EU-SILC and the potential for synthetic panel estimates

Colgan, Brian

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Colgan, B. (2023). EU-SILC and the potential for synthetic panel estimates. *Empirical Economics*, *64*(3), 1247-1280. https://doi.org/10.1007/s00181-022-02277-7

Empirical Economics (2023) 64:1247–1280
https://doi.org/10.1007/s00181-022-02277-7

EU-SILC and the potential for synthetic panel estimates

Brian Colgan[1]

Received: 16 August 2021 / Accepted: 20 June 2022 / Published online: 6 August 2022

Abstract
In the absence of panel data, researchers have devised alternative methods for estimating synthetic poverty dynamics using repeated cross section surveys. These methods are not only salient in the absence of panel data, but also in contexts where there are concerns over the quality of panel data and/or the panel data are of insufficient length to analyse medium- to long-term mobility trends. Both of these issues afflict the longitudinal element of the European Survey on Income and Living Conditions (EU-SILC) (Hérault and Jenkins, J Econ Inequ 17(1):51–76, 2019). Using the longitudinal element of EU-SILC, this paper assesses the accuracy of the synthetic panel approach put forth by Dang and Lanjouw (2021). For most conventional poverty lines, the DL approach is found to be highly accurate when the true ρ is known. Similar to Hérault and Jenkins (J Econ Inequ 17(1):51–76, 2019) the pseudo-panel approach for estimating ρ is found to be highly sensitive to cohort definition. The longitudinal element of EU-SILC, however, offers a unique route for overcoming this shortcoming.

Keywords Synthetic panel · Pseudo-panel · Poverty · Poverty dynamics · EU-SILC

JEL Classification C53 · D31 · I32

Introduction

Panel datasets provide important insights into the nature of poverty and the household characteristics associated with poverty and its re-occurrence. Despite the benefits of panel data, large-scale projects over long time periods are typically found in only a small number of developed countries. Even where panel data are available, they can suffer from small sample sizes and high attrition rates. These issues call into question the representativeness of some panel samples and may lead to inconsistent estimates

✉ Brian Colgan
b.p.colgan@vu.nl

[1] Vrije Universiteit, Amsterdam, Netherlands

of population means (Ashenfelter et al. 1986). This paper examines a method for estimating synthetic income dynamics using cross-sectional data.

The obvious drawback of using cross-sectional data is that the same individuals are not followed over time, however, as Verbeek (2008) states:

"Repeated cross-sections suffer much less from typical panel data problems like attrition and non-response, and are very often substantially larger, both in number of individuals or households and in the time period that they span."

Deaton (1985), in the seminal paper, "Panel Data from Time Series of Cross Sections", proposed grouping individuals with common time invariant characteristics and treating the averages within these cohorts as observations in a pseudo-panel. This paper gave rise to the pseudo-panel approach and although there have been various refinements since its inception, it is still limited to the examination of inter-cohort dynamics; on the subject of intra-cohort mobility it is silent. Estimates at the cohort level may also suffer from bias if cohorts are heterogeneously effected by events such as migration and death (Fields and Viollaz 2013).

"Synthetic" panel approaches offer an alternative bridge to overcome the gap in panel data; one which does not focus solely on inter-cohort dynamics.[1] Dang et al. (2014) (henceforth DLLM) introduced a synthetic panel technique capable of constructing upper and lower bounds for poverty transitions. Similar to pseudo-panel techniques, time invariant characteristics are used to link two rounds of independent cross-sectional data. Rather than comparing cohort averages, the DLLM method estimates income models at the household level consisting of only time invariant characteristics for both rounds of cross-sectional data (t and $t + 1$). For the households observed in period $t + 1$, period t income can be predicted using the coefficients predicted from the income model for time t (or vice versa). The question then becomes how to treat the residuals from these income models and in particular how to treat the correlation between residuals over time.

DLLM propose both nonparametric and parametric approaches for treating the correlation of residuals in order to estimate bounds. The parametric approach imposes a bi-variate normal distribution on the residuals and uses either zero/perfect correlation or lower and higher correlation estimates from alternative datasets or neighbouring countries to provide lower and upper bounds. Both the nonparametric and parametric bounds have been shown to be successful at encompassing true joint and conditional poverty estimates (Dang et al. 2014; Cruces et al. 2015; Perez et al. 2015). Unfortunately, the approach can be of limited use to policy makers due to the size of the bounds (Hérault and Jenkins 2019).

Dang and Lanjouw (2021), henceforth DL, builds upon the DLLM method and incorporates pseudo-panel techniques to estimate residual correlation and arrive at point estimates of transitions. The key innovation of the DL approach is to approximate the true income correlation using the correlation between cohort level average incomes. Thus far Dang and Lanjouw (2021), Garcés Urzainqui (2017), and Hérault and Jenkins (2019), have examined the performance of the DL approach in compar-

[1] For the purposes of this paper, I consider pseudo-panel techniques to relate to cohort levels means and their comparison over time. Synthetic panels techniques are those which move beyond cohort means to try and estimate household level dynamics. The distinction, therefore, lies in the techniques ability to estimate intra-cohort dynamics.

ison to true panel data. The findings have been mixed with the method being shown to be sensitive to a number of practical considerations. HJ suggest the method may be less accurate in high income settings; however, one must consider how accuracy is measured. Typically, accuracy of synthetic panel estimates is measured by counting how many of the synthetic panel estimates lie within the 95% confidence interval of the true panel estimate. While intuitively appealing, this benchmark will vary across setting due to differences in the size and quality of the survey data used for validation. Previous validation papers and this potentially inconsistent benchmark are further discussed in Sect. 2.

Beyond validation, both the DLLM and DL approaches have been implemented in a wide range of contexts and time periods. Some prominent examples include: Ferreira et al. (2012) focus on Latin America, Dang and Ianchovichina (2018) examine income mobility and the Arab Spring, Dang and Dabalen (2019) explore the nature of poverty in Africa, and Dang and Lanjouw (2018) look at long term income mobility in India.

In a general sense it is the purpose of this paper to provide further clarity as to the accuracy and practical implementation of the DLLM approach and the DL proposition for approximating $\rho_{y_{i1}y_{i2}}$. Throughout the analysis particular attention is paid to the evaluation of synthetic panel estimates in the absence of true panel data. Validation will be carried out using EU-SILC, a pan-European survey containing both longitudinal and cross-sectional elements. The analysis in this paper will primarily focus on France, Poland, and Greece for the period 2005-2016. The scope, in terms of both countries and time, offers new insight into the performance of the DL approach in different settings and economic conditions. Furthermore, the shortcomings of EU-SILC, such as high attrition rates, short panel length and the absence of certain ad hoc modules from the longitudinal element, mean that there is great potential for the DL approach to be of practical usefulness for EU-SILC users (Hérault and Jenkins 2019).

In order to give structure to the validation, two questions are asked of the DL approach; *does it work when the true* $\rho_{y_{i1}y_{i2}}$ *is known?* If so, *can* $\rho_{y_{i1}y_{i2}}$ *be accurately approximated?*

The first question can be seen to examine two important components of the DL approach, namely the choice of income model and the bivariate normal assumption. Using the true panel correlation of the residual terms, the sensitivity of estimates with respect to the choice of income model can be examined. If, given the true correlation term, the DL approach fails to produce a good approximation of joint and conditional poverty probabilities, this suggests that the bivariate normal assumption is too strict a structural form to impose on the residual and that there is little added value in exploring approaches to approximating the correlation term. The DL approach is found to be fairly insensitive to the choice of income model. Given the true $\rho_{y_{i1}y_{i2}}$, the DL approach can produce accurate estimates of joint and conditional poverty probabilities although, much depends on the normality of the residuals. Standard techniques for addressing issues of normality can have a significant impact on the accuracy of estimates. In line with previous research, estimates are found to be less accurate the higher the poverty line is set.

The second question concerns the approximation of $\rho_{y_{i1}y_{i2}}$. The pseudo-panel approach proposed by DL for approximating $\rho_{y_{i1}y_{i2}}$ is strongly related to the more common-place pseudo-panel measure for estimating whether incomes within a coun-

try have been converging or diverging over time (Dang and Lanjouw 2021). In this paper this link is made explicit and both the DL approximate for $\rho_{y_{i1}y_{i2}}$ and the related pseudo-panel measure of convergence/divergence (δ) are estimated.

Both the approximation of $\rho_{y_{i1}y_{i2}}$ and δ are found to be highly sensitive to a number of practical decisions. Accurate approximates can be produced, but the problem faced by practitioners is how, in the absence of panel data, to select the best cohort definition.

The final contribution of this paper is to explore alternative sources of $\rho_{y_{i1}y_{i2}}$. The short longitudinal element of EU-SILC begs the question as to how to predict $\rho_{y_{i1}y_{i2}}$ into the future. This paper presents a framework for producing upper and lower bound of $\rho_{y_{i1}y_{i2}}$ as well as point estimates. Using correlation estimates provided in Ball (2016), the framework is found to be highly accurate at tracking $\rho_{y_{i1}y_{i2}}$ into the future.

This paper consists of 6 sections. Section 1 provides a brief overview of the methodology. Section 2 provides an overview of the evidence to date. Section 3 presents the data; the challenges and opportunities it presents. Section 4 explores the accuracy of synthetic panel estimates when the true $\rho_{y_{i1}y_{i2}}$ is known. Section 5 explores means for approximating $\rho_{y_{i1}y_{i2}}$. Section 6 concludes and summarizes the key findings. Supplementary material is also available in the form of an online "Appendix".

1 Methodology

The DL and DLLM approaches can be decomposed into two elements; income models and residual autocorrelation. In the absence of true panel data, both approaches require income models containing only time invariant household characteristics in order to link independent cross-sectional data. The DLLM approach provides nonparametric and parametric bounds under certain assumptions concerning the autocorrelation of the residuals from the aforementioned income models. The DL synthetic panel approach builds upon the methodology of DLLM by providing a method for approximating the autocorrelation between residuals which allows for point estimates of transitions. In what follows I will provide a brief overview of the DLLM approach and the DL innovation. What follows borrows heavily from the in depth exposition provided by Dang and Lanjouw (2021).

1.1 Income models

Household income can be estimated using linear projections of income for the two rounds of cross-sectional data:

$$y_{i1} = \beta_1^{'} x_{i1} + \epsilon_{i1} \tag{1}$$

$$y_{j2} = \beta_2^{'} x_{j2} + \epsilon_{j2} \tag{2}$$

where x_k denotes the set of time invariant household characteristics observable in both rounds and y_k refers to household income.[2]

[2] In all applications of the DL and DLLM approaches thus far, the log transformation of either household income or household consumption has been used.

1. Using the data in survey round 1 obtain predicted coefficients $\hat{\beta}_1$ and predicted residuals $\hat{\epsilon}_{i1}$ from the linear income model (1)
2. For each household in round 2 predict round 1 income using the predicted coefficient $\hat{\beta}_1$

Step 2 implies that households observed in period 2, with a certain set of time invariant characteristics, would have achieved the same average level of income in period 1 as similar households observed in period 1. This requires the first of the two key assumptions underpinning the DL approach.

Assumption 1 (A1): the underlying population sampled must be the same in survey round 1 and survey round 2.

This assumption is unlikely to hold in the case of large population shifts in terms of births, deaths and migration. As the interval between cross sections grows there is naturally greater scope for such changes to occur, however previous research has found the accuracy of estimates to be fairly insensitive to the length of interval under consideration (Dang and Lanjouw 2021; Hérault and Jenkins 2019).

1.2 Residual autocorrelation

The income models allow unobserved income to be predicted on the basis of time invariant household characteristics, but the question still remains as to how to treat the residuals. These residuals will be comprised of unobserved time invariant characteristics and time varying factors; it is therefore likely that residuals are correlated over time. DLLM provide both a nonparametric and a parametric treatment of the residual correlation.

1.2.1 Nonparametric bounds

The nonparametric bounds rest upon the assumption of non-negative correlation between ϵ_{i1} and ϵ_{i2}. In the case of students it is easy to imagine a scenario where low income during a period of study is then followed by higher income upon the completion of studies, however such a scenario is unlikely to prevail for the average household. Furthermore, steps, such as excluding those under or above a certain age from the sample, can be taken in order to reduce this risk. Under this assumption an upper bound can be estimated assuming the residuals are completely independent while a lower bound can be estimated assuming the residuals are perfectly correlated.

1.2.2 Parametric bounds

The bounds on poverty transitions can be further narrowed through the application of the second assumption underpinning the DLLM parametric approach.

Assumption 2 (A2): ϵ_{i1} and ϵ_{i2} have a bivariate normal distribution with (partial) correlation coefficient ρ and standard deviations $\sigma_{\epsilon 1}$ and $\sigma_{\epsilon 2}$.

In the absence of true panel data A2 cannot be tested. Validation studies thus far have found that while residuals typically fail formal tests of bi-variate log-normality, visual

inspection of the residuals indicates that it still provides a reasonable approximation. A2 results in the following parametric estimation framework:

$$P(y_{i1} \sim z_1 \text{ and } y_{i2} \sim z_2) = \Phi_2 \left(d_1 \frac{z_1 - \beta_1^{'} x_{i2}}{\sigma_{\epsilon 1}}, d_2 \frac{z_2 - \beta_2^{'} x_{i2}}{\sigma_{\epsilon 2}}, \rho_d \right) \tag{3}$$

where d_j is an indicator function equal to 1 if household is poor and -1 if the household is non-poor, $\rho_d = d_1 d_2 \rho$. In order to estimate parametric bounds, one requires an upper and lower estimate for ρ. DLLM suggest that the correlation coefficient from alternative panel data sets within the same country or from similar countries could be used.

Steps:

1. Estimate the income model for the two periods
2. Estimate Eq. 3 using upper and lower values for ρ taken from external sources

1.2.3 DL—point estimates

If the true partial correlation were known, then using Eq. 3, it would be possible to arrive at point estimates for transitions. DL show that the partial correlation can be derived from the simple correlation using the following formula:

$$\rho = \frac{\rho_{y_{i1} y_{i2}} \sqrt{var(y_{i1}) var(y_{i2})} - \beta_1^{'} var(x_i) \beta_2}{\sigma_{\epsilon 1} \sigma_{\epsilon 2}} \tag{4}$$

In the absence of panel data the simple correlation will not be available. Therefore, DL propose an approximation using cohort level averages. Assume household income follows a simple linear dynamic data generating process (AR(1)) given by:

$$y_{i2} = \alpha + \delta y_{i1} + \eta_{i2} \tag{5}$$

This is a fairly standard assumption in the pseudo-panel literature. The coefficient δ is a common measure of unconditional mobility with $\delta < 1$ indicating convergence and $\delta > 1$ indicating divergence of incomes (Antman and McKenzie 2007)[3]. Equation 6 sets out the relationship between the δ term and the statistic of interest $\rho_{y_{i1} y_{i2}}$.

$$\rho_{y_{i1} y_{i2}} = \sqrt{\frac{var(y_{i1})}{var(y_{i2})}} \delta \tag{6}$$

Pseudo-panel techniques replace individual level observations with cohort level averages.

$$\tilde{y}_{c(t),2} = \alpha + \delta \tilde{y}_{c(t-1),1} + \tilde{\eta}_{c(t),2} \tag{7}$$

[3] See Antman and McKenzie (2007) for a detailed account of pseudo-panel estimation. In the pseudo-panel literature the δ is commonly referred to as the β coefficient. I continue to use δ in order to maintain comparability with the previous synthetic panel validation exercises.

The simple correlation coefficient $\rho_{y_{i1}y_{i2}}$ can then be approximated by either directly calculating the cohort-level simple correlation coefficient $\rho_{y_{c1}y_{c2}}$ or via Eq. 8 the cohort alternative to Eq. 6.

$$\rho_{y_{c1}y_{c2}} = \sqrt{\frac{var(y_{c1})}{var(y_{c2})}}\delta \tag{8}$$

Equation 8 highlights the link between the DL approximate and the pseudo-panel approach for estimating δ. The DL approximate uses the ratio of the cohort level variances $\frac{var(y_{c1})}{var(y_{c2})}$, therefore even if δ were to be accurately estimated the ρ_c may prove to be inaccurate due to $\frac{var(y_{c1})}{var(y_{c2})} \neq \frac{var(y_{i1})}{var(y_{i2})}$. This leads to a related measure of ρ_c:

$$\rho_{y_{c1}y_{c2}} = \sqrt{\frac{var(y_{i1})}{var(y_{i2})}}\delta \tag{9}$$

This related measure, which Garcés Urzainqui (2017) refers to as the indirect DL, will be explored further in Sect. 5.

For both measures the grouping methods discussed are equivalent to instrumental variables (IV) methods where the IVs are the grouping variables—cohort dummy variables and the time dummy variables (Verbeek, 2007). In the first stage, individual income is regressed on cohort dummy variables, calculating the average income level for each cohort. These cohort averages are then used in Eq. 7.

In order for Eq. 7 to be consistently estimated via OLS, the standard criteria for valid instruments must be met. Consistent estimation of $\rho_{y_{i1}y_{i2}}$ requires IVs which are relevant, exogenous and satisfy the exclusion restriction. The exclusion restriction requires that the instrument affects the outcome variable only through the instrumented variables. Verbeek and Vella (2005) show that this requirement will not be met if cohort effects are present in the residual represented by η_{i2} in Eq. 5.[4] Pseudo-panel approaches also impose a full rank requirement which simply states that there must be sufficient variation in cohort level averages.

Typical cohort definitions used in the pseudo-panel literature are birth cohort and sex. The exclusion restriction and full rank requirement suggest that the number of cohorts should be sufficiently large so as to counter potential systematic differences in individual characteristics and to allow for sufficient variation. However, researchers must also be wary of small sample bias when cohorts do not contain sufficient individuals. This leads to the classic trade-off found in the pseudo-panel literature between the number of cohorts and the minimum number of individuals in each cohort (Antman and McKenzie 2007). Verbeek and Nijman (1992) proposes that a cohort should contain a minimum of 100-200 observations, whereas Devereux (2007) suggests that much larger cohorts are required (500+).

[4] Verbeek and Vella (2005) propose an augmented IV approach to overcome this issue, but it requires at least 3 periods of observation.

2 Validation approaches and findings

Previous validation exercises of the DL method have been conducted using two empirical strategies; within panel analysis and rotating panel analysis.

Within panel analysis uses actual panel data to mimic the conditions under which synthetic panel techniques may be applied. This is done by randomly splitting the data in half; one half of the data is used for the income model in period 1 while the other half is used for the income model in period 2. This process is repeated R times to avoid spurious results relating to any particular split. While such an approach has the advantage of automatically satisfying assumption 1, it does constrain one of the key advantages of cross-sectional data, namely larger sample size (Dang and Lanjouw 2021). The smaller sample size resulting from using only half of a panel sample will reduce the accuracy of the income model, reduce the number and size of potential cohorts for the approximation of the partial correlation coefficient, and may have implications for the bi-variate normal assumption.

The alternative to splitting true panel data is to use surveys with a rotating panel design. A rotating survey design is one which combines cross-sectional and longitudinal elements. A subset of households from each survey round are followed for multiple time periods. This approach affords the synthetic approach the larger sample size of cross-sectional data, but assumption 1 is no longer automatically satisfied.

In what follows I will discuss the empirical validation exercises in terms of the two key components of the DL approach. I will first discuss findings relating to the income models and performance when the true rho is used before discussing the performance of the DL innovation for approximating $\rho_{y_{c1}y_{c2}}$.

2.1 Results with true rho

Table 1 provides an overview of the validation studies thus far, outlining the key choices made by the authors. Validation exercises have been conducted using data sets from 8 countries; 5 developing countries and 3 developed countries. There is an equal split between consumption as the measure of well-being and income; however, it is important to note that all 3 developed countries use income as the measure of well-being.

The age column indicates the age range for the household head for which analysis is conducted. The DL approach links households over time via the time invariant characteristics of the household head. Households with younger household heads may be more unstable due to a greater incidence of dissolution and formation. Similarly older households may face a greater risk of dissolution due to higher death rates. Consistent with the pseudo-panel literature, DL restrict analysis to households where the household head is aged between 25 and 55 while Garcés Urzainqui (2017) is more lax about the choice of age range opting for the wider range of 25-70. HJ consider two age ranges throughout their analysis; 25-55 and 25-75. They find that the performance of the DL method is sensitive to the choice of age range and that the 25-75 range produces more accurate estimates.

Table 1 Overview of the empirical validation of the DL method

	Validation Approach	Country	Time period	Y/C	Age	Income Model	Cohort Definition
DL 2017	Within	Bosnia-Herzegovina	2001, 2004	C	25–55	Age, sex, education level, ethnic majority, urban	yob(1)
		US	2004, 2007, 2009	Y			
	Rotating	Lao PDR	2002, 2007	C			
		Peru	2004, 2006	C			
		Vietnam	2004, 2006, 2008	C			
Urzanqui 2017	Within	Thailand	2006–2007	Y	25–70	Sex, religion, civil status, education level	yob(3) by region of birth
HJ 2018	Within	UK	2001–2015	Y	25–55 and 25–75	Sex, yob(5), education level, ethnicity/cob	yob(5)*sex
		Australia	1991–2008 (1 yr intervals)	Y			yob(5)*cob

Notes. The interval column lists the years of data used to calculate the true panel and synthetic panel estimates. For example in DL a 3-year synthetic panel is constructed from 2004-2007 for the US. The Y/C column indicates whether household income or household consumption are used to measure well-being. The abbreviations yob and cob refer to year of birth and country of birth respectively. All income models contain an indicator for education however depending on the availability of data the number of categories varies. A1 is automatically satisfied for the within panel countries, while for the rotating panel countries it is satisfied for all countries exception Lao PDR

In all evaluations, income models have contained variables relating to age, sex, education, and religion/ethnicity/country of birth. Garcés Urzainqui (2017) provides a sensitivity analysis of synthetic poverty transition estimates with respect to the choice of income model. Moving from parsimonious models to more complex models, which contain variables whose claim of time invariance is more doubtful, Garcés Urzainqui (2017) finds that given an accurate estimate of ρ even parsimonious models may give a good impression of poverty transitions.

The evidence with respect to the choice of income model is thus quite encouraging. However, one important gap in the literature is the sensitivity of sub-population estimates to the choice of income model. Parsimonious models may work well in aggregate, but less so for certain sub-populations particularly if time invariant characteristics specific to that sub-population are not included in the income model. Furthermore, when using the bounds approach, the size of the bounds will increase with more parsimonious income models.

DL do not provide estimates of poverty transitions using the true ρ, instead they report transitions using an approximation of ρ based upon 1 year age cohorts. The precision of estimates is evaluated by counting the number of times point estimates lie within the 95 % confidence interval of the true estimate. The authors find that all of the estimated conditional probabilities and 18 out of the 20 joint probabilities lie within the 95 % confidence of the true estimate. The worst performing country by this measure of precision, and the more testing measure of lying within one standard deviation of the true estimate, is the US—the only high income country. Garcés Urzainqui (2017) finds, using the preferred income model, that all joint poverty estimates lie within the 95 % confidence interval of the panel estimate.

The validation exercise of HJ is by far the largest in terms of the number of time periods considered and it is also much less optimistic. They find that even using the true panel rho a much higher number of synthetic panel estimates lie outside the 95 % confidence interval of the true panel estimate than has previously been found. For certain years, the DL approach performs well with all estimates lying with the CI, but for other years less than half of the probabilities are accurately estimated. Importantly though, chronic poverty is almost always accurately estimated. The accuracy of synthetic panel estimates is found to vary with the poverty line used.

HJ also examine the potential for parametric bounds estimated using rho values of 0.5 and 0.9. The authors find that while these bounds are highly successful at encompassing the true panel estimate, they are often too large to be of practical use. This is indeed true, however, this need not always be the case. The parametric bounds calculated by HJ are based on a sample which is half the size of the available longitudinal data set. In practice there is no need to split the data set and cross-sectional data sets are typically larger than longitudinal; parametric bounds may therefore be of practical use.

One interpretation of the results so far is that the DL method works well in developing countries, but less well in high income settings. However, the precision of existing panel estimates, as conceptualized by the confidence intervals, will be determined by the size of the panel, the value of the proportion in question ($p(1-p)$), and the survey design. This results in an inconsistent benchmark. Furthermore, in DL the standard errors for the true poverty dynamics using US data are calculated without taking into

account complex survey design resulting in extremely narrow confidence intervals. In terms of the absolute difference between the true panel values for the US and synthetic panel estimates, the synthetic panel estimates for the US are among the most accurate. In HJ the true poverty dynamics and their confidence intervals in both the UK and Australia are derived from the full longitudinal sample. The results presented in HJ, therefore, compare the accuracy of synthetic panel estimates to a longitudinal dataset which has twice as many households. The implication of differing survey design effects and the calculation of confidence intervals somewhat erodes the conclusion of relatively poor performance of the DL approach in high income countries.

Inconsistent benchmarks beg the question as to how one should evaluate synthetic panel estimates. If users are primarily interested in levels of chronic poverty then evaluation based upon absolute differences between true panel and synthetic panel estimates could be appropriate. If trends in poverty dynamics are of interest than it may be sufficient to examine whether the true and synthetic panel estimates move in similar directions over time. Or if policy makers are interested in identifying at risk groups then it may be sufficient if the synthetic panel estimates can accurately rank important sub-populations in terms of their riskiness to persistent poverty.

In what follows I will continue to assess the accuracy of DL estimates using the confidence interval surrounding the true panel estimates, but will also consider the aforementioned user needs. A broader approach to validation should ultimately help to identify the practical usefulness and limitations of the DL approach.

2.2 Considerations in defining cohorts

DL do not perform an empirical analysis of the sensitivity of their findings with respect to the definition of cohorts. Garcés Urzainqui (2017) examines 9 alternative cohort definitions using birth cohorts of different length and interacting these cohorts with sex and region of birth. The findings are encouraging in the sense that examining cohort level correlations in average income can give a good approximation of the true correlation, and concerning in that the correlation is sensitive to the choice of cohort definition. Garcés Urzainqui (2017) also considers an alternative synthetic panel approach proposed by Bourguignon and Moreno (2020). Bourguignon and Moreno (2020) propose incorporating the estimation of ρ into the synthetic panel procedure. The discussion and application of this approach is beyond the scope of this paper, but Garcés Urzainqui (2017) finds that the DL approach produces more accurate estimates of ρ and more accurate synthetic panel estimates.

HJ find significant volatility in the DL ρ depending upon the cohort definition used. In the words of Hérault and Jenkins (2019), this is, 'worrying because researchers applying the DL method (and without longitudinal data) might choose the 'wrong' definitions and sample selection criterion, if only because of data constraints.' Furthermore, cohort definitions which may perform well on average for all the years considered may perform poorly for one specific year suggesting, that when possible, the cohort rho should be averaged over multiple time periods. The volatility in cohort ρ is particularly large when examining the age range 25-55.

As noted in Eq. 8 the pseudo-panel approximation of ρ relates directly to the means based approach for measuring convergence/divergence in incomes via the δ term in Eq. 7. It is therefore informative to examine the means based unconditional mobility literature. Antman and McKenzie (2007) examines the accuracy of the means based approach using Mexican household data. Restricting the analysis to households where the household head is aged 25-49, they find that cohorts defined by yob(5)*education, where education is divided into 3 categories, produces realistic δ coefficients.

Fields and Viollaz (2013) use Chilean data to examine the accuracy of the means based approach. Restricting their sample to households where the household head is aged between 20-65, they find that cohorts defined by yob(2)*sex can accurately approximate the unconditional δ coefficient. Both Fields and Viollaz (2013) and Antman and McKenzie (2007) fail to address how the most accurate cohort definition can be selected in the absence of true panel data.

In summation, it appears that pseudo-panel techniques can be used to approximate ρ, but the accuracy is sensitive to the definition of cohorts. This volatility need not be a hindrance if guidance emerges as to how to identify the best performing cohort definition. From the perspective of the practical application of the DL method it is not the sensitivity of the DL rho to the choice of cohort definition which is relevant, but rather whether a particular cohort definition performs well across a number of different settings or if certain characteristics of cohort definitions are associated with more accurate estimates.

3 Data

Synthetic panel approaches require at least two rounds of cross-sectional data. The wording of questions, sampling approach and treatment of data must be similar in both periods to credibly link these independent samples. Furthermore, the validation of synthetic panel estimates requires true panel data, constructed in a similar manner as the cross-sectional data, for comparison. This paper makes use of both the longitudinal and cross-sectional releases of EU-SILC (2006-2017) made available by Eurostat in order to validate the DL approach.

EU-SILC is a household survey which covers 28 EU and non-EU countries over varying lengths of time. It is fundamental to the measurement of poverty and inequality and also helps to shape future policy. It consists of two types of data; cross-sectional and longitudinal. The longitudinal element of EU-SILC is comprised of a four year rotating panel for each country. For a given cross-sectional data set at time t, 75% of the original households will be re-interviewed in t+1, 50% in t+2 and 25% in t+3.

The quality of the longitudinal element of EU-SILC is undermined by two factors; attrition and survey design variables. Evidence from 3 high quality panel surveys, the British Household Panel Survey (BHPS), the Household, Income and Labour Dynamics (HILDA) survey in Australia, and the German Socio Economic Panel, suggests that the retention rate for a three year panel should be approximately 80 % (Jenkins (2010); Kroh and Spieß (2006); Watson and Wooden (2006); Lynn et al.

(2006)). The majority of countries included in the 2009-2012 EU-SILC longitudinal dataset fail to match the retention rates of high quality panels.[5]

At present the publicly available version of the longitudinal element of EU-SILC does not include a variable detailing which stratum a household belongs too (Iacovou et al. 2012). A PSU variable is available, but incorporating clustering while ignoring the strata can lead to an overestimation of standard errors in the context of poverty rates (Howes and Lanjouw 1998). Point estimates of poverty transitions will not be affected, but constructing confidence intervals without taking clustering and strata into consideration will lead to erroneous estimates of precision.

The probability weights accompanying the longitudinal element of EU-SILC are also a source of concern. These weights are re-calibrated for each year to account for attrition and ensure that the sample remains representative of the true population. Despite these efforts a number of countries report discrepancies between the poverty rate calculated using the cross-sectional data and the longitudinal data for the same period (Jenkins and Van Kerm 2017). Such discrepancies cannot be explained by changes in the underlying population and are therefore indicative of poor data quality.

While certain characteristics of EU-SILC show the potential usefulness of synthetic panel approaches they also present a challenge for validation exercises. Within panel validation can overcome the problem posed by high levels of attrition as both constructed cross-sectional data and the panel data will be drawn from the same sample. However, the length of within panel data which can be considered will be limited by the rotating panel design of EU-SILC. For each additional year added to the panel length the sample size decreases by approximately one third. Given this declining sample size the within panel approach to validation is most feasible for one year panels and for countries with larger longitudinal samples. It is worth bearing in mind that the sample size of a one year panel will be approximately three quarters the size of the cross-sectional data for a given year. For within panel validation this longitudinal sample must be randomly split in half, resulting in an effective sample size which will be approximately three eights of what would be available for a true application of the DL approach.

France, Poland and Greece are the three countries selected for within panel validation. These three countries are selected due to their large sample sizes with each one year panel containing approximately 4500–5500 households.[6] The three countries selected also represent different levels and patterns of poverty over time.

In line with common practice I use the equivalized household disposable income variable provided by Eurostat as the measure of living standards.[7] Following the approach of Van Kerm and Alperin (2013) I record as missing any income smaller than 75 % of the lowest percentile or higher than 125 % of the top percentile. This

[5] Please see A.1 in the "Appendix" for 3 year retention rates.

[6] At the beginning of the time period considered Greece has a relatively small sample size, however, this increases dramatically over the time period considered.

[7] Household disposable income includes all income from work, private income from investments and property, transfers between households, and all social transfers received. This measure of annual household disposable income is then divided by the number of single adult equivalents in the household (according to the modified- OECD equivalence scale) to arrive at an individual measure of single adult equivalent disposable income attributed to all household members.

process effects a very small number of households and helps to ensure results are not driven by potentially misreported incomes.

In keeping with Eurostat's 'headline' poverty statistics, the primary poverty line used throughout the paper is 60 % of contemporary national median income.[8] This poverty line is comparable to that used by HJ in that it is a relative line, but differs from the absolute poverty lines used by DL and Garcés Urzainqui (2017). In Sect. 4, I examine the sensitivity of the synthetic poverty estimates to alternative poverty lines.

4 Findings with true $\rho_{y_{i1}y_{i2}}$

In this section the performance of the DL approach is examined when the true $\rho_{y_{i1}y_{i2}}$ is known. Using the true $\rho_{y_{i1}y_{i2}}$ isolates the performance of the income model and bivariate normal assumption allowing one to abstract from performance issues related to the ability of pseudo-panel approaches to approximate $\rho_{y_{i1}y_{i2}}$. In particular this section examines the accuracy and sensitivity of estimates to alternative income models, poverty lines, and sub-population decompositions.

4.1 Income models

Alongside the estimation of residual autocorrelation, the income model plays a key role in the DL synthetic panel approach. The challenge to researchers utilizing this method is how to assess such an important component. Some guidance is provided by DL, who state that regressors in time t+1 should be either time invariant or time-varying but easily recalled for period t.

The strict time invariance of household characteristics is the key to linking independent cross sections under the synthetic panel approach. For some variables the case is clear cut. Year of birth does not change over time and thus is certainly time invariant. The number of children under a certain age threshold is not time invariant, however it is, to a certain degree, predetermined. An individual's education level plays an important role in determining their income and yet is not time invariant. Restricting the age of the household head considered in the sample can strengthen the claim of time invariance for education—in many countries individuals rarely add to their formal education after the age of 25. Furthermore, using information on when an individual finished their education it is possible to ensure that no changes in education level occurred in the interval between cross sections.

These are the kinds of issues that a researcher must consider when constructing the income model. Income models can be, and often are, assessed in terms of their predictive power; however, first and foremost researchers must ensure the variables in their models are time invariant otherwise the validity of their findings are undermined.

In order to test the sensitivity of poverty estimates to the income model, I construct models on the basis of a tiered approach similar to Garcés Urzainqui (2017). These tiers begin with the most assuredly time invariant variables to those which are likely

[8] In line with common practice these thresholds are generated using the cross-sectional element of EU-SILC.

time invariant, but may not be for a small section of the population. The most basic model includes: sex, 5 year birth cohort, country of birth.

In line with previous findings, the accuracy of synthetic panel estimates are robust to the choice of income model. Given this robustness, one may be tempted to conclude that the most parsimonious model should therefore always be favoured as it is most assuredly time invariant. However, it is important to note that parsimonious models will perform poorly for certain sub-populations.[9]

In what follows an income model comprised of sex, 5 year birth cohort, country of birth, and interaction terms between sex and education and sex and birth cohort is used. All included variables are time invariant or can be made time invariant. This is comparable to the income model used by HJ, however, the adjusted R^2 for tier 3 for each country is higher than those reported by HJ.[10]

4.2 Normality of residuals

When the true $\rho_{y_{i1}y_{i2}}$ is known, inaccuracies in synthetic panel estimates can arise from the two underlying assumptions; A1 stable population and A2 bivariate normal assumption for the residuals. The within panel validation approach ensures that A1 is met, therefore the bivariate normal assumption is the likely culprit behind inaccurate estimates. This raises the question as to how the bivariate normal assumption can be assessed in the absence of true panel data. One straightforward approach for assessing the bivariate normal assumption is to examine the normality of residuals for the income model in each period. In order for the residuals from the income models to follow a bivariate normal distribution, the year specific residuals must be normally distributed. While strict normality of residuals is unlikely to hold, visuals representations of the residual distribution through p-p plots, q-q plots and histograms can help practitioners to predict if synthetic panel estimates will be accurate and for which dynamics inaccuracies may occur.

Figure 1a presents the p–p plot of the residuals for Greece (2014). P–P plots present the cumulative probabilities at certain values for the empirical distribution of residuals in comparison to the normal distribution. For Greece the empirical distribution lies above the 45 degree line in the lower half of the distribution. Where the normal distribution expects a cumulative probability of 25%, the empirical distribution reports approximately 15% for Greece. Using the normal distribution would thus overstate the probability of negative residuals in comparison to the empirical distribution. In terms of poverty dynamics, the use of the bivariate normal assumption for residuals should lead poverty in Greece to be overstated in the years considered.

There are a number of techniques available to practitioners for improving the normality of residuals. One such technique is to identify overly influential data points and remove them from the income model. Given that income is generally considered to be more volatile than consumption, this may be a more salient issue when income is the

[9] Please see "Appendix A2" for a comparison the performance of income models for a given year, over time, and for one particular subpopulation.

[10] Using a comparable model to Tier 3, HJ find adjusted R^2 values ranging from 0.14-0.17 depending on the year considered.

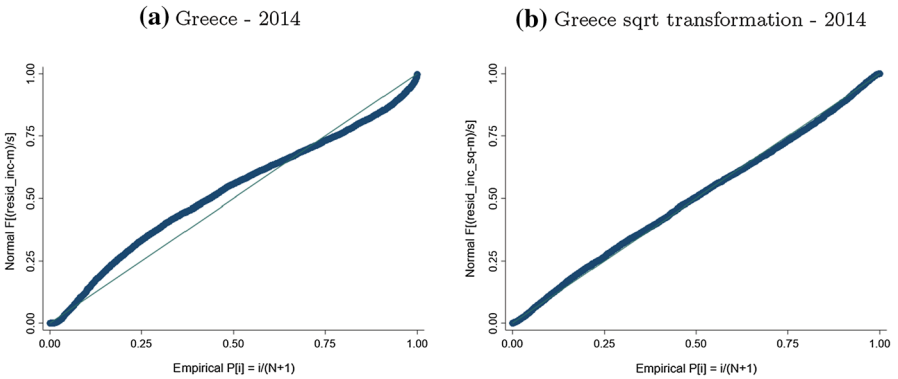

Fig. 1 Normality of residuals in base year—ln transformation. The reported residuals are calculated using income model 3 and the longitudinal sample for the given year

measure of welfare (World Bank 2014). Typically, one is reluctant to exclude outliers unless there are genuine concerns over erroneous values; however, since the income model is used to predict income, it is possible to exclude outliers from the income model whilst still including them in the estimation of synthetic panel estimates. Alternatively one could transform the variable of interest. Every application of the DLLM approach takes the logarithm of the variable of interest. This certainly helps, but as Fig. 1a shows it is not always sufficient. There are alternative transformations, such as taking the square root which may be more suitable.

Figure 1b presents P–P plots where the square root transformation rather than the log is taken for Greek household income. The effect of this alternative transformation is dramatic. The empirical distribution of residuals now closely matches the normal distribution in terms of cumulative probabilities. This lends support to the bivariate normal assumption.

Prior to the estimation of synthetic income dynamics, it is also informative for practitioners to examine the predicted poverty rate under the assumption of normally distributed residuals. Figure 2 compares the true poverty rate and synthetic poverty rate for Greece. The true and synthetic poverty rates both refer to period t. The synthetic poverty rates are calculated using the predicted incomes for period t. Assuming the residuals are normally distributed a synthetic poverty rate for time period t is estimated.[11] Comparing the true and synthetic poverty rates for Greece, again reveals that the synthetic poverty rate estimated using the logarithmic transformation overestimates poverty. The square root transformation results in more accurate synthetic poverty rates for Greece. In what follows the square root transformation is used for both Greece and Poland.[12]

[11] A similar level of accuracy is found when synthetic poverty rates for $t-1$ are estimated using period t time invariant household characteristics and are compared to the true poverty rates for $t-1$.

[12] Please see section A.3 of the "Appendix" for a comparison of true and synthetic poverty rates for France and Poland—synthetic estimates are found to be highly accurate.

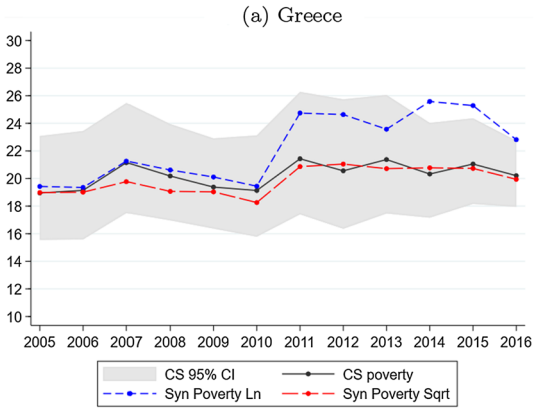

Fig. 2 Comparing true and synthetic poverty rates. All of the reported point estimates and confidence intervals are the average values for the 51 random splits of the longitudinal data

4.3 Poverty dynamics

In this section poverty dynamics for France, Poland and Greece are presented. Figure 3 presents the findings for France when the true $\rho_{y_{i1}y_{i2}}$ is known. Alongside estimates using the true $\rho_{y_{i1}y_{i2}}$, parametric bounds are estimated with $\rho_{y_{i1}y_{i2}}$ fixed at either 0.6 or 0.9. For the majority of country settings included in EU-SILC this bound encompasses the true one year $\rho_{y_{i1}y_{i2}}$. The inclusion of estimates using a fixed ρ has the additional benefit of allowing for a visual inspection of whether it is changes in the returns to time invariant characteristics of household heads or whether it is changes in the $\rho_{y_{i1}y_{i2}}$ term from year to year which drives the observed trends.

In line with prior research, the DL approach is found to be extremely accurate at measuring the rate of chronic poverty; all estimates lie within the 95% confidence interval of the true estimate.[13] Importantly, the DL estimates also accurately capture the upward trend in chronic poverty in France due to the onset of the recession. The three remaining joint probabilities are estimated with a reasonable degree of accuracy, with estimates lying within the 95% confidence intervals or marginally outside. It is worth bearing in mind that were survey design variables available one could reasonably expect all of the DL estimates to lie within the 95% confidence interval of the panel estimate. Similar to HJ, the DL approach is found to systematically underestimate the probability of being non-poor in both periods; however, it does successfully capture the trend.

The parametric bounds successfully capture all of the true joint poverty dynamics with the exception of those persistently non-poor. It is interesting to note that the synthetic panel estimates using the true $\rho_{y_{i1}y_{i2}}$ are more accurate at capturing changes in trends than the parametric bounds with fixed ρ. This can be most clearly seen in the 2006-2008 period for the (poor, non-poor) joint probability. A similar level of accuracy is found for both Poland.[14]

[13] Confidence intervals are calculated as the average of the 51 true panel splits. This ensures that the synthetic panel estimates are being compared to a panel dataset of comparable size.

[14] Please see Fig. 5 in "Appendix A.3".

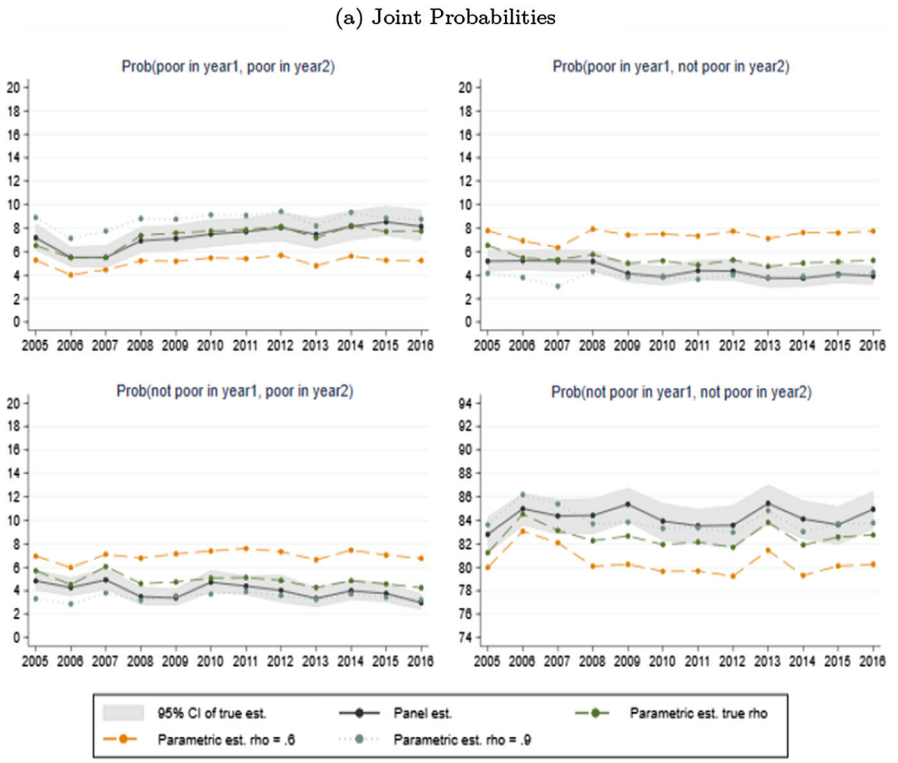

Fig. 3 France—poverty line set at 60 % of median income. All of the reported point estimates and confidence intervals are the average values for the 51 splits

Figure 4 presents the joint probabilities for Greece using both the log transformation and square root transformation of household income. The difference between the two synthetic panel estimates is dramatic. While the log transformation continues to accurately capture the rate of chronic poverty, the estimates for the three remaining joint probabilities are largely inaccurate and erratic. In total 22 out of 48 of the synthetic joint probabilities estimates lie outside the 95% confidence interval of the true panel estimates. By comparison, using the square root transformation results in only 2 out a total of 48 joint probabilities lying outside the 95% confidence interval of the true panel estimates. This highlights the important role of examining the normality of the residuals from the income models and taking action to improve normality.

4.4 Conditional probabilities

Figure 5 presents the most common conditional probabilities for France.[15] Similar to HJ, no drop in accuracy is found when examining conditional probabilities. The number of synthetic conditional probabilities which lie outside the 95% confidence

[15] A similar level of accuracy is found for Poland and Greece—please see Fig. 6 in the A.4.

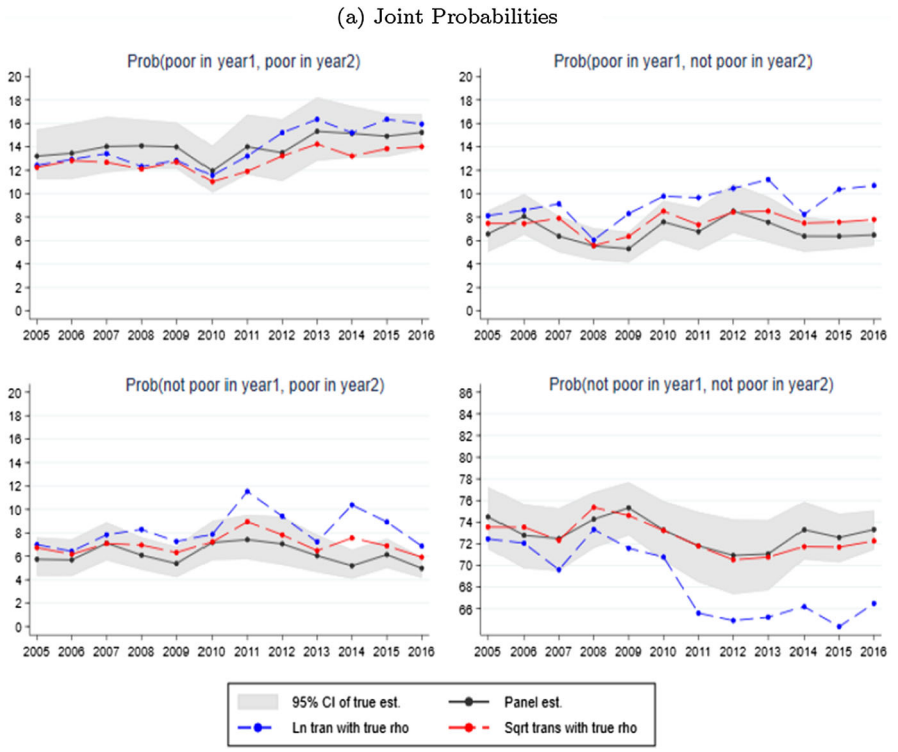

Fig. 4 Greece—poverty line set at 60 % of the median. All of the reported point estimates and confidence intervals are the average values for the 51 splits

interval of the true probability for the exit and entry rates closely matches the accuracy of the respective joint probabilities. This is unsurprising as the joint probabilities are used in the calculation of the conditional probabilities.

There is, however, a much larger absolute difference between the synthetic and true exit rates. This is particularly evident in the case of France. The large absolute differences observed for France are due to the small discrepancies between the true and synthetic joint probabilities for being poor in t-1 and non-poor in t being magnified by the small denominator of being poor in period t-1. This highlights that greater caution should be exercised when examining synthetic panel estimates of exit rates in countries with low levels of poverty in the initial period.

4.5 Alternative poverty lines

The headline findings identify an individual as poor if their income is below 60% of the median income. This is the poverty line used by Eurostat when analysing the Europe 2020 goals, however this is not the only poverty line in use in developed countries. The OECD (2019), for example, defines poverty as an income below 50% of the median income.

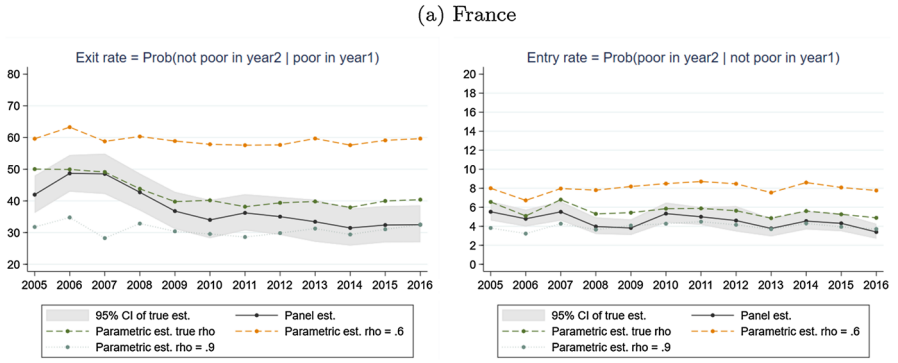

Fig. 5 Conditional probabilities. *Notes* All of the reported point estimates and confidence intervals are the average values for the 51 splits

Figure 6 presents the joint probabilities when the poverty line is set at 80% of the median income. Similar to HJ, I find a deterioration in the performance of the DL estimates as the poverty line rises.[16] Given that the 60% and 50% of median poverty lines are those most commonly used in the developed countries settings this does not represent a significant concern for poverty analysis, yet it does raise doubts over the suitability of the DL approach for broader mobility measurement. A similar pattern is observed for both Poland and Greece (see "Appendix A.4.1").

4.6 Alternative age range

When examining estimates for the 25-55 age range, it is worth noting that the confidence intervals for the narrower age range should be larger due to the smaller sample size; this make the binary measure of accuracy less demanding. The findings for the 25-55 age range are comparable to those found for the 25-75 age range in terms of estimates lying with the 95% confidence interval, absolute differences and in ability to replicate significant changes in trends. This robustness to alternate age ranges when the true $\rho_{y_{i1}y_{i2}}$ is known is found for all countries (see "Appendix A.4.2").

4.7 Sub-populations

Policy makers are not solely concerned with overall poverty transitions; rather, they often wish to identify which sub-populations are particularly vulnerable to persistent poverty so that policy interventions may be better targeted. The DL approach is found to accurately capture poverty dynamics for subpopulations determined by the sex and education level of the household head (see "Appendix A.4.3").

[16] Using a lower poverty line does not decrease accuracy—please see Fig. 7 in the "Appendix A.4.1".

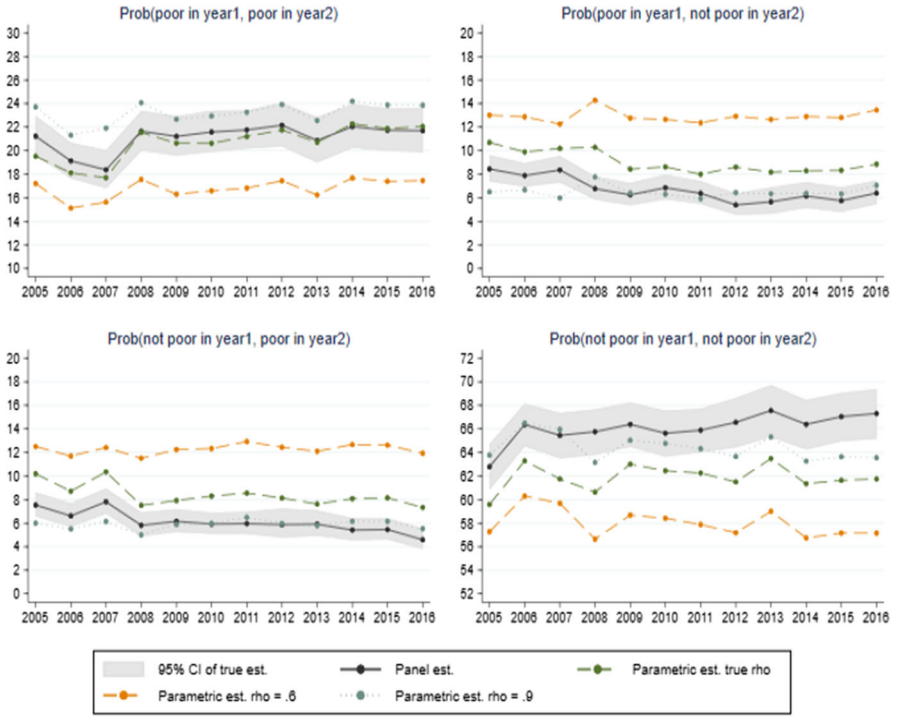

Fig. 6 France—joint probabilities mw80. All of the reported point estimates and confidence intervals are the average values for the 51 splits

4.7.1 Urban versus rural poverty dynamics

Applications of the DL approach thus far have tended to produce household level probabilities using the entire sample. These probabilities are then aggregated to the national or sub-national level. An alternative approach, which is particularly salient for comparing urban and rural poverty dynamics, is to estimate separate income models for urban and rural households to arrive at probabilities of poverty dynamics. This separate income model approach allows for differences in the returns to time invariant household characteristics between the regions and also allows the use of alternative $\rho_{y_{i1}y_{i2}}$ estimates for the two areas. However, the use of separate income models comes at the cost of fewer households for predicting income which may have consequences for the normality of residuals.

Figure 7 compares the estimates of poverty dynamics for urban areas using aggregate and separate income models for Poland. For the aggregate estimates a dummy variable for region is included in the income models. The use of separate income models for urban and rural areas results in more accurate synthetic poverty estimates. This improved precision is driven by allowing different returns to time invariant households

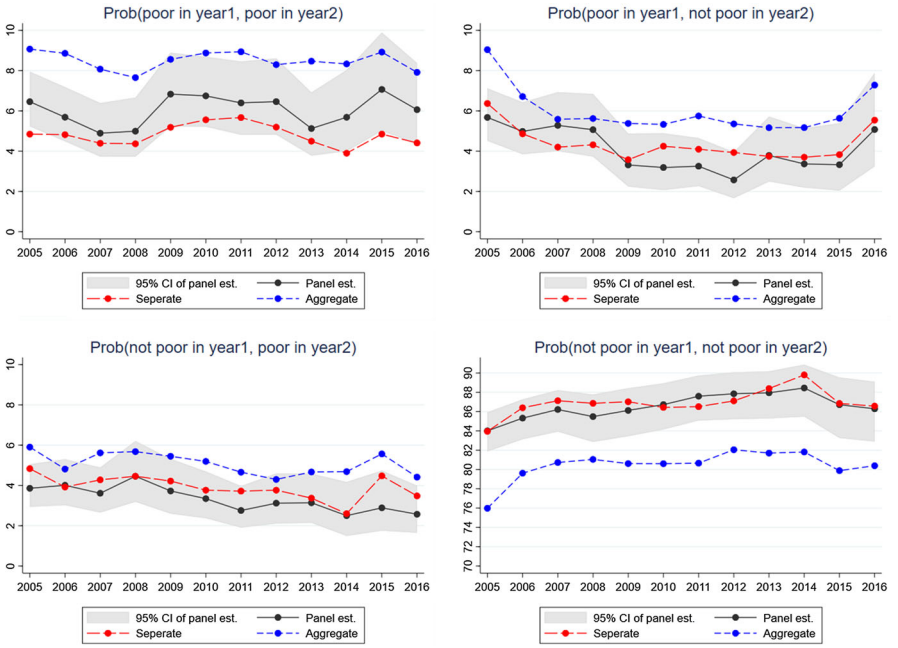

Fig. 7 Poland urban poverty dynamics. All of the reported point estimates and confidence intervals are the average values for the 51 splits

characteristics, rather than through the use of different ρ values for the two regions. A similar improvement is found for synthetic panel estimates for rural Poland.[17]

The findings for Poland open up an alternative path to estimating synthetic poverty dynamics; separate income models could be estimated for subpopulations based on region or education level of the household head. Aggregate estimates could then be produced by weighting subpopulation estimates with respect to their share of the population and summing up. This approach would require subpopulation estimates for $\rho_{y_{i1}y_{i2}}$ and may fall foul of small sample sizes, but in certain settings it may be a viable approach. The analysis of residuals and the comparison of synthetic poverty and true poverty rates would also provide important checks for subpopulation synthetic panel estimates.

5 Approximating $\rho_{y_{i1}y_{i2}}$ and δ

Alongside the income model, the approximation of $\rho_{y_{i1}y_{i2}}$ is a key component of the DL method. In this section pseudo-panel techniques for approximating $\rho_{y_{i1}y_{i2}}$, using neighbouring countries $\rho_{y_{i1}y_{i2}}$, and using rotating panel elements to approximate $\rho_{y_{i1}y_{i2}}$ into the future are explored.

[17] Please see Fig. 32 in "Appendix A.4.3".

5.1 Pseudo-panel techniques for approximating $\rho_{y_{i1}y_{i2}}$

DL propose using pseudo-panel techniques for approximating $\rho_{y_{i1}y_{i2}}$. Equation 8 highlights the link between the common place pseudo-panel estimate of unconditional convergence and the DL pseudo-panel approach for approximating $\rho_{y_{i1}y_{i2}}$. Table 2 presents estimates of both these measures derived using alternative cohort definitions and the 3 year panel element of EU-SILC. In the analysis which follows the rotation groups present in both cross sections are dropped. As highlighted by Garcés Urzainqui (2017), the presence of the same households in both cross sections may bias the findings.

Unlike HJ and Garcés Urzainqui (2017), the DL ρ is not directly reported here, rather the correlation between the income measures in two time periods is reported. This allows for the accuracy of the DL pseudo-panel approach to estimating $\rho_{y_{i1}y_{i2}}$ to be examined alongside the pseudo-panel approach for approximating the unconditional convergence (δ) and the indirect route to approximating ρ expressed in Eq. 9.

Table 2 presents the average true panel $\rho_{y_{i1}y_{i2}}$ and δ and the average approximated $\rho_{y_{i1}y_{i2}}$ and δ obtained from alternative cohort definitions for France, Poland and Greece. Three types of cohort definition are examined: year of birth (yob) cohorts, yob and sex cohorts, and yob and education cohorts.

Within each type of cohort definition, it is unsurprising to find that the average approximated $\rho_{y_{i1}y_{i2}}$s and δs increase as the cohort definition used increases in size and decreases in number. Similar to the findings of HJ, in both France and Poland there are a number of cohort definitions which can on average accurately approximate $\rho_{y_{i1}y_{i2}}$, however, the question facing practitioners is how to identifying the best performing cohort. There are three summary statistics which can, in theory, help to guide practitioners; the adjust R^2, the average cohort size (size), and the number of cohorts(N). The relationship between these three elements and the accuracy of the pseudo-panel estimates is, unfortunately, ambiguous.

Focusing on cohorts defined by year of birth, yob(4) on average provide reasonably accurate approximates of $\rho_{y_{i1}y_{i2}}$ for both Poland and France for the 25-55 age range. This is despite large differences in the adjusted R^2 between France and Poland as well as differences in the average number of observations. In both countries, however, the average number of observations per cohort exceeds 100, the criteria set forth by Verbeek and Nijman (1992). The average number of observations per cohort may conceal some cohorts with very small numbers of observations, but this should not be a source of significant volatility as the analysis is performed using weights.

Moving to the 25-75 age range, it is interesting to note that while both the true $\rho_{y_{i1}y_{i2}}$ and the pseudo-panel approximates increase as the age range incorporates older households, this increase is much more pronounced for the pseudo-panel approximates. A consequence of this larger increase in the pseudo-panel estimates is that the cohort definitions which are most accurate for the 25-55 age range are not always the most accurate for the 25-75 age range. This is particularly evident in Poland where yob(2) is the least accurate approximate for the 25-55 age range and yet is the most accurate approximate for the 25-75 age range. It is clear that pseudo-panel estimates are sensitive to the age range examined.

Table 2 Pseudo-ρ—three-year panels

	Age 25–55					Age 25–75				
	ρ	δ	Adj R^2	Size	N	ρ	δ	Adj R^2	Size	N
(a) France										
Panel	0.71	0.72				0.73	0.72			
	(0.68, 0.73)	(0.69, 0.74)				(0.71, 0.75)	(0.70, 0.74)			
yob(2)	0.47	0.46	0.018	139	15	0.58	0.51	0.022	129	25
yob(3)	0.51	0.49	0.018	190	11	0.66	0.60	0.022	189	17
yob(4)	0.67	0.67	0.017	261	8	0.74	0.69	0.021	247	13
yob(5)	0.72	0.80	0.016	347	6	0.76	0.73	0.020	321	10
yob(3)*Sex	0.50	0.48	0.025	94.8	22	0.59	0.53	0.027	94.5	34
yob(4)*Sex	0.62	0.61	0.023	130	16	0.65	0.59	0.027	124	26
yob(5)*Sex	0.70	0.75	0.021	174	12	0.71	0.67	0.025	161	20
yob(10)*Sex	0.81	1.00	0.018	347	6	0.83	0.86	0.023	321	10
yob(1)*Ed	0.71	0.66	0.11	22.5	93	0.74	0.67	0.11	21.1	152
yob(2)*Ed	0.84	0.81	0.098	46.3	45	0.85	0.79	0.11	42.9	75
(b) Poland										
Panel	0.67	0.63				0.69	0.65			
	(0.63, 0.70)	(0.59, 0.66)				(0.66, 0.72)	(0.63, 0.68)			
yob(2)	0.52	0.63	0.0090	365	15	0.68	0.81	0.023	340	25
yob(3)	0.61	0.76	0.0087	497	11	0.74	0.92	0.023	499	17
yob(4)	0.69	0.95	0.0083	684	8	0.77	1.00	0.023	653	13
yob(5)	0.77	1.10	0.0079	912	6	0.80	1.05	0.022	849	10
yob(2)*Sex	0.39	0.44	0.010	182	30	0.68	0.76	0.029	170	50
yob(3)*Sex	0.44	0.52	0.0095	249	22	0.73	0.83	0.028	250	34

Table 2 continued

	Age 25–55					Age 25–75				
	ρ	δ	Adj R^2	Size	N	ρ	δ	Adj R^2	Size	N
yob(4)*Sex	0.53	0.69	0.0090	342	16	0.77	0.90	0.028	327	26
yob(5)*Sex	0.55	0.74	0.0084	456	12	0.79	0.93	0.027	424	20
yob(10)*Sex	0.77	1.14	0.0078	912	6	0.86	1.02	0.026	849	10
yob(1)*Ed	0.89	0.87	0.096	58.8	93	0.89	0.86	0.11	55.5	153
(c) Greece										
Panel	0.58	0.56				0.59	0.57			
	(0.52, 0.64)	(0.50, 0.63)				(0.55, 0.63)	(0.52, 0.61)			
yob(2)	0.034	0.032	0.0085	208	15	0.25	0.24	0.016	213	25
yob(5)*Sex	0.056	0.0065	0.0083	260	12	0.32	0.28	0.017	267	20
yob(1)*Ed	0.75	0.77	0.095	33.6	93	0.76	0.73	0.11	34.9	153

The estimated ρs for France and Poland are the average of each three year panel from 2004–2007 to 2013–2016. Due to data limitations the estimates for Greece are the average of each three year panel from 2005–2008 to 2013–2016. The Adj R^2 column reports the average adjusted R^2 over all years resulting from regressing $\ln(y_1)$ on cohort dummies for the full sample. Size refers to the average number of observations per cohort. All estimates are weighted, but not adjusted for survey design

The findings for cohorts defined by yob and sex are broadly in line with those defined by yob alone. Cohort definitions with 100 or more observations on average come closest to approximating the true $\rho_{y_{i1}y_{i2}}$ and the range of values is quite narrow. The exception to this, is the estimates for Poland for the 25-55 age range. Despite having average cohort sizes greater than 100, a comparable adjusted R^2 to the yob only cohorts, and a number of cohorts which is sufficient to produce accurate approximates for other age ranges and country settings, the estimates for Poland using the cohorts defined by yob*sex exhibit a high degree of variability. At least in the case of the 25-55 age range for Poland, pseudo-panel estimates appear to be sensitive to the cohort definition even when the average cohort size exceeds 100.

Cohorts defined by yob*education provide an accurate approximate of the average $\rho_{y_{i1}y_{i2}}$ for France, but not for Poland or Greece. This is despite: the adjusted R^2 being broadly similar in all three countries, the average cohort size being larger in Poland, and the number of cohorts being almost identical. This, again, highlights the concern raised by HJ—it is not that pseudo-panel techniques cannot accurately approximate ρ, but rather that in the absence of panel data it is not possible to identify when and which pseudo-panel estimates are accurate.

Turning to the indirect approach to approximating $\rho_{y_{i1}y_{i2}}$, there appears to be little difference in the two approaches in the context of France. The approximates for $\rho_{y_{i1}y_{i2}}$ and δ exhibit a comparable level of accuracy. In Poland, however, there is a large discrepancy between the true δ and the pseudo-panel approximates; this discrepancy is much less pronounced for the true $\rho_{y_{i1}y_{i2}}$. It appears that in the case of Poland $\frac{var(y_{c1})}{var(y_{c2})} \neq \frac{var(y_{i1})}{var(y_{i2})}$, however, using the true $\frac{var(y_{i1})}{var(y_{i2})}$ does not, as hypothesized, increase accuracy.

The findings for France and Poland appear to support the use of pseudo-panel approximates for the 25-75 age range. Using estimates derived from cohort definitions with an average cohort size greater than 100 and cohorts defined by sex*yob, one can arrive at potentially useful estimates for parametric bounds. This finding does not, however, hold for Greece where no cohort definition approximates the true $\rho_{y_{i1}y_{i2}}$. This poor performance can to some extent be identified through the inspection of the adjusted R^2 which suggests the cohort definitions based upon yob and yob*sex are weak instruments.

Table 2 compares the average $\rho_{y_{i1}y_{i2}}$ over the time period considered to the average pseudo-panel approximates, however, cohort definitions which perform well on average may perform poorly for specific time periods and may fail to track changes in the true $\rho_{y_{i1}y_{i2}}$ over time. Figure 8 compares the true $\rho_{y_{i1}y_{i2}}$ with the best performing cohort approximations from each category of cohort definition for each country and age range. For both France and Poland, the true panel and $\rho_{y_{i1}y_{i2}}$ exhibit little variation over time. Approximates for $\rho_{y_{i1}y_{i2}}$ exhibit considerably more variation. Some cohort definitions produce approximates for $\rho_{y_{i1}y_{i2}}$ which are almost 0.2 from the true value for France and 0.5 in Poland for certain years. To put such differences in perspective, in Fig. 3 the lower parametric bound uses a value for $\rho_{y_{i1}y_{i2}}$ which is approximately 0.2 points lower than the true $\rho_{y_{i1}y_{i2}}$. Using this lower bound produces synthetic panel estimates which, not only, lie outside the 95% confidence interval of the true poverty

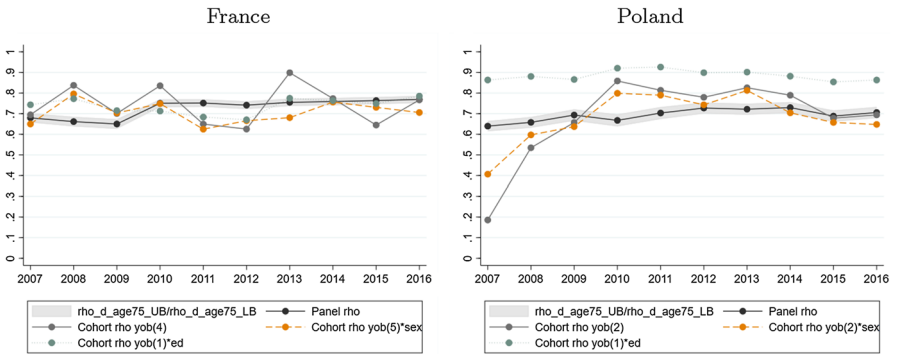

Fig. 8 ρ comparison—3 year panel. *Notes* Age range 25–75. The estimates above 2007 represent the δ for the period 2004–2007

dynamics, but can also be almost twice as large as the true joint probability as is the case for the probability of being (non-poor, poor) in 2011.

HJ suggest that where possible practitioners should use the average pseudo-panel approximated calculated from a series of cross sections. Given that the true $\rho_{y_{i1}y_{i2}}$ appears be quite stable over time, there is merit to this suggestion. As is evidenced in Fig. 8, the average will, however, be sensitive to the time period used to calculate the average. Furthermore, using average pseudo-panel estimates requires repeated cross-sectional data which is a significant data requirement and one which is likely to become less feasible if one wishes to examine longer-run poverty dynamics.

Approximates, even on average, are highly sensitive to the choice of cohort definition. The problem for potential practitioners is how to identify the best performing cohort definition. Within a given category of cohort definition, the optimal choice will depend upon the average cohort size, number of cohorts and the prevailing adjusted R^2, however, the relationship varies from country to country and there does not appear to be one statistic or combination of statistics which successfully identifies the best performing cohort definition.

Pseudo-panel techniques are intended to track convergence in the absence of panel data, however, in the absence of panel data it is not possible to identify the best performing cohort definition. In this key sense pseudo-panel techniques do not appear to be a viable alternative to true panel data and great caution should be exercised when drawing conclusions from such techniques.

5.2 Neighbouring countries and $\rho_{y_{i1}y_{i2}}$

In DLLM, the authors suggest that $\rho_{y_{i1}y_{i2}}$ estimates from neighbouring countries could be used as parametric bounds for the estimates of synthetic poverty dynamics. While neighbouring countries may not exhibit similar rates of correlation, countries with similar institutional characteristics might. The modified varieties of capitalism (VoC) framework, proposed by Amable (2003), groups countries which share key institutional characteristics in terms of welfare, education and health. Amable (2003)

identifies 6 varieties of capitalism in Europe; social democracy, corporatist, liberal, Southern European, and Eastern European.

Table 3 presents the findings for $\rho_{y_{i1}y_{i2}}$ grouped by variety of capitalism. The framework appears to be highly successfully at grouping countries which exhibit similar levels and patterns of $\rho_{y_{i1}y_{i2}}$ overtime. Greece, however, is an exception. Despite sharing certain institutional features, Spain and Italy do not provide an accurate approximate of $\rho_{y_{i1}y_{i2}}$ for Greece. This is likely due to the stark differences in the economic conditions across these countries during the time period considered.

Given the inaccuracies of pseudo-panel techniques when approximating $\rho_{y_{i1}y_{i2}}$, the VoC approach offers some hope to practitioners if panel data is available for a nearby country which shares certain institutional characteristics and contains comparable income or consumption measures. These requirements will, however, make this approach impractical in many settings.

5.3 Approximating $\rho_{y_{i1}y_{i2}}$ using shorter panel data

The discussion of $\rho_{y_{i1}y_{i2}}$ thus far has focused on how to approximate $\rho_{y_{i1}y_{i2}}$ in the absence of panel data. This is clearly a pressing issue for practitioners, but does not address the question of how to approximate $\rho_{y_{i1}y_{i2}}$ beyond what is available from current panel data.[18]

Let us consider a situation where we wish to approximate ρ between 2004-2010. Using the 2011 longitudinal release of EU-SILC, it is possible to calculate the household level income correlation 2007-2010. This needs to be extended in order approximate income correlation for the 2004-2010. An examination of the development of $\rho_{y_{i1}y_{i2}}$ for the countries included in EU-SILC shows that income correlation typically declines over time and that the percentage decline in ρ as the panel is extended tends to decrease (both these features are evident in Table 3). The point estimate for $\rho_{y_{i2007}y_{i2010}}$ thus provides a useful upper bound estimate for $\rho_{y_{i2004}y_{i2010}}$.

In order to arrive at a lower bound and point estimate for $\rho_{y_{i2004}y_{i2010}}$, one must consider alternative approaches to depreciating $\rho_{y_{i2007}y_{i2010}}$ overtime. Using 2011L, one potential rate of depreciation can be calculated as:

$$\lambda = \frac{\rho_{y2007\,y2010} - \rho_{y2007\,y2009}}{\rho_{y2007\,y2009}} \tag{10}$$

If, for example, $\rho_{y_{i1}y_{i2}}$ declined by 5 % when moving from $\rho_{y2007\,y2009}$ to $\rho_{y2007\,y2010}$ then one could depreciate $\rho_{y07\,y10}$ by 5 % to arrive at an estimate for $\rho_{y06\,y10}$. To arrive at an estimate for $\rho_{y05\,y10}$, one could then take the estimate for $\rho_{y06\,y10}$ and depreciate it again by the same rate of depreciation. This assumes that $\rho_{y_{i1}y_{i2}}$ is log-linear over time.

An improved approach is to calculate the rate of depreciation in $\rho_{y_{i1}y_{i2}}$ using alternative longitudinal data sets which cover the earlier time period. For estimating the rate of depreciation in income correlation between 2007 and 2006, the 2010 longitu-

[18] In this context $\rho_{y_{i1}y_{i2}}$ is taken to indicate the correlation across any two time periods rather than 1 year income correlation.

Table 3 ρ by varieties of capitalism

	Social Dem.		Corporatist		Liberal		S. Europe		E. Europe		
	Norway	Sweden	Austria	France	UK	Ireland	Italy	Spain	Slovakia	Poland	Greece
1	0.87	0.86	0.82	0.83	0.65	0.74	0.78	0.74	0.81	0.85	0.63
	(0.85, 0.88)	(0.84, 88)	(0.8, 0.84)	(0.83, 0.84)	(0.61, 0.68)	(0.69, 0.78)	(0.77, 0.79)	(0.72, 0.76)	(0.79, 0.83)	(0.84, 0.86)	(0.59, 0.66)
2	0.75	0.80	0.77	0.81	0.65	0.68	0.72	0.70	0.78	0.77	0.63
	(0.73, 0.77)	(0.78, 0.82)	(0.74, 0.79)	(0.8, 0.82)	(0.62, 0.68)	(0.63, 0.74)	(0.71, 0.74)	(0.67, 0.72)	(0.75, 0.8)	(0.75, 0.79)	(0.59, 0.66)
3	0.72	0.74	0.77	0.78	0.59	0.64	0.68	0.69	0.71	0.72	0.53
	(0.69, 0.74)	(0.71, 0.77)	(0.74, 0.79)	(0.77, 0.79)	(0.55, 0.63)	(0.57, 0.69)	(0.66, 0.7)	(0.67, 0.71)	(0.67, 0.73)	(0.7, 0.74)	(0.49, 0.57)

The rho estimates reported are the 1–3 income correlations estimated using the 2014 longitudinal release of EU-SILC. All estimates are weighted, but not adjusted for survey design. Only households whose head is aged 25–75 in the base year are included

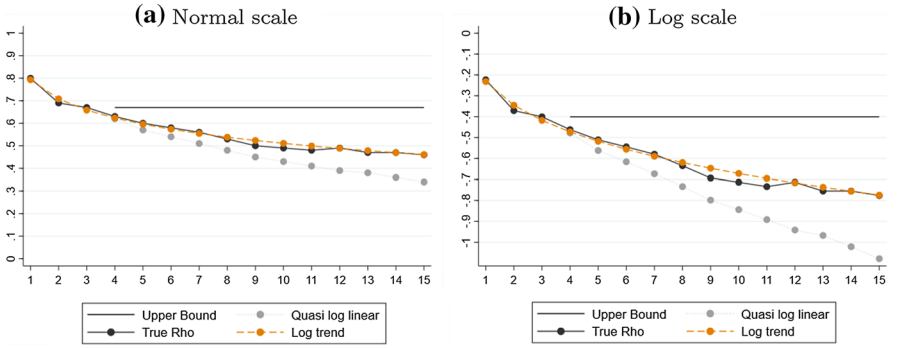

Fig. 9 Predicted income correlation Ball (2016)

dinal release can be used to calculate a new λ. Given that this rate of depreciation will directly relate to the economic conditions of the period of interest, it will more closely resemble the true rate of depreciation—namely the depreciation between $\rho_{y_{06}y_{10}}$ and $\rho_{y_{07}y_{10}}$. This no longer assumes $\rho_{y_{i1}y_{i2}}$ is log-linear over time, but rather implies quasi log-linearity.

The appropriateness of the quasi log-linear approach can be further examined using the correlation data provided in Ball (2016). Ball (2016) performs an extensive analysis of the correlation of individual income in New Zealand using Inland Revenue Department (IRD) tax data. The IRD data reports income on all people in New Zealand who have reported income to the tax authorities since 2000.

Figure 9 applies the quasi log-linear approaches to New Zealand data. Treating 2000 as the base year, one can use the later releases to estimate the rate of depreciation. In order to match the data availability of EU-SILC, the analysis is limited to panels spanning 4 years, with each "panel" producing 1–3 year estimates of ρ). Figure 9 presents the findings. The quasi log linear line becomes less accurate as the panel is extended.

The decreasing rate of depreciation observed in Table 3 as well as the pattern of sharp decreases in early periods followed by stagnant $\rho_{y_{i1}y_{i2}}$ found for the New Zealand data, indicates that the rate of depreciation may be better approximated by a logarithmic trendline:

$$\rho_{y_{i1}y_{i2}} = \beta_0 + \beta_1 ln(period) \tag{11}$$

where period indicates the panel length. This simply fits the logarithmic trend which best explains the observed depreciation in the true correlation and extrapolates beyond the available panel length.

While the quasi log-linear line becomes less accurate over time in Fig. 9, the logarithmic trendline is highly accurate throughout. Figure 9 provides support for using a logarithmic trend for point estimates, assuming quasi log-linearity for a lower bound and using the last available true estimate of $\rho_{y_{i1}y_{i2}}$ as an upper bound. This support comes with the caveat that the findings for New Zealand are at the individual level rather than at the household level.

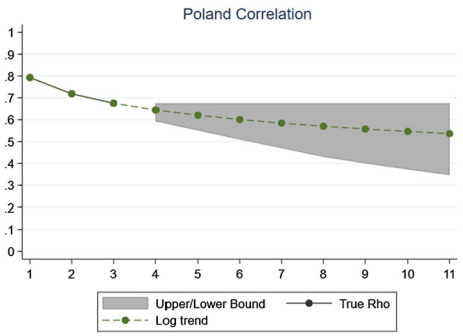

Fig. 10 Income correlation Poland. Notes. Age range 25–55

Figure 10 presents the findings from applying this approach to EU-SILC data for Poland. Given the pattern in $\rho_{y_{i1}y_{i2}}$ found in the longitudinal element of EU-SILC, the logarithmic trend provides a reasonable approximation. Extending ρ using shorter panels does, however, requires certain assumptions about the behaviour of ρ in the longer term. The upper and lower bound assumptions find strong support in the limited panels of EU-SILC and should, therefore, always be shown. The logarithmic trendline requires a stronger assumption and results using this approach should be interpreted with greater caution. Practitioners should also keep in mind the purpose of their analysis when extending ρ. If the intention, is to rank sub-populations within a country in terms of chronic poverty then using the logarithmic trendline should be sufficient as the same assumption will be applied to all groups and the ranking will be unaffected. Cross-national comparisons, however, would require that the logarithmic assumption holds in both countries which may be too strong an assumption. In such cases, the bounds may be more useful.

6 Conclusion

This paper examines the performance of the DL synthetic panel approach for approximating poverty transitions and the means based pseudo-panel approach for establishing whether incomes have been diverging or converging over time. Accuracy is primarily measured by counting how many of the synthetic panel estimates lie within the confidence interval of the true estimate. The shortcomings of such a benchmark are discussed in Sect. 2.

When analysing the performance of the DL approach two key questions are considered, does the approach work when the true $\rho_{y_{i1}y_{i2}}$ is known? Can the true $\rho_{y_{i1}y_{i2}}$ be accurately approximated using pseudo-panel techniques?

The answer to the first of these questions is a conditional yes. The DL approach is reasonably accurate at estimating joint and conditional poverty probabilities when the true $\rho_{y_{i1}y_{i2}}$ is known. Key to the accuracy of the synthetic panel estimates is the normality of the residuals from the respective income models. Inspecting these residuals and taking measures to improve the assumption of normality can have a significant positive impact on the accuracy of estimates. This suggests that a new step should be added to the standard DL approach. Following the estimation of the income

models and prior to the prediction of income, the residuals should be inspected and steps taken to improve their normality.

In light of this additional step, the accuracy of synthetic panel estimates is insensitive to the income model and age range selected, but does deteriorate as the poverty line increases. The example of urban and rural poverty dynamics for Poland highlights that there is some benefit in estimating separate income models for different subpopulations.

Given the conclusions from the income model component of the DL approach, analysis turns to the second question. While pseudo-panel techniques are capable of accurately approximating $\rho_{y_{i1}y_{i2}}$, estimates are particularly sensitive to the cohort definition. Furthermore, in the absence of panel data, there is no statistic or combination of statistics which successfully identify the best performing cohort definition. This raises serious concerns as to the usefulness of the DL approximate for $\rho_{y_{i1}y_{i2}}$ and the means-based approach for approximating δ when household income is the variable of interest.

Despite the inaccuracy of the DL approximate for $\rho_{y_{i1}y_{i2}}$, there are still a number of settings where the DL synthetic panel approach can be of use. In the case of EU-SILC and other panel surveys with rotating panel designs, the DL approach can offer an alternative set of poverty dynamic estimates which are at least partially protected against the effects of attrition, has a much larger sample size enhancing the reliability and feasibility of sub-population analysis and can avail of variables present only in the cross-sectional data (such as EU-SILC's ad hoc modules).[19] Furthermore, if limited panel data is available, parametric bounds and point estimates can be estimated under certain assumptions concerning the development of $\rho_{y_{i1}y_{i2}}$ overtime. This opens the door for longer run analysis of poverty dynamics to take place.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00181-022-02277-7.

Funding The author have no relevant financial or non-financial interests to disc lose.

Availability of data and material In order to access the European Union Survey on Income and Living Conditions (EU-SILC), researchers must agree to confidentiality and data security requirements. Among these data restrictions is that the data cannot be shared with individuals outside the research project. For this reason the datasets generated and analysed during the current study cannot be made publicly available. A simplified version of the do files used to generate the findings can be made available upon request. For further information with regards to accessing the EU-SILC data please use the following link: https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

Declarations

Conflict of interest The author have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article.

[19] The DL method can only be said to partially account for attrition as the true panel correlation term may still be effected by non-random attrition. Visual inspection of the relationship between retention rates and correlation indicate no clear pattern over time or across countries. It is not clear that countries with higher rates of attrition systematically experience higher or lower correlation over time.

References

Amable B (2003) The diversity of modern capitalism. Oxford University Press on Demand

Antman F, McKenzie DJ (2007) Earnings mobility and measurement error: a pseudo-panel approach. Econ Dev Cult Change 56(1):125–161

Ashenfelter O, Deaton A, Solon G (1986) Collecting panel data in developing countries: Does it make sense? living standards measurement study working paper no. 23. ERIC

Ball, C. (2016). Estimating income dynamics from cross-sectional data using matching techniques. Working Papers in Public Finance 06

Bourguignon F, Moreno, H (2020) On the construction of synthetic panels. HAL SHS working paper

Cruces G, Lanjouw P, Lucchetti L, Perova E, Vakis R, Viollaz M (2015) Estimating poverty transitions using repeated cross-sections: A three-country validation exercise. J Econ Inequal 13(2):161–179

Dang H-AH, Dabalen AL (2019) Is poverty in africa mostly chronic or transient? evidence from synthetic panel data. J Dev Stud 55(7):1527–1547

Dang H-AH, Ianchovichina E (2018) Welfare dynamics with synthetic panels: the case of the arab world in transition. Rev Income Wealth 64:S114–S144

Dang H-AH, Lanjouw PF (2018) Poverty dynamics in India between 2004 and 2012: insights from longitudinal analysis using synthetic panel data. Econ Dev Cult Change 67(1):131–170

Dang H-A, Lanjouw P (2021) Measuring poverty dynamics with synthetic panels based on repeated cross-sections. mimeo, update of Dang and Lanjouw (2013)

Dang H-A, Lanjouw P, Luoto J, McKenzie D (2014) Using repeated cross-sections to explore movements into and out of poverty. J Dev Econ 107:112–128

Deaton A (1985) Panel data from time series of cross-sections. J Econom 30(1–2):109–126

Devereux PJ (2007) Small-sample bias in synthetic cohort models of labor supply. J Appl Econom 22(4):839–848

Ferreira FH, Messina J, Rigolini J, López-Calva L-F, Lugo MA, Vakis R, Ló LF, et al (2012) Economic mobility and the rise of the latin american middle class. World Bank Publications

Fields G, Viollaz M (2013) Can the limitations of panel datasets be overcome by using pseudo-panels to estimate income mobility? Universidad Cornell-CEDLAS

Garcés Urzainqui D (2017) Poverty transitions without panel data? an appraisal of synthetic panel methods. In: Seventh meeting of the society for the study of economic inequality (ecineq). retrieved from http://www.ecineq.org/ecineqnyc17/filesx2017/cr2/p447.pdf

Hérault N, Jenkins SP (2019) How valid are synthetic panel estimates of poverty dynamics? J Econ Inequal 17(1):51–76

Howes S, Lanjouw JO (1998) Does sample design matter for poverty rate comparisons? Rev Income Wealth 44(1):99–109

Iacovou M, Kaminska O, Levy H (2012) Using eu-silc data for cross-national analysis: strengths, problems and recommendations. ISER working paper series

Jenkins SP (2010) The british household panel survey and its income data

Jenkins SP, Van Kerm P (2017) How does attrition affect estimates of persistent poverty rates? The case of eu-silc. Monitoring social inclusion in Europe, 401

Kroh M, Spieß M (2006) Documentation of sample sizes and panel attrition in the german socio economic panel (soep)(1984 until 2005). DIW Data Documentation

Lynn P, Buck N, Burton J, Laurie H, Urhig NS (2006) Quality profile: British household panel survey. ISER, Uni& versity of Essex, Colchester

OECD. (2019). Poverty rate. In https://data.oecd.org/inequality/poverty-rate.html

Perez V et al (2015) Moving in and out of poverty in mexico: What can we learn from pseudopanel methods? Understanding Society at the Institute for Social and Economic Research
Van Kerm P, Alperin MNP (2013) Inequality, growth and mobility: the intertemporal distribution of income in european countries 2003–2007. Econ Model 35:931–939
Verbeek M (2008) Pseudo-panels and repeated cross-sections. The econometrics of panel data. Springer, Berlin, pp 369–383
Verbeek M, Nijman T (1992) Can cohort data be treated as genuine panel data? In Panel data analysis. Springer, Berlin, pp 9–23
Verbeek M, Vella F (2005) Estimating dynamic models from repeated cross-sections. J Econom 127(1):83–102
Watson N, Wooden M (2006) Modelling longitudinal survey response: the experience of the hilda survey. In: Acspri social science methodology conference. pp 10–13
World Bank D (2014) Measured approach to ending poverty and boosting shared prosperity: concepts, data, and the twin goals. World Bank Publications