing from bounds to point estimates is rarely straightforward.

Dang and Lanjouw (2013, 2021) consider the possibility of proxying the relevant intertemporal correlations based only on the observed correlation of income at the cohort level within the available cross-section surveys. This approach has been questioned by Elbers (2021), who suggests that the underlying assumptions upon which it is contingent are strong, and not obviously plausible. Similarly, Bourguignon and Moreno (2015) and Bourguignon, Moreno and Dang (2021) propose looking at both the first and second moments of the model residuals across cohorts as a means of extracting sufficient insight into the intertemporal association of incomes at the household level to support estimation of point estimates of poverty dynamics. Further research is warranted and is ongoing.¹

In this paper we summarise the synthetic panel methodology introduced in DLLM and in Dang and Lanjouw (2013, 2021). We draw on these, and additional, papers to present some early evidence validating the method in a variety of settings. We also discuss caveats that have emerged from several independent assessments of the approach. For example, Hérault and Jenkins (2019) and Garcés-Urzainqui (2017) document examples of cases where strict statistical criteria are not satisfied when the point estimate approach of Dang and Lanjouw (2013, 2021) is implemented in the absence of support from ancillary data. In these two studies, the broad, qualitative findings on poverty dynamics were still found to be largely valid. However, they were predicated on cohort definitions that had been separately determined to be particularly successful. In more general, practical, applications there is little guidance regarding how best to define cohorts. The literature on the synthetic panel methodology is still very young, and much of it is not yet published; clearly, the dust has not yet settled on the various approaches. This primer provides an overview of the current state of the debate, with the explicit objective of also stimulating further scrutiny.

The plan of this paper is as follows. We first present the initial idea of the synthetic panel approach, as presented in DLLM, in the next section. We then describe, in Section 3, the parametric extension to point estimates discussed in DLLM and in Dang and Lanjouw (2013, 2021). We describe the additional assumptions imposed by this extension, as well as their plausibility. In Section 4 we report on several validation studies that assess the performance of the DLLM approach, showing that the bounds estimates do generally sandwich the true panel-based poverty transitions in the settings considered. We indicate that corroboration of

¹ See also Garcés-Urzainqui (2017) and Lucchetti, Corral, Ham and Garriga (2018).
key parameters from ancillary data is desirable in order to proceed to plausible point estimates of poverty transitions. We note that the papers assembled in this symposium journal issue report such point estimates, and that most, though not all, do draw on available ancillary data of some kind. Final remarks are offered in Section 5.
2 Synthetic panel method: the basics and bound estimates

In this section we offer a summary of the method introduced in DLLM, referring the reader interested in additional details to that paper. The summary presented here is an extended version of that provided in Rongen et al. (2021).

Central to the analysis of poverty dynamics are the joint and conditional probabilities of the four possible poverty transitions, which are often represented in a 2x2 transition matrix. For example, we want to estimate the likelihood of chronic poverty, i.e. that an individual is poor in both survey round 1 and round 2. This is a joint probability which we can write as follows, denoting per capita income by $y_i$, the poverty line by $z$ and the respective round by subscript 1 or 2:

$$ P(y_{i1} \leq z_1 \text{ and } y_{i2} > z_2). $$

The remaining three joint probabilities are specified in a similar way. We may also wish to estimate conditional probabilities. For instance, we wish to estimate the likelihood of poverty exit, conditional on being poor in round 1. This conditional probability can be written as:

$$ P(y_{i2} > z_2 | y_{i1} \leq z_1). $$

As a rule, we do not observe incomes in round 1 and round 2 for the same unit $i$ across rounds of a cross-sectional survey. Hence, we would normally not be able to estimate poverty transition quantities. The synthetic panel method, however, draws on time-invariant household characteristics in cross-sectional data to produce upper and lower bounds for each of the four poverty mobility outcomes. This results in estimates of the four joint probabilities, after which the corresponding conditional probabilities can easily be computed.

We model the logarithm of per capita income in each round as a linear function of time-invariant household characteristics and an individual error term. Schematically:

$$ y_{i1} = \beta'_1 x_{i1} + \epsilon_{i1} $$

and

$$ y_{i2} = \beta'_2 x_{i2} + \epsilon_{i2}. $$

---

2 Studies of poverty dynamics are in principle interested not only in poverty transitions but also in analysing the contribution of uncertainty and risk to these outcomes. For example, Ligon and Schechter (2003) introduce a panel data based-method to separate out the impact on welfare of poverty and of different sources of uncertainty.
The basic idea of the method is to use the vector of time-invariant household characteristics $x_i$ to predict household per capita income for the survey round in which the household was not interviewed. By their nature, such characteristics remain unchanged across the two rounds. They will mostly be characteristics of the household head, including their year of birth, educational attainment, ethnicity or religion, and birth district, depending on how rich the questionnaire is. In addition, if the survey includes retrospective questions, these could also be exploited: for example, if the household owned a car or television at the time of the previous survey.

As we have cross-sectional data for each round, we are able to estimate the two coefficient vectors $\beta_1$ and $\beta_2$ on the selected time-invariant explanatory variables. By definition, it holds that $x_{i1} = x_{i2}$, which is why we can predict the round 1 income for households in the round 2 survey by applying the coefficient vector resulting from the round 1 regression, $\hat{\beta}_1$, to the characteristics of households in round 2, $x_{i2}$, as follows:

$$\hat{y}_{i1} = \beta_1' x_{i2} + \hat{\epsilon}_{i1}.$$ 

Before we discuss how estimates for the error term are obtained, we first turn to the assumptions underlying the approach. We will see that upper and lower bounds are derived from the two boundary cases of our assumption regarding the relation between the error terms in our model.

The DLLM method for estimating bounds on poverty transitions is predicated on two important assumptions. First, we need to assume that the underlying population from which the data are sampled is the same in both survey rounds. This implies that the method cannot be applied in cases where there are large shocks to the population: for example, if there was massive migration into or out of the country. It follows that estimates based on long intervals between survey rounds should be interpreted more cautiously. Moreover, to help ensure the stability of the reference population, the sample is typically limited to households whose heads are in a certain age range—say between 25 and 55 years of age—in round 1. This range needs to be chosen in such a way that the formation of new households and the dissolution of existing households is at a minimum. Moreover, the lower age limit should be set such that heads will generally have attained their maximum level of education. The age range for the round 2 sample needs to be shifted accordingly.

Second, assumptions are required regarding the relationship between the error terms $\epsilon_{i1}$ and $\epsilon_{i2}$. Specifically, it is assumed that these are on average positively correlated over time. In the presence of household fixed effects and the persistence of shocks, this seems to be a

---

3 Alternatively, predictions can be made forwards, predicting income in round 2 for households in the round 1 survey. Sensitivity analysis by DLLM shows that in their case results were equivalent.
4 Strictly, we need to assume that the $t=1$ regression results are applicable to predict $t=1$ incomes for households sampled in the $t=2$ survey.
reasonable assumption. We do not make further assumptions about the shape of the distribution of the error terms; in that sense, these bound estimates are non-parametric.

The upper- and lower-bound estimates for poverty transitions follow from two extreme cases of the assumed relationship between the error terms: no correlation whatsoever ($\rho = 0$); and perfect positive correlation ($\rho = 1$). Note that DLLM define the synthetic panel bounds in reference to mobility: the upper-bound estimates indicate maximum mobility. Hence, these correspond to higher probabilities of the poor–non-poor and non-poor–poor transitions, and correspondingly lower probabilities for the immobile categories (poor–poor and non-poor–non-poor). In the following, upper-bound estimate and lower-bound estimate refer to upper and lower bounds on mobility.

At one extreme, we consider the case of $\rho = 0$, which yields the upper-bound estimates. In this scenario, the errors are uncorrelated and mobility will be at a maximum, both for movements out of poverty and into it. The upper bound will be strictly larger than actual mobility, because we assume that in practice errors are positively correlated. The predicted upper-bound round 1 income for household $i$ surveyed in round 2 will be:

$$\tilde{y}_{12}^{2U} = \beta_1^* x_{i2} + \tilde{\epsilon}_{i1},$$

where $\tilde{\epsilon}_{i1}$ is a residual randomly drawn from the actual distribution of residuals resulting from estimating the income model on the round 1 observations. Evaluation of these income predictions against the poverty line yields a predicted poverty status in round 1, which then determines the transition category, given that round 2 poverty status is observed.

Aggregating these quantities at the population level, we obtain the country-level poverty dynamics. Due to the random nature of this process, these steps are repeated $R$ times, say 500 times, and averaged to obtain a robust estimate of upper-bound poverty mobility.

At the other extreme, the lower-bound estimates assume a perfect positive correlation between the errors: $\rho = 1$. In this case, mobility will be at its minimum. For the prediction, this means we simply take the estimated round 2 residual, scaled by a factor $\gamma$, and add it to the fitted round 1 income:

$$\tilde{y}_{12}^{2L} = \beta_1^* x_{i2} + \gamma \tilde{\epsilon}_{i2}.$$

---

5 Note that this is but one way of considering maximum mobility; the literature describes many different concepts. See Shorrocks (1978) and Fields (2008) for further discussion.

6 We use a scaling factor $\gamma = \widehat{\sigma}_{i1} / \widehat{\sigma}_{i2}$ to adjust for the difference in error variance between the two rounds ($\widehat{\sigma}_e$ is the standard error of the residuals).
After aggregation, but without averaging over repetitions, we obtain the lower-bound population-level poverty dynamics estimates. These quantities can also be aggregated at the level of subgroup, for example by geographic regions.

7 Repetition is not necessary in this case because adding the scaled round 2 residual for the same unit is not a stochastic process.
3 Synthetic panel method: point estimates

DLLM, and Dang and Lanjouw (2013, 2021), explore the feasibility of producing point estimates of poverty mobility. They start with the more strict assumption that the regression errors $\varepsilon_1$ and $\varepsilon_2$ follow a bivariate normal distribution.\(^8\) We can then express the likelihood of, for example, chronic poverty as follows:

$$P(y_{i1} < z_1 \text{ and } y_{i2} < z_2) = \phi_2\left(\frac{z_1 - \beta_1' x_{i2}}{\sigma_{\varepsilon_1}}, \frac{z_2 - \beta_2' x_{i2}}{\sigma_{\varepsilon_2}}, \rho\right)$$

where $\phi_2(\cdot)$ indicates the cumulative density function of the bivariate normal distribution; standard deviations of the errors are denoted $\sigma_{\varepsilon_1}$ and $\sigma_{\varepsilon_2}$, and their correlation coefficient again as $\rho$.\(^9\) This last parameter is a critical element in the estimation procedure, which cannot be estimated directly nor validated in the cross-section.

DLLM suggest that in some settings ancillary panel data, fielded at a different time or involving a different survey organisation, might be available. If so, these could be scrutinised to reveal a plausible value for $\rho$. Dang and Lanjouw (2013, 2021) show, on the basis of Monte Carlo evidence, that final point estimates are rather stable in the face of moderate departures from the true $\rho$ in the estimation procedure. Dang and Lanjouw also show that it is sufficient to observe the intertemporal correlation of income or consumption in the ancillary data – removing the need to model income or consumption in those ancillary data. Proposition 1 in their paper provides an expression from which the intertemporal correlation of errors can be calculated as a function of the correlation of incomes and of the estimated parameters from the income models in the cross-sectional surveys:

$$\rho = \frac{\rho_{y_{i1},y_{i2}} \sqrt{\text{var}(y_{i1}) \text{var}(y_{i2}) - \beta_1' \text{var}(x_i) \beta_2}}{\sigma_{\varepsilon_1} \sigma_{\varepsilon_2}},$$

where $\rho_{y_{i1},y_{i2}}$ is the simple intertemporal correlation coefficient of incomes. Given this $\rho$, the likelihood of all four poverty transitions can be estimated for each household. Aggregation to the population or subgroup level then provides overall probabilities.

Since the intertemporal correlation of incomes cannot be directly observed without actual panel data, Dang and Lanjouw (2013, 2021) explore the feasibility of proxying $\rho_{y_{i1},y_{i2}}$ by working with the correlation between cohort-level average incomes across the two rounds. Calculating the intertemporal correlation of incomes at the cohort level is clearly possible across the two cross-sectional surveys, even though the sampled households differ.

---

\(^8\) Elbers (2021) warns that in the likely event that the residuals reflect omitted time-invariant variables homoscedasticity is unlikely to hold. He outlines an approach that takes this into account explicitly (see further below).

\(^9\) We refer to Dang and Lanjouw (2013, 2021) for the formulae for the other transition categories.
Proceeding in this way is appealing as it means the analyst does not have to extract the correlation from ancillary data. It has been argued, however, that the cohort-level correlation cannot serve as a consistent proxy of the correlation across households because it does not contain any information on within-cohort correlation of errors (Elbers, 2021). Interestingly, several validation exercises find that the Dang and Lanjouw cohort-level correlation approach does produce estimates of $\rho$ that are quite close to their true value, and thus point estimates of mobility that are correspondingly accurate. But this is not universally the case. Hérault and Jenkins (2020), in their analysis of validation data from the UK and Australia, report estimates of $\rho$ that in some cases diverge markedly from true values, and show in particular that they are sensitive to the cohort definition being used (see also Garcés-Urzainqui (2017), who links this lack of robustness to the role of cohorts as instrumental variables). They note further that there is little guidance available regarding how best to define cohorts.

In the absence of any kind of validation data against which to check, caution is warranted with respect to point estimates derived using the Dang and Lanjouw (2013, 2021) cohort-based approach. It would seem advisable in this case to avoid excessive reliance on the precise value of these point estimates, and to emphasise rather those empirical findings that are robust to the dynamics captured by the upper and lower bounds.10

**Moving beyond cohort-level averages**

Certain alternatives to the estimation of this key parameter have been put forward. In essence, they all explore variations of the original idea of aggregating over cohorts that was introduced in Dang and Lanjouw (2013, 2021). Bourguignon, Moreno and Dang (2021) propose a procedure based on two equations, jointly estimated in a non-linear equation system. First, they argue that cohort averages are best employed to estimate the AR(1)-coefficient delta of consumption in a model that includes the time-invariant covariates used to predict consumption in the first place. This avoids the indirect procedure in Dang and Lanjouw (2013, 2021) that requires proxying the unconditional correlation of consumption – which they argue will deliver biased estimates. Second, they introduce the idea of working with cohort-level variances, and in particular the within-cohort variance of residuals from the income prediction models. They argue that the Dang and Lanjouw approach will overestimate intertemporal correlation, and they present evidence to show that their approach is superior to Dang and Lanjouw, using data from Mexico.

Garcés-Urzainqui (2017, 2021) notes that both cohort averages and cohort variance approaches require a particular structure. He suggests working with the means equation put

10 Dang and Lanjouw (2013, 2021) provide formulae for calculating standard errors on the point estimates. These serve to further underscore that point estimates are best viewed as ‘fuzzy’. Elbers (2021) emphasises the importance of ensuring that standard errors capture not only the uncertainty about the regression residuals but also the uncertainty that is derived from the errors in parameter estimation.
forward in Bourguignon, Moreno and Dang (2021) in an explicit two-sample two-step least squares framework (TS2SLS). This estimator is subject to similar assumptions as the means equation in Bourguignon, Dang and Moreno (2021) in order to deliver consistent estimates, but has a number of theoretical and practical advantages when those assumptions hold.

Elbers (2021) prefers to identify the AR(1) parameter $\delta$ linking residuals from different periods via exclusive reliance on cohort-level variances. He suggests that any effects of cohort membership on income levels should be incorporated in the income models (Equations IM), as they can be flexibly estimated with the repeated cross-sections at hand. Further, he argues that omitted variables and non-linearities are likely to induce heteroskedasticity across cohorts, so that it may be possible to estimate the parameter $\delta$ in the Bourguignon, Moreno and Dang variance equation. For that purpose, he introduces a Generalized Methods of Moments/ Maximum Likelihood approach that can take into account cohort sizes and measurement error, improving on the ordinary least squares estimation of Bourguignon, Moreno and Dang (2021).

These papers also provide a number of improvements beyond the estimation of intertemporal correlation. Bourguignon, Moreno and Dang (2021) provide a more flexible, although computationally costly, residual calibration and simulation procedure that allows for departure from assumptions of bivariate normality and matches more closely with the synthetic income distribution. They also point out that the analysis can be extended to full mobility tables. Elbers (2021) describes a systematic implementation procedure that permits the calculation of standard error estimates that take into account both model and sampling error.

Further research is needed to assess when and to what extent the alternatives and refinements presented above should be implemented in practical applications – thereby preserving a means of retrieving the estimate of $\rho$ or $\delta$ that is needed for point estimates of poverty mobility without having to resort to parameters derived from ancillary panel data. All of these papers present some encouraging evidence, in the form of simulations or empirical data exercises, in favour of the methods they introduce. However, as approaches based on aggregation over cohorts, they are potentially subject to the lack of robustness over different cohort definitions that has emerged as the main problem for the Dang and Lanjouw (2013, 2021) method in independent validations, as we shall see below. It remains to be seen whether these innovations are able to provide more robust estimates, and at the current stage it is also yet to be seen to what extent they are able to relax the strong underlying assumptions of Dang and Lanjouw (2013, 2021).
4 Validation evidence

A number of empirical studies probe the validity of the DLLM synthetic panels approach. These validation studies start by examining actual panel data and deriving estimates of poverty transitions from them. The data are then re-analysed as though they were repeated rounds of cross-section data, and the synthetic panels method is applied. The synthetic panel estimates are then compared to the ‘true’ estimates in order to assess the validity of the method. Note that implementation of the synthetic panels method is based on sub-samples from the actual panel data, chosen in such a way as to ensure that household-level observations in each sub-sample occur just once. Table 1 illustrates findings from DLLM (2014) and Cruces et al. (2014) indicating that the true poverty estimates of poverty transitions, as calculated from actual panel data, are consistently ‘sandwiched’ by the estimates of bounds produced using the synthetic panels approach in the same data. In these validation exercises, the actual panels are first used to directly estimate poverty transitions, and then the panel structure of the surveys is deliberately ignored to produce estimates based on the synthetic panel method.

What is also clear from Table 1 is that the bounds on poverty transitions are often quite wide. This could impede the drawing of strong conclusions concerning poverty transitions. Rongen et al. (2021) show, however, that in the case of Malaysia, where multiple rounds of cross-section data are available, interesting trends in poverty dynamics are still discernible from bounds estimates. In the absence of ancillary data to provide insights into the likely value of $\rho$, the intertemporal correlation of errors, and hesitation in imposing the assumptions needed to produce point estimates, bounds-based analyses are often as far as one can go.

When point estimates are pursued based on the Dang and Lanjouw (2013, 2021) approach, strong assumptions are unavoidable. Despite this unappealing requirement, Dang and Lanjouw implement the approach and compare it to true panel results in a selection of countries. They find that the point estimates can be, at times, quite close to the truth. Table 2 reports point estimates based on the Dang and Lanjouw (2013, 2021) approach and compares these to true estimates of poverty mobility. The table shows that even when estimates of $\rho$ based on examination of the intertemporal correlation of cohort average residuals do not exactly match true values, the final poverty mobility estimates line up quite closely with the truth in Bosnia-Herzegovina, Lao PDR, Peru, the United States and Vietnam (Dang and Lanjouw, 2013, 2021).

In the most systematic validation study of the Dang and Lanjouw (2013, 2021) approach available to date, Hérault and Jenkins (2020) consider the cases of the UK and Australia over 15 and 18 wave-years, respectively. Here, estimates of $\rho$ are found to be quite sensitive to the cohort definitions employed. Hérault and Jenkins (2020) show, further, that even for the cohort definition that produces the best estimates of $\rho$ on average, the corresponding poverty mobility estimates are found, at times, to deviate significantly from their true, panel
data-based, values. The accuracy of estimates varies rather widely with the particular time interval being considered, as well as the age range of household heads included in the income model, which seems to be mainly driven by oscillations in the cohort estimate for $\rho$.\footnote{This may be partially explained by the fact that their ‘leading case’ results in few, large, cohorts, which implies larger standard errors for estimates of $\rho$.} Estimated accuracy is also found to be sensitive to the poverty line employed. It should be noted, however, that even if the synthetic panel estimates disappoint with respect to the statistical evaluation criterion employed by Hérault and Jenkins – lying within the 95% confidence interval of the rather precisely estimated panel benchmark\footnote{In turn, confidence intervals for synthetic panel estimates that would adequately reflect the underlying uncertainty might be more likely to contain the panel benchmark, especially when the estimate for $\rho$ is not too inaccurate.} – the broad, qualitative findings regarding poverty dynamics do not differ markedly, except in cases where the cohort $\rho$ is particularly far off the actual value. Overall, Hérault and Jenkins (2020) caution against relying solely on the Dang and Lanjouw (2013, 2021) point estimates approach. At a minimum they recommend employing the DLLM (2014) bounds approach – which they confirm reliably ‘sandwiches’ the true panel estimates – and attaching most emphasis to findings and results that are consistent with patterns captured by the bounds estimates. The fact that the stability of $\rho$ – the range of variation over the whole study period for both the UK and Australia – is similarly broad when compared to the typical confidence interval of an estimate based on cohort averages is an encouraging signal for the extrapolation of this parameter from ancillary data.

Hérault and Jenkins (2020) further document that the relative accuracy of poverty transition estimates varies across population subgroups, defined by age. This might be an additional aspect worth exploring in future validation work, as one of the most common applications of synthetic panel applications is the analysis of differences in poverty transitions across groups – so-called poverty profiles. Here, the homogeneity assumption, i.e. the estimation of a single model and a single intertemporal correlation for the whole population, might be an additional source of errors.

In another empirical investigation, Garcés-Urzainqui (2017) compares the performance of the approaches put forward in Dang and Lanjouw (2013, 2021) and Bourguignon and Moreno (2015) on panel data from Thailand.\footnote{Bourguignon and Moreno (2015) present a set of methods that is slightly different to Bourguignon, Moreno and Dang (2021). However, both estimating equations in Bourguignon, Moreno and Dang (2021) are part of the analysis, and none of them can outperform the Dang and Lanjouw (2013, 2021) approach in its own} That study also documents that methods to estimate $\rho$ or $\delta$ are very sensitive to the way cohorts are defined. However, the performance of the Dang and Lanjouw approach seems, if anything, superior to the alternative methods put forward in Bourguignon and Moreno (2015) – and in particular the variance equation tends to underestimate residual persistence. This validation also provides evidence in favour of the flexible residual calibration procedure in Bourguignon, Moreno and Dang (2021) with respect to the bivariate normality assumption in Dang and Lanjouw (2013, 2021).
5 Concluding remarks

Recent calls to end the ‘dark ages of inequality statistics’ have raised awareness of the negative consequences for society of a lack of information and transparency regarding the income of top earners, and regarding income inequality statistics more generally (Alvaredo et al., 2020). A similar concern can be expressed for the other end of the global income distribution. While new innovative methods of working with ‘big data’ are being eagerly explored to produce timely, detailed and accurate poverty assessments, in most of the developing world, policymakers and researchers are still confronted with a dearth of the nationally representative panel data that are the standard tool used in the study of welfare dynamics in high-income countries.

It is in this context that synthetic panel methods need to be understood, namely as a palliative to this pressing data scarcity. The short conclusion is that synthetic panel methods can generally be trusted to produce bounds that encompass the actual dynamics. On the other hand, in the absence of relevant ancillary data, the assumptions required to produce point estimates are stronger, and, in particular, results can be quite sensitive to the cohort definition used to estimate the intertemporal correlation of incomes (or income residuals).

We have outlined some alternatives to the estimation of this key parameter introduced in the literature. While some of them may come, with time, to replace or provide further robustness to the Dang and Lanjouw (2013, 2021) approach, the latter has been to date the standard method in the literature and it is also the one adopted throughout this special issue. The adoption of other improvements to abstract from the normality assumption and produce valid standard errors may also help to enhance the credibility of synthetic panel estimates.

While a cautious researcher might thus prefer to focus on the insights that result from bounds, policymakers are often more interested in point estimates due to their ease of interpretation and communication, among other reasons. In such cases, researchers should be careful to adequately convey the amount of uncertainty surrounding these estimates, and discuss the key modelling choices taken.

Taking a step back, it is unsurprising that some of the assumptions that need to be made in order to conduct dynamic analysis without longitudinal data are strong. Hopefully, it will eventually become possible to study topics such as intergenerational income mobility, global inequality in income or wealth, or poverty dynamics in developing countries, with high-quality data that minimise the need for strong assumptions. The recent COVID-19 pandemic underlines the importance of understanding the patterns, drivers and consequences of vulnerability to shocks. Wider availability of panel data in developing countries would surely contribute to a deeper understanding of the factors shaping transitions into and out of

---

poverty. In the meantime, synthetic panels can, with the necessary caveats, be a useful tool to gain some valuable insights.
References


### Table 1: Validation of the DLLM bounds estimates method

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Poverty dynamic</th>
<th>Results (%) of base-year population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Truth</td>
</tr>
<tr>
<td>Dang et al. (2014)</td>
<td>Vietnam (2006–8)</td>
<td>Chronic poverty</td>
<td>9.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>5.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>4.9%</td>
</tr>
<tr>
<td>Dang et al. (2014)</td>
<td>Indonesia (1997–2000)</td>
<td>Chronic poverty</td>
<td>7.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>10.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>8.3%</td>
</tr>
<tr>
<td>Cruces et al. (2015)</td>
<td>Chile (1996–2006)</td>
<td>Chronic poverty</td>
<td>4.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>19.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>3.0%</td>
</tr>
<tr>
<td>Cruces et al. (2015)</td>
<td>Nicaragua (2001–05)</td>
<td>Chronic poverty</td>
<td>35.7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>3.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>26.1%</td>
</tr>
<tr>
<td>Cruces et al. (2015)</td>
<td>Peru (2008–09)</td>
<td>Chronic poverty</td>
<td>23.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>10.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>10.0%</td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>Bosnia and Herzegovina (2001–04)</td>
<td>Chronic poverty</td>
<td>10.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>12.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>10.5%</td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>Lao PDR (2002–07)</td>
<td>Chronic poverty</td>
<td>13.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>14.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>10.9%</td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>United States (2007–09)</td>
<td>Chronic poverty</td>
<td>6.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty exit</td>
<td>3.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poverty entry</td>
<td>4.6%</td>
</tr>
</tbody>
</table>
Table 2: Validation and application of Dang and Lanjouw (2013, 2021) point estimate method

<table>
<thead>
<tr>
<th>Study</th>
<th>Country (years)</th>
<th>Intertemporal correlation, $\rho$</th>
<th>Poverty dynamic</th>
<th>Mobility estimate (% of base-year population)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Truth</td>
<td>Estimate</td>
<td></td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>Bosnia-Herzegovina (2001–04)</td>
<td>0.45</td>
<td>0.40</td>
<td>Chronic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>Lao PDR (2002–07)</td>
<td>0.43</td>
<td>0.46</td>
<td>Chronic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>Peru (2005–06)</td>
<td>0.66</td>
<td>0.63</td>
<td>Chronic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>United States (2007–09)</td>
<td>0.70</td>
<td>0.79</td>
<td>Chronic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dang and Lanjouw (2013, 2021)</td>
<td>Vietnam (2006–08)</td>
<td>0.63</td>
<td>0.78</td>
<td>Chronic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>