Stress-Testing Survey-to-Survey Imputation: Understanding When Poverty Predictions Can Fail*

Paul Corral, Andres Ham, Peter Lanjouw, Leonardo Lucchetti, and Henry Stemmler†

July 22, 2025

Abstract

Accurate and timely poverty measurement is central to development policy, yet the availability of up-to-date high-quality household survey data remains limited—particularly in countries where poverty is most concentrated. Survey-to-survey (S2S) imputation has emerged as a practical response to this challenge, allowing practitioners to update poverty estimates using recent surveys that lack direct welfare measures by borrowing information from other comprehensive surveys. A critical review of the method is provided, revisiting its statistical underpinnings and testing its limitations through extensive model-based simulations. Through model-based simulations, the analysis demonstrates how violations of parameter stability, omitted variable bias, and shifts in survey design can introduce substantial errors—particularly when imputing across time or under economic and structural change. Results show that standard corrections such as re-weighting or covariate standardization may fail to eliminate these biases, especially when imputing across time or under structural change. The performance of alternative model specifications is also evaluated under various methods, including performance under heteroskedastic errors, non-normality. The findings offer practical guidance for practitioners on when S2S imputation is likely to succeed, when it should be reconsidered, and how to communicate its limitations transparently in the context of poverty monitoring and policy design.

Key words: Poverty, Inequality, Poverty imputation, missing data

JEL classification: I32, C53

*The authors acknowledge financial support from the World Bank. We thank Chris Elbers, Roy van der Weide, Sergio Olivieri, Hai-Anh Dang, and Nobuo Yoshida for comments. We also thank participants of the S2S workshop held at the World Bank. Additionally, we thank Luis-Felipe Lopez Calva, Gabriela Inchauste, and Maria Eugenia Genoni for providing support and space to pursue this work. Any error or omission is the authors' responsibility alone.

†Paul Corral, Leonardo Lucchetti, and Henry Stemmler are part of the World Bank. Andres Ham is of Universidad de los Andes, and Peter Lanjouw is part of Vrije Universiteit Amsterdam. Corresponding author: Paul Andres Corral Rodas, pcorralrodas@worldbank.org.

1 Introduction

The measurement and monitoring of poverty are central to assessing global development progress. For the World Bank, whose mission is to eradicate extreme poverty on a livable planet, the ability to track poverty reduction is fundamental for measuring institutional effectiveness and guiding policy decisions. Yet, a persistent challenge hampers this crucial task: the limited availability of recent, high-quality household survey data that includes information that allows for comparable measures of expenditure or income across time. This challenge is particularly acute in countries where poverty is most concentrated (Dang et al., 2017; World Bank, 2024).

The issue affects many countries, and hinders poverty monitoring due to the lack of data in some of the world's most populous and poverty afflicted economies. For example, India and Nigeria account for a substantial share of the world's poor, meaning that they heavily influence global poverty trends. In India, official consumption survey data was unavailable between 2011 and 2022.[1] In Nigeria comparable survey data for poverty monitoring was absent between 2009 and 2018, and as of 2025 has no new publicly available consumption survey data. These data gaps not only affect our understanding of poverty at the country level but compromise our ability to produce reliable global poverty estimates and assess progress toward poverty reduction goals.

Traditional poverty measurement relies on household surveys that collect detailed consumption or income data. These surveys represent substantial investments in both financial and human resources. They require extensive preparation, careful implementation, and can place considerable burden on responding households, who must either maintain detailed consumption diaries or participate in comprehensive recall interviews. The method of data collection itself can introduce significant biases into poverty estimates which may compromise comparability of estimates over time.[2] These methodological challenges add another layer of complexity to the already demanding task of poverty measurement.

The challenges of collecting traditional living standards surveys for poverty measurement become particularly acute during crises when standard survey operations are disrupted or impossible to conduct. Armed conflicts, natural disasters, health emergencies, and other humanitarian crises often prevent face-to-face household interviews precisely when poverty monitoring becomes most crucial. The COVID-19 pandemic provided a stark illustration of this challenge, as household survey efforts were halted globally at a time when understanding welfare impacts was most critical. Similar data collection constraints arise in conflict zones, where security concerns prevent enumerator access to households, or during natural disasters that displace populations and disrupt statistical operations. During such crises, policymakers and international organizations must resort to alternative methods for estimating poverty.

Survey-to-survey (S2S) imputation has emerged as a methodological response to these data limitations. The method is often used as a way to produce poverty numbers when traditional welfare data collection is unavailable due to costs, complexity, conflict, or other reasons. S2S, often combined with rapid phone surveys or other alternative data collection methods, have emerged as a key tool for maintaining poverty monitoring during these challenging periods where traditional surveys cannot be collected.

S2S methods build upon techniques originally developed for small area estimation (SAE) in poverty mapping, pioneered by Hentschel et al. (1998) and further refined by Elbers et al. (2003). The method

[1] The 75th round of the NSO Survey of Consumption Expenditure was collected in 2017-18 but was not released until after 2024: https://www.mospi.gov.in/unit-level-data-report-nss-75th-round-july-2017-june-2018-schedule-250social-consumption-health.

[2] For instance, consumption diaries – while theoretically more accurate – can lead to respondent fatigue and underreporting over time. Recall modules, on the other hand, may suffer from memory bias, with longer recall periods typically resulting in lower reported consumption. Research has shown that simply changing the recall period or the number of consumption items can lead to substantially different poverty estimates within the same population (Beegle et al. 2012).

imputation enables poverty estimation using surveys that lack direct welfare measures by:

1. Developing a predictive model of household welfare using a survey with consumption or income data (the "source" survey)
2. Applying this model to a different survey containing similar household characteristics but lacking direct welfare (the “target” survey)
3. Generating poverty estimates based on the predicted welfare distribution

While S2S and small area estimation share methodological foundations, they serve distinct purposes. Unit-level small area estimation of poverty, as developed by Elbers et al. (2003) and advanced by Molina and Rao (2010), typically applies nationally estimated models (from a household survey) to population census data to generate precise poverty estimates for small geographic areas where survey estimates are often too imprecise for reliable estimates. In contrast, S2S focuses on applying the model to a separate survey to predict welfare, regardless of geographic disaggregation. One might view traditional poverty mapping as a special case of S2S where the target dataset are the different areas of the census and the primary goal is obtaining geographically disaggregated estimates.

Although a potentially an attractive solution, S2S approaches come with their own limitations and potential biases that must be carefully considered. This paper first examines the limitations and potential pitfalls of survey-to-survey imputation through rigorous analysis of its fundamental assumptions. This is a technical paper that relies on simulated data to conduct experiments. In these experiments, the focus is not on traditional concerns such as variable selection or model specification, but rather on the method’s core limitations, particularly for measuring poverty across extended time periods or different populations. The analysis relies on simulated data to isolate and demonstrate specific mechanisms that can generate biased estimates. This approach allows for clear illustration of three critical findings:

1. Standard sampling bias-correction techniques, such as re-weighting to match population means, may be insufficient when source and target surveys differ fundamentally.
2. The method’s tendency to replicate the welfare distribution of the source survey can make it unreliable for measuring changes in inequality and poverty.[3] Thus, better communication behind the limitations of the methods are crucial. These insights are particularly relevant given the increasing reliance on survey-to-survey imputation to fill data gaps in poverty monitoring.
3. Third, omitted variable bias will likely affect poverty predictions, particularly when imputing across time periods that include significant economic shocks or structural changes. This last finding has important implications for applications that use "fast-moving" variables to capture welfare changes,[4] as these variables may introduce bias if they are correlated with unobserved factors that affect welfare.

Through systematic examination of these limitations using simplified examples, the paper provides practitioners with a framework for understanding when and why S2S methods can fail. The results here are of relevance since survey-to-survey imputation has become more prevalent in recent times to overcome data scarcity.

[3] The model captures the joint distribution between y and x in the source data. The parameters corresponding to the source data are then applied to the target data, where the only difference is the distribution of covariates, x. Thus, unless all the change is captured by changes in the covariates the method will likely give biased estimates.

[4] Yoshida et al. (2021) refer to expenditure related covariates as fast moving. These include dichotomous variables indicating if a household reported purchasing a product or not.

2 The effectiveness of S2S across time: A summary of the evidence

Estimating poverty relies on household survey data that collects household consumption or income. Collecting household consumption and income data is costly and complex. For consumption, modules typically span multiple domains—food, housing, health, education, and more—each requiring detailed recall and careful categorization. These surveys are usually fielded over a full year to represent seasonal consumption patterns. As Deaton, Grosh, et al. (1998) observe, these modules are among the most demanding components of household surveys, often encompassing hundreds of items and multiple recall periods. This not only increases respondent burden but also drives up implementation costs. Ensuring accuracy adds another layer of complexity: respondents may forget or misreport expenditures, and cultural and regional differences often require localized adaptations in survey design. Moreover, as Beegle et al. (2012) show, even small changes in how surveys are administered can yield significant differences in reported poverty and consumption levels.

Because of the complexity and cost, few countries collect comparable consumption or income data frequently. Long-term, harmonized consumption trends are the exception, not the norm. Yet, policymakers, donors, and development practitioners continually need up-to-date poverty statistics. To fill this gap, survey-to-survey imputation has become an increasingly used method for estimating poverty when consumption or income data are unavailable.

S2S builds upon the idea proposed by Elbers et al. (2003) and Hentschel et al. (1998) where a linear regression model is fit on a survey where consumption or income are available to determine the joint distribution of consumption or income, and a given set of household characteristics – this survey is typically referred to as the source or training data. The model parameters are then applied to a separate dataset (the target data) where the same set of household characteristics can be found with the aim of replicating the welfare distribution (consumption or income) via imputation. With a simulated welfare distribution in hand, poverty estimates and other welfare indicators can be derived from the target data. The methods can be applied to contemporaneous surveys or to surveys corresponding to a different point in time.

S2S is useful when applied to contemporaneous surveys to obtain statistics across the welfare distribution for indicators that are not available in the living standards survey. For example, the Demographic and Health Surveys (DHS) provide a litany of health related indicators that are not available in most living standards surveys that collect welfare for poverty estimation. This is where S2S comes into play. A model trained on the living standards survey and applied to the DHS can help us inform about the difference in stunting rates between poor and non-poor households or determine how it varies across the welfare distribution.

The method is often treated as a prediction exercise and unlike much of econometrics, the parameters of the linear regression are not intended to capture only the direct effect of the characteristics on welfare (Elbers et al. 2003, p356). Nevertheless, the original intent of Elbers et al. (2003) was to predict on to the same population.

When imputing to a different year, the method rests on the assumption that model parameters remain stable—that is, the relationship between household characteristics and welfare has not changed. In other words, the method assumes that model parameters remain stationary over time – meaning that any observed changes in poverty are solely attributable to changes in the model's covariates, rather than shifts in unobservable factors or changing returns to these covariates (Dang, Lanjouw and Serajuddin

2019). As Christiaensen et al. (2012) note, such assumptions may be especially problematic in rapidly growing economies, where structural economic changes can alter the relationship between poverty and its predictors. It can also be problematic when a strong shock has occurred between the survey periods – for example, a global pandemic or changes in a country's fragility. Mathiassen (2009) emphasizes that the validity of imputations rests on a strong assumption of parameter stability and explicitly warns against extrapolating too far across time or applying the method when significant changes in welfare dynamics or survey design are likely.

Researchers and practitioners have proposed various approaches to address the limitations of S2S in predicting poverty over time. Stifel and Christiaensen (2007) advocate for including time-varying variables such as rainfall and prices to capture temporal changes. Yoshida et al. (2021) suggest incorporating variables that track economic conditions more directly, like indicators on whether the household consumed a given good. However, as Yoshida et al. (2021) emphasize, without updated training data to re-estimate model parameters, the risk of missing significant economic changes remains substantial, even with these additional covariates. Dang et al. (2025) further note that evidence on the utility of granular consumption predictors remains inconclusive, and instead advocate for parsimonious models that are easier to implement in surveys and in challenging field conditions.

The effectiveness of S2S imputation across time has been tested in a range of contexts, with mixed results. For example, Dang, Lanjouw and Serajuddin (2019) evaluate S2S using Jordan's Household Expenditure and Income Survey and the Employment and Unemployment Survey; Stifel and Christiaensen (2007) conduct similar tests for Kenya; and Dang et al. (2025) apply the method across multiple countries—including Ethiopia, Malawi, Nigeria, Tanzania, and Vietnam—reporting estimates generally within acceptable error margins. While some studies report estimates within acceptable margins of error (Dang et al. 2025, Mathiassen 2013), others find significant discrepancies between predicted and observed poverty rates (Newhouse et al. 2014, Mathiassen 2013). A useful example comes from Malawi, where Mathiassen and Wold (2021) document their efforts to produce comparable poverty estimates between 2004 and 2014. While their method performed well within individual surveys and with careful seasonal adjustment, comparability across survey rounds proved fragile. Shifts in survey design and implementation such as shorter questionnaires, periods of data collection, changing field protocols, and others led to unstable predictor relationships and inconsistent poverty trends over time. Even in this relatively well-structured case, the authors conclude that S2S imputation should only be used when comparability of predictors and survey instruments can be assured (Mathiassen and Wold 2021).

Christiaensen et al. (2012) undertakes an empirical validation of survey-to-survey imputation methods over time. The authors perform survey-to-survey imputation in scenarios where there is comparable expenditure data which provides a "true" estimate of poverty. The authors validate their approach with data for Vietnam and for China using rural household panel data. In Vietnam, the authors obtain a model using the 1992/93 data and predict poverty using the 1997/98 data. They note that the method works relatively well and depending on the covariates used, differences between predicted and observed poverty rates were on average 3.4 percentage points during a period where poverty fell by 23.2 percentage points. For the Chinese regions where the method was tested, the authors also find that the methods work relatively well. However, depending on the model used, differences between predicted and observed rates were considerable. This highlights the method's conundrum: different models may generate different findings, and at least some models may get it right. However, ex-ante, it is very difficult to determine which model will yield the most accurate estimates or even the right estimates. So far, no foolproof methods or diagnostics exist to reliably identify the ideal model, raising fundamental concerns about the robustness of imputed poverty estimates.

Applications of survey-to-survey (S2S) imputation have also made their way into global poverty monitoring—for example, until recently, every poverty number for India after 2011 was based on an imputation. However, evidence from real-world applications, particularly when time gaps are large or covariate distributions change substantially, reveals mixed findings and highlights the need for caution. These findings underscore the importance of understanding the context and data very well when applying S2S across different periods and economic environments. To better appreciate how such estimates are produced and the assumptions they rest on, the next section outlines the foundations of survey-to-survey imputation, tracing its origins in the broader field of missing data methodologies and clarifying its application in the context of poverty measurement.

3 The basics behind survey-to-survey imputation (S2S)

Imputation is a method for filling in missing data. According to Van Buuren (2018), the first instance of a statistical method to replace a missing value dates to 1930,[5] and the first widespread use of the term "imputation" comes from Madow et al. (1983).[6] The method was originally proposed to fill in missing observations and considered the nature of the missing data (Dempster et al. 1977). The goal of multiple imputation is not to create a single prediction, but rather multiple imputations to reflect the uncertainty around the actual value (Van Buuren 2018). By generating plausible values for missing data, multiple imputation allows for valid estimation and the construction of appropriate confidence intervals (Van Buuren, 2018). For example, in the case of a regression on agricultural yields the variable capturing plot sizes may have missing values and these must be imputed with the aim of not losing information as well as obtaining a valid estimate of the coefficient for the relationship between land and yields. Using multiple imputation in this example, as opposed to just the predicted land size, is expected to reduce the rate of false positives.[7]

The case of a variable that is entirely missing in the dataset was not considered by the original multiple imputation literature. The academic background for predicting an entirely missing variable in a given dataset is more aligned to the small area estimation literature (see Rao 2005 and Rao and Molina 2015). Perhaps the first instance of survey-to-survey imputation, as is applied in this paper, comes from the work of Hentschel et al. (1998) which is more aligned to small area estimation and noted by the authors. The key difference to small area estimation up to that point was that Hentschel et al. (1998) predicted the variable of interest, consumption, at the household level and from that they obtained aggregate statistics based on the prediction of consumption. While earlier examples exist of combining data from different sources to predict household-level variables (e.g., Arellano and Meghir 1992), the innovation in Hentschel et al. (1998) lies in applying models estimated from survey data to census data to replicate the full consumption distribution. This approach allows for the estimation of poverty and other welfare indicators as if welfare data were available in the census itself. After refinements by Elbers et al. (2003), the work became the basis of what later was referred to as poverty mapping in the World Bank. The key difference between survey-to-survey imputation and small area estimation is that the latter aims to replicate the welfare distribution for each specific area or group of interest, rather than only for the entire population. One of the first applications of the methods from Elbers et al. (2003) to predict national level poverty can be seen in the work of Simler et al. (2004) who use the methods to track changes in poverty in Mozambique.

Survey-to-survey (S2S) imputation for poverty measurement relies on the assumption that a population's

[5] Allan and Wishart (1930)
[6] As noted by Van Buuren (2018).
[7] Type I errors, which occur when a null hypothesis is incorrectly rejected.

welfare distribution can be captured by a linear model which is estimated on the *source* data. For simplicity, the subscript *target* is used to indicate when parameters are applied to the target data. Hence, the assumed data generating process (DGP) for transformed welfare $\ln y_i$ is:[8]

$$\ln y_i = x_i\beta + e_i;\ e_i \sim N(0, \sigma_e^2) \tag{1}$$

where x_i is a vector of independent variables common to the source survey and the target survey, and e_i is a random disturbance term that is assumed to be distributed i.i.d. and $N(0, \sigma_e^2)$. The β as well as the σ_e^2 parameters are estimated using the source survey or the training data since this is the only data where the welfare vector of interest ($\ln y_i$) is available. These estimated parameters are then applied to the target data to predict $\ln y_i$, and from that poverty or other welfare related indicators.

Because normally distributed errors are assumed in Eq. 1, although this can be relaxed as later illustrated, for any given household i in the target survey, the probability of being poor for household i is entirely dependent on its expected welfare, $x_i\beta$, and the distribution of its idiosyncratic term, e_i, which is assumed to follow $e_i \sim N(0, \sigma_e^2)$

$$Prob\left(poor_i\right) = \Phi\left(\frac{\ln z - x_i\beta}{\sqrt{\sigma_e^2}}\right) \tag{2}$$

where $\ln z$ is the natural log of the poverty line,[9] and Φ is the standard normal distribution. Consequently, what the method calculates is each household's probability of being poor. The average probability of being poor across households corresponds to the national poverty rate.[10]

The implementation of welfare predictions to the target data can follow different approaches. The most common method, grounded in the multiple imputation literature, generates several welfare vectors to reflect prediction uncertainty. This approach is aligned to what was implemented in poverty mapping through the PovMap software (Zhao 2006) and is similar to the methodology used in Stata's multiple imputation commands (`mi regress`). Under this approach, each imputed welfare vector is generated through a five-step process:[11]

1. Obtain model parameter estimates for β, and σ_e^2 via ordinary least squares fit on the training data (Eq. 1), noted below as $\hat{\beta}$ and $\hat{\sigma}_e^2$.

2. The distribution of the error terms is drawn from its posterior Chi-square distribution, where n is the number of observations in the training or source data, and K is the number of covariates in the model:

$$\sigma_e^{2*} \sim \hat{\sigma}_e^2 \frac{(n-K)}{\chi^2_{n-K}},$$

[8] Transformed data is often used as the dependent variable to ensure the model's assumptions hold. For simplicity, throughout this document the assumed transformation is the natural logarithm, although many others are possible. Meeting the model's statistical assumptions is crucial for reliable estimation. Corral et al. (2022) demonstrate that violations of these assumptions, particularly the normality of residuals, can lead to biased poverty estimates. Their analysis shows that appropriate transformation of the dependent variable (household welfare) can significantly reduce such bias. When the standard logarithmic transformation proves insufficient, alternative transformations may be necessary to better approximate normality in the model's error term.

[9] Note that if the transformation of the dependent variable is not the natural logarithm, then this is not valid. The poverty line must be transformed in a similar manner as the dependent variable.

[10] Note that because the only thing differing across households in Eq. 2 are the characteristics, x_i, a proxy means test (PMT) approach that relies only on $x_i\beta$ will yield the same household ranking as the survey-to-survey approach if the same model is used for either. Under a PMT, the threshold is chosen at a given percentile of $x\beta$ so that by construction it yields a desired proportion of people eligible.

[11] Taken from `mi regress` StataCorp (2023)

3. With σ_e^{2*} in hand it is possible to update the variance covariance matrix of the β parameters:

$$\beta^* \sim MVN\left(\hat{\beta}, \sigma_e^{2*}\left(x'x\right)^{-1}\right),$$

4. Then errors are drawn from:

$$e_i^* \sim N\left(0, \sigma_e^{2*}\right)$$

5. Each imputed transformed welfare for household i in the *target* data under imputation m is given by:

$$\ln y_{im}^* = x_{i_{target}}\beta_m^* + e_{im}^* \tag{3}$$

Hence, for each imputed vector, a new σ_e^{2*} is drawn and used to draw β^* and a household specific residual, e_i^*. The multiple imputation simulation approach is not necessarily aligned to reducing the prediction mean squared errors (MSE), but as noted by Van Buuren (2018) the purpose is to minimize the likelihood of false positives, since normally the imputed vectors are frequently used for regression analysis. Alternatively, step 4 may be skipped and the imputed transformed welfare can be obtained from $\ln y_{im}^* \sim N\left(x_{i_{target}}\beta_m^*, \sigma_e^{2*}\right)$.

An alternative approach, derived from the small area estimation literature, treats the source survey as the best representation of the true welfare distribution. Under this approach, parameters estimated from the source survey are applied directly to the target data through Monte Carlo simulation. For straightforward indicators like poverty rates, practitioners can simply apply the asymptotic formula for expected values (Eq. 2). The steps are straight forward:

1. Fit Equation 1 to the training data using ordinary least squares to obtain estimates of the model parameters, $\hat{\beta}$ and $\hat{\sigma}_e^2$.

2. Each transformed welfare for household i in the Monte Carlo simulation is given by – note how the coefficients remain constant across Monte Carlo simulations:[12]

$$\ln y_{im}^* \sim N\left(x_{i_{target}}\hat{\beta}, \hat{\sigma}_e^2\right) \tag{4}$$

Recent methodological advances have expanded model fitting options to include machine learning (ML) techniques. However, since most ML methods do not make explicit assumptions about error term distributions, they cannot directly replicate the multiple imputation approach described above. Nevertheless, bootstrap-based alternatives exist where draws from the empirical residuals are used to obtain a welfare vector ($e_i^* = \hat{e}_j$,where $j \sim \text{Uniform}\{1, 2, \ldots, n\}$). These were implemented in both PovMap (Zhao 2006) and Stata's `sae` command by Nguyen et al. (2018). This bootstrapping approach has proven particularly valuable for regularized regression techniques, as demonstrated by Lucchetti et al. (2024) with lasso regression,[13] and may be extended to other ML methods such as gradient boosting, random forests, or Bayesian additive regression trees.

An additional method that can be paired with machine learning techniques as well as OLS, is predictive mean matching (PMM). The approach compares the predicted values from the target data to those from the training data, specifically by minimizing the absolute difference between linear predictors $|x_{i_{target}}\hat{\beta} - x_i\hat{\beta}|$. Once a specified number of nearest neighbors is identified, one of the actual observed welfare values is randomly drawn (Van Buuren, 2018). PMM is notably less sensitive to model misspecification and is

[12] Monte Carlo simulations are not always necessary and for poverty one could immediately apply Eq.2. However, for more complex parameters Monte Carlo simulations allow for indicators that are difficult to handle analytically.

[13] https://github.com/pcorralrodas/lassopmm

robust to back transformation (ibid). For instance, models estimated on $\log(y)$ typically yield similar results to those based on $\exp(y)$ (Van Buuren 2018), although differences in explained variance can still affect poverty estimates. The number of donors (i.e. neighbors) is determined by the practitioner. While setting this number to one can result in repeated values—especially when model fit is weak—Van Buuren (2018) recommends using five neighbors. He also emphasizes that this choice should be adaptive, depending on the sample size and the model's explanatory power, as measured by its R^2.

3.1 Imputing to a contemporaneous survey

Imputation of poverty to a contemporaneous survey can be a useful tool. For example, assume a country collects a living standard survey annually and has also collected a Demographic and Health Survey (DHS). If one would like to determine how nutritional indicators among children vary across the welfare distribution, survey to survey imputation may be of use to determine this. Nevertheless, practitioners could also impute the indicator of interest to the living standards survey. For example, with the goal of determining energy poverty in Bulgaria, Rude and Robayo (2024) impute energy spending from the household budget survey (HBS) to the European Survey of Income and Living Conditions to determine profiles of the energy poor.[14]

The successful application of survey-to-survey imputation depends on meeting several critical preconditions, first outlined by Hentschel et al. (1998) and later refined by Elbers et al. (2003). These preconditions ensure that the fundamental assumption – both surveys represent the same underlying population – holds true. Having 2 contemporaneous surveys which represent the same population are needed for the approach to work well. In this instance, the tendency of S2S methods to replicate the distribution of the source data in the target data is considered an advantage.

Covariate comparability

The variables used to predict welfare must be present and measured consistently in both source and target surveys, they should be representative of the same population for a given period. This requirement extends beyond simple presence of the covariates to encompass:

1. Identical definitions of key variables
2. Consistent measurement approaches
3. Similar survey implementation protocols

Distribution Alignment

For the method to work effectively, both surveys must exhibit three types of consistency:

1. Similar distributions of predictor variables. Each covariate should have similar summary statistics across surveys. This includes both averages and measures of spread/variation
   (a) Example: The distribution of education levels should be comparable across surveys. Not just means, but other moments should be aligned.
2. Stable structural relationships between variables
   (a) The correlation patterns between predictors should be preserved

[14] These applications resemble a problem considered by Rubin (1986) where no single data set contains the complete set of variables for inference, where imputations are made both ways.

i. Example: If education strongly predicts formal employment in the source survey, this relationship should hold in the target survey. Changes in these relationships may indicate structural economic changes that could invalidate the model's predictions.

ii. For predictions in the same time period this is a desirable feature as it will allow for an accurate replication of the welfare distribution. In the case of imputing to a Demographic and Health Survey (DHS) it would allow for accurate estimation of stunting rates across welfare quantiles.

A consistent definition of covariates across training and target surveys is also crucial. For instance, household size must be defined uniformly across surveys. If one survey counts only members sharing five or more meals per week while another uses a different definition, this violates not just the first requirement (similar distributions) but potentially the second as well, as it affects both the variable's distribution and its relationships with other predictors.

Model Assumptions

The reliability of welfare predictions depends critically on meeting the model's statistical assumptions, particularly:

1. Normal distribution of residuals, although this can be relaxed as noted in the previous section
   (a) When the normality assumptions are violated, Corral et al. (2022) provide evidence that appropriate transformations of the dependent variable may help achieve normality. In cases where transformations prove insufficient, practitioners may draw residuals from their empirical distribution, though this requires the assumption of symmetric errors around zero.
2. Homoscedasticity of error terms, although this can be relaxed by modeling for heteroskedasticity
   (a) Both Elbers et al. (2002) and Harvey (1976) offer methods that allow for modelling heteroskedasticity
3. Linear relationships between predictors and welfare, this too can be relaxed
   (a) Using polynomials in a linear model or the use of methods like random forest or gradient boosting can help relax the linearity assumption

3.2 Imputing across time

Survey-to-survey (S2S) imputation provides a valuable tool for filling gaps in welfare data, but its use over time requires greater caution than for contemporaneous applications as discussed in section 2. The core assumption—that changes in welfare reflect only shifts in observable characteristics, with stable relationships between covariates and welfare—becomes increasingly fragile as the time between surveys grows. Structural economic change, evolving labor markets, and shifting household behaviors can all violate this assumption, leading to imputation errors that distort actual welfare dynamics. Moreover, S2S methods tend to reproduce the welfare distribution of the source survey: coefficients capture the joint distribution of y and x from the source period, and errors are drawn from that same model in the source data. As a result, households with identical characteristics in the source and target periods are assigned identical poverty probabilities, regardless of contextual changes. The same limitations that affect contemporaneous S2S imputation—chiefly, its tendency to replicate the structure of the source data—are likely to be amplified when imputing across time, further constraining the method's ability to

track true changes in poverty and inequality. Given the growing reliance on S2S for poverty monitoring, even subtle distortions can mislead policy. Practitioners should therefore apply these methods with caution, rigorously test validity across time periods, and clearly communicate the method's limitations.

The Variance Decomposition Challenge

The core challenge lies in the decomposition of empirical welfare variance into two components:

- The explicable empirical variance captured by the model: $\text{var}\left[x\hat{\beta}\right] = \frac{1}{n}\sum_i \left(x_i\hat{\beta} - \bar{x}\hat{\beta}\right)^2$, and
- The unexplained random component: (σ_e^2)

In well-specified models, the R^2 typically ranges from 0.40 to 0.60, meaning that a substantial portion of welfare variation remains unexplained. The R^2 statistic represents the proportion of variance explained by the model:

$$R^2 = \frac{\text{var}\left[x\hat{\beta}\right]}{\left(\text{var}\left[x\hat{\beta}\right] + \hat{\sigma}_e^2\right)}$$

When imputing across time, practitioners must rely on an error distribution (σ_e^2) estimated from historical data, same as β. This implicitly assumes that households with the same characteristics face the same probability of being poor across different time periods. However, a critical limitation arises from the assumption of constant parameters over time. In reality, parameters often change, especially in rapidly developing countries or over long time horizons, as highlighted in micro decomposition techniques such as those of Bourguignon et al. (2008). This implies that the estimated contributions of changes in characteristics may not fully reflect actual poverty or inequality levels but rather the role of compositional shifts in shaping welfare evolution. In other words, instead of measuring the true welfare distribution, the method may only capture the effect of compositional shifts – how the population's characteristics are changing over time. The assumption of stable parameters when characteristics change significantly becomes increasingly restrictive, leading to potential bias in predicted welfare levels.[15]

A possible avenue for improvement is to explore methods that allow for parameter variation over time. This could involve explicitly modeling β_t as a function of time or estimating it using repeated cross-sections or panel data, where available. Such an approach could help relax the assumption of constant coefficients and possibly improve the reliability of survey-to-survey imputations in dynamic contexts. However, such an approach is not readily found in the literature.

Population-Level Implications

The predicted poverty rates for the population depend on:

- The distribution of household characteristics in the target period
- The stability of relationships between these characteristics and welfare
- The assumed consistency of unobserved factors affecting welfare

Assuming that welfare is log-normal and independence between the mean $(\bar{X}\beta)$ and the idiosyncratic component, e, then for the population, poverty is given by:

[15]This concern can also be seen directly from the regression equation: $\beta = (X'X)^{-1}X'y$. When the distribution of covariates X changes substantially (whether due to structural transformation, demographic shifts, or policy reforms) both $(X'X)$ and $X'y$ change.

$$FGT_0 = \Phi\left(\frac{\ln z - \bar{X}_{target}\hat{\beta}}{\sqrt{\text{var}\left[X_{target}\hat{\beta}\right] + \hat{\sigma}_e^2}}\right) \tag{5}$$

Mathematically, this translates to predictions, for any given poverty line, depending on both the mean transformed welfare $\left(\bar{X}_{target}\hat{\beta}\right)$ and its variation $\left(\text{var}\left[X_{target}\hat{\beta}\right] + \hat{\sigma}_e^2\right)$.

Sources of Temporal Instability

Several factors can undermine the method's reliability across time:

- Structural changes in welfare determinants (e.g., educational convergence)
- Sampling differences between surveys
- Policy changes (e.g., new welfare programs)
- Economic shocks or systemic changes (e.g., currency reforms)
- The reliability of survey design over time, including consistency in variable definitions, data collection methods (e.g., in-person vs. phone), and other methodological changes that may affect comparability

These implications extend beyond model specification. Many S2S applications employ stepwise regression or regularization techniques (like lasso or ridge regression) to optimize model fit, often measured by R^2. While some argue that endogeneity and omitted variable bias are less concerning in predictive modeling,[16] these issues become particularly problematic when applying models across time periods. A model trained on data from one year may produce biased estimates when used to predict welfare in another year, as both omitted variables and endogeneity can significantly impact the model's temporal stability and will likely cause biased estimates limiting the usefulness of S2S to predict poverty in years where a welfare aggregate is unavailable (see Annex 6.1).

Implications for Inequality Measurement

The challenges noted in previous sections extend to inequality measurement. The reliance on error distributions obtained from a different point in time, particularly affects inequality measurement. Under the common assumption of log-normally distributed welfare, the Gini coefficient depends critically on the welfare distribution's standard deviation. Using an outdated error distribution can thus produce misleading inequality estimates, even when mean predictions appear reasonable.

Assuming that welfare is log-normally distributed then Gini is equal to (Crow and Shimizu 1987):

$$Gini = 2\Phi(\frac{\sigma}{\sqrt{2}}) - 1$$

where σ is the empirical standard deviation of $\ln y$. Consequently, the imputed Gini across time is also dependent on $\hat{\sigma}_e^2$, which is estimated in an older survey and the resulting empirical variance of the linear fit,var $\left[x\hat{\beta}\right]$, in the target survey since the empirical variance of the lognormally distributed y could be decomposed by:

[16] Dang et al. 2025 note that endogeneity is not a concern. An endogenous variable is frequently defined as an explanatory variable that may be correlated with the error term (Wooldridge, 2009 p88). Omitted variables are related as these are correlated to the error term and a covariate.

$$\sigma^2 = \text{var}\left[x_{target}\hat{\beta}\right] + \hat{\sigma}_e^2$$

and consequently, is also subject to changes in the sample's distribution of the observed characteristics used in the model.

The method's tendency to replicate the welfare distribution of the source survey can make it unreliable for measuring changes in inequality and poverty. This suggests that survey-to-survey imputation across time periods should be approached with considerable caution, particularly when economic conditions or social structures have changed significantly between the source and target periods. Given the increasing reliance on these methods to fill data gaps in poverty monitoring, it is crucial to clearly communicate their limitations to avoid misinterpretation of trends. Moreover, imputed poverty estimates should be systematically validated against other proxies for household welfare, such as national accounts, labor market indicators, and administrative data, to ensure consistency and identify potential discrepancies.

4 Model Based Simulations

This section builds on the concepts introduced in Section 3, focusing on the effects of violating the underlying assumptions of survey-to-survey imputation on imputed poverty estimates. Using simulated data, a controlled environment is created to systematically examine these effects. Model-based simulations leverage the assumptions underlying the imputation models to investigate their robustness and identify potential points of failure. Unlike real-world data, model based simulations allow for precise manipulation of individual components, enabling a clearer understanding of how specific changes impact the model's estimates.

4.1 Creating populations

We create 1,000 populations of 20,000 households where the welfare of the population is generated with the following data generating process (DGP):

$$\ln y_i = 3 + 0.1x_{1_i} + 0.5x_{2_i} - 0.25x_{3_i} + 0.2x_{4_i} - 0.15x_{5_i} + e_i \tag{6}$$

where $e_i \sim N(0, 0.5^2)$

1. x_1 is a discrete variable, simulated as the rounded integer value of the maximum between 1 and a random Poisson variable with mean $\lambda = 4$
2. x_2 is a binary variable, taking value 1 when a random uniform number between 0 and 1 is less than 0.2
3. x_3 is a binary variable, taking value 1 when a random uniform number between 0 and 1 is less than 0.5 as long as $x_2 = 1$, otherwise it is equal to 0
4. $x_4 \sim N(2.5, 2^2)$
5. x_5 is a variable drawn from a Student's t distribution with 5 degrees of freedom and scaled by 0.25

The Gini for this distribution is ~0.38, and the covariates explain roughly 47 percent of the variation of $\ln y_i$. For the examples presented in the following sections, source and target sample data will be taken from each population. In this experiment, we take a grid of 20 poverty thresholds, corresponding to every 5th percentile up to the 95th percentile of each population.

In Annex 6.4, we present a different DGP which follows an assumed double nested error model similar to that of small area estimation based on Marhuenda et al. (2017) although covariate generation has been updated to follow the one presented here. This adjusted DGP is used to illustrate the limited impact on bias when the DGP clearly follows a clustered or nested error yet the modeling ignores this.

4.2 How to Impute?

The typical imputation approach undertaken for S2S work done to predict poverty, follows the literature on multiple imputation (MI). A similar method was followed in the original software implementation of small area estimation proposed by Elbers et al. (2003), PovMap (Zhao 2006). Under the MI approach, the parameters estimated on the training data are not applied directly to the target data, instead the parameters to be applied to the target data are drawn from their posterior distributions as illustrated in section 3. For small area estimation, the imputation approach from Elbers et al. (2003) has been updated to follow the approach from Molina and Rao (2010).[17] Under the method of the latter, the parameters estimated using the training data are applied directly to the data and noise is estimated via a parametric bootstrap which is aligned to the model's assumptions (González-Manteiga et al., 2008).

To compare different imputation methods, a simple random sample (SRS) comprising 20 percent of the data generated in Section 4.1 is used as both source and target data. In the results presented in this section, the same dataset is used for both model development and prediction. This approach eliminates potential additional noise introduced by sampling variability, allowing for a more controlled assessment of imputation performance that is entirely driven by the model fit and its assumptions. The process is applied across 1,000 generated populations allowing us to assess the method's empirical bias.

4.2.1 Why simulate the errors?

Survey-to-survey (S2S) imputation of poverty is not simply about predicting the dependent variable. Relying solely on the predicted value of welfare to derive poverty rates often yields poor results. This section uses the synthetic population introduced in Section 4.1 to illustrate this point.

We fit a linear model on a simple random sample (SRS) comprising 20 percent of the population. This model is then used to generate two sets of predicted values for the full data:

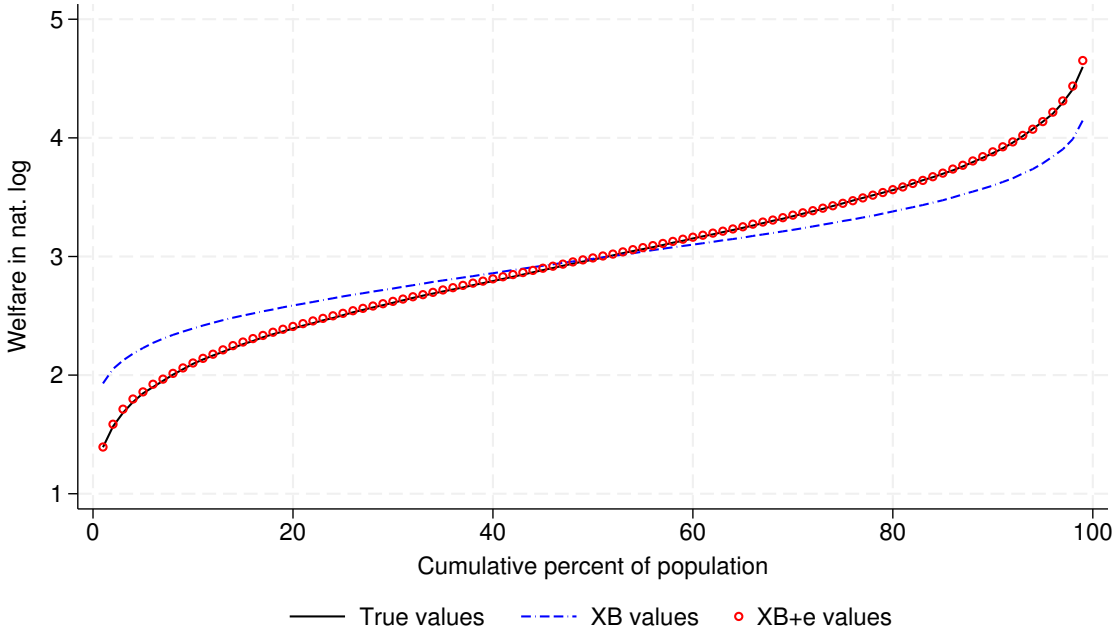
1. The linear prediction, $\hat{y_i} = x_i\hat{\beta}$
2. The linear prediction plus a randomly drawn error: $y_i^* = x_i\hat{\beta} + e_i^*$, where e_i^* is drawn from $N(0, \hat{\sigma}^2)$.

Figure 1 shows that relying solely on the linear prediction, $x_i\hat{\beta}$, fails to replicate the true distribution of welfare. For example, if the true poverty line is 2.39—corresponding to a 20 percent poverty rate in the actual data—then the linear prediction alone would imply a poverty rate of only 10 percent. This gap persists despite the fact that the linear prediction has a lower mean squared error (MSE = 0.25) than the version with stochastic error (MSE = 0.5). This suggests that minimizing MSE is not sufficient for reproducing the distributional properties needed to estimate poverty accurately. Moreover,

[17] See a detailed discussion in Corral et al. (2021) and Corral Rodas et al. (2021)

the Spearman rank correlation is higher for the linear prediction than for the version with added error. In other words, correctly ranking households is also not enough to replicate the full welfare distribution or derive accurate poverty statistics.

Figure 1: Replicating the actual distribution, with and without simulated errors



Note: Data are generated as described in 4.1. True values correspond to those generated following Eq. 6, XB are the predicted linear fit on to the full data, and XB+e is the predicted linear fit plus a random error following the estimated distribution from the model.

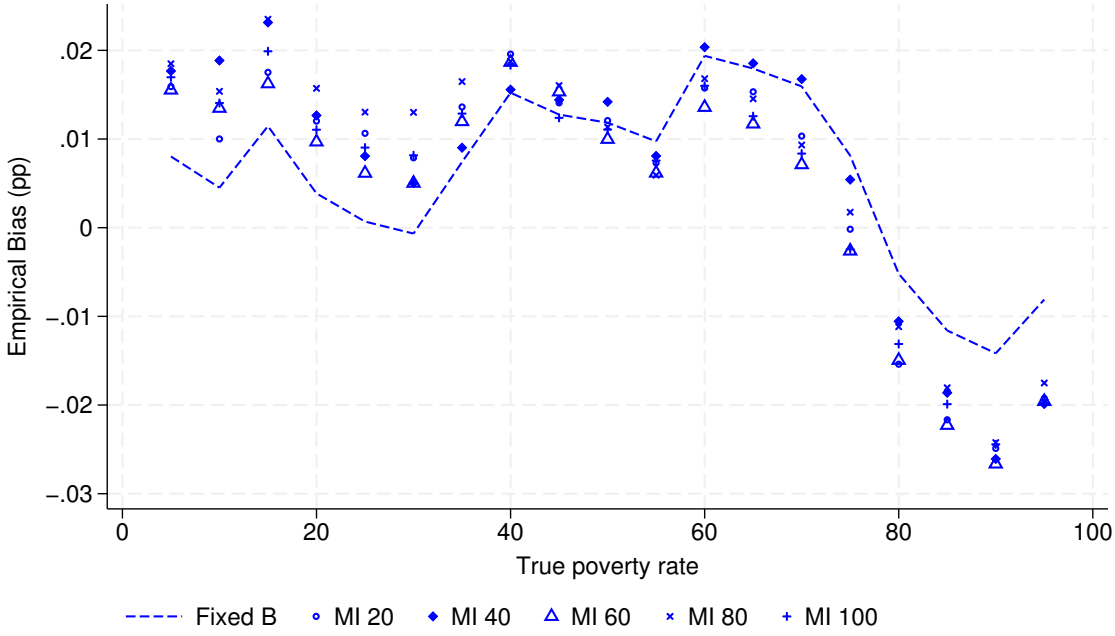
This underscores a central point: accurate estimation of poverty through S2S imputation requires more than good point predictions or rankings. It requires replicating the underlying distribution of welfare.

4.2.2 Under a normal error DGP

The baseline imputation methods assume that welfare is linearly related to a set of characteristics and that errors follow a normal distribution with same variance, σ^2. The data generation process (DGP) used to simulate the dataset adheres to these assumptions.

Under this simulation setup, the prediction bias across different approaches remains minimal across various poverty lines, indicating that all methods effectively replicate the underlying welfare distribution (Figure 2). Although the "Fixed B" method – where parameters are directly applied (Eq. 4) – produces the lowest bias, the differences between methods are negligible. Even in the most biased case observed in this simulation, the discrepancy is less than 0.025 percentage points. Additionally, the number of imputations performed under multiple imputation (MI) appears to have minimal impact on bias (Eq. 3).

Figure 2: Bias in FGT0 under MI and direct application of parameters



Note: Data are generated as described in 4.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5. **Fixed B**: uses the method described in Eq. 4 to generate predictions. For each of 100 Monte Carlo (MC) simulations, poverty is estimated at each threshold and then averaged across simulations to produce the final estimate. **MI X**: implements X imputations following Eq. 3, poverty is estimated for each threshold under each imputed vector, and the results are averaged across simulations to obtain the final estimate.

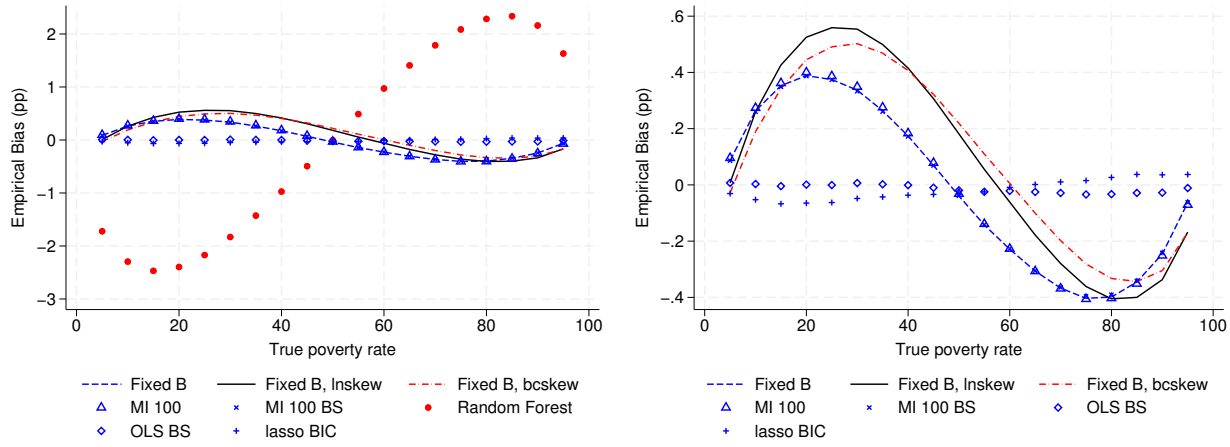
4.2.3 Under non-normally distributed errors

It is essential to ensure the imputation method chosen is the one best aligned to the data at hand. Data transformations, to ensure the model assumptions are met can help (Corral et al. 2021; Rojas-Perilla et al. 2020), but there are instances where transformations are of little use. If the residuals do not follow a normal distribution, an alternative is to draw from the empirical residuals. Stata's `mi regress` includes a bootstrap option that estimates posterior parameters from bootstrap samples, addressing concerns about asymptotic normality when parameter assumptions are questionable (StataCorp, 2023 `mi impute regress` p2.). Similarly, the `hetmireg` command,[18] inspired from PovMap methods (Zhao 2006), generates bootstrap samples with errors drawn from the empirical distribution, though it omits the area random effect required for area-level estimates necessary for small area estimation.

Recent advancements have introduced machine learning (ML) techniques for estimation, but since most ML methods lack explicit assumptions about error term distributions, they cannot be directly applied. Instead, bootstrap-based alternatives, implemented in PovMap (Zhao 2006) and Stata's `sae` command (Nguyen et al., 2018), have proven effective, particularly for regularized regression techniques like lasso regression as shown by Lucchetti et al. (2024). These methods could be extended to other ML approaches, including gradient boosting, random forests, and Bayesian additive regression trees. Verme (2024) offers a useful primer on the performance of random forests and other ML approaches in poverty prediction.

[18] `hetmireg` (Corral - mimeo) command supports heteroskedasticity following the alpha model described in Elbers et al. (2002). Corral (mimeo) - https://github.com/pcorralrodas/hetmireg

Figure 3: Bias in FGT0 of different methods under non-normal errors



Note: Data are generated as described in 4.1, but errors are simulated from a Student's t-distribution with 10 degrees of freedom and scaled for a SD of 0.5. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5. **Fixed B**: uses the method described in Eq. 4 to generate predictions. For each of 100 Monte Carlo (MC) simulations, poverty is estimated at each threshold and then averaged across simulations to produce the final estimate. **Fixed B lnskew**: follows the same approach as "Fixed B" but the dependent variable is transformed using a zero-skewness log transformation. **Fixed B bcskew**: follows the same approach as "Fixed B" but the dependent variable is transformed using a Box-Cox transformation. **MI X**: implements X imputations following Eq. 3, poverty is estimated for each threshold under each imputed vector, and the results are averaged across simulations to obtain the final estimate. **MI X BS**: same as "MI X" but uses Stata's bootstrap option. **OLS BS**: Fits the model on bootstrap samples drawn using simple random sampling (SRS), applies the estimated parameters to the target data, and draws residuals from their empirical distribution. **lasso BIC**: applies parameters from lasso model where λ is selected using the Bayesian information criterion (BIC) function and residuals for each imputation are randomly drawn from the predicted residuals. **Random Forest:** applies parameter estimates from a random forest model to the target data, with residuals for each imputation randomly drawn from the distribution of predicted residuals.

Results from a simulation where the DGP is adjusted and errors are simulated from a Student's t-distribution with 10 degrees of freedom and scaled for a SD of 0.5 are presented in Figure 3. Results from this simulation seem to suggest that random forest prediction coupled with residuals drawn from their empirical distribution yield the most biased estimates, although if the poverty line were at the 50th percentile it would be deemed unbiased if other percentiles are ignored (Fig. 3, left).[19] Under this type of errors, the applied data transformations are of limited use (Fig. 3, right - "Fixed B, lnskew" and "Fixed B, bcskew").[20] Additionally, Stata's `mi regress` with the bootstrap option does not yield considerable improvement under this simulation.[21] Drawing residuals from their empirical distribution seems to yield the best results under this simulation where errors follow a Student's t-distribution with 10 degrees of freedom and scaled for a SD of 0.5. Both the lasso model fitting, paired with residuals drawn from the empirical distribution and an OLS with residuals drawn in the same way yield the least biased estimates.[22]

Motivated by the unexpectedly poor performance of random forests in predicting poverty levels, particularly given the results presented in Verme (2024), we also examined their classification ability in a separate simulation using normally distributed errors.[23] This distinction is relevant for Proxy Means Testing (PMT), where the goal is not to predict continuous consumption or income, but rather to assign

[19] Perhaps under better tuning of the random forest model the bias can be reduced, although for simplicity the basic options of the command are used here. Nevertheless, if classification were the objective, random forest appears to do a much better job than OLS (Tables 2 and 3).

[20] Zero-skewness log (lnskew) or Box-Cox transformation (bcskew). Both can be easily implemented in Stata by using the commands: lnskew0 and bcskew0. If weights are needed, users can use lnskew0w, a modified version of lnskew0 within Stata's SAE package.

[21] Stata's documentation for the option is unexpectedly short and provides few details on the method's implementation.

[22] The OLS implementation relies on the `hetmireg` command by Corral (mimeo) - https://github.com/pcorralrodas/hetmireg.

[23] The data used follows those presented in Annex 6.4

households to eligibility categories. As Grosh and Baker (1995) emphasize, PMT prioritizes classification accuracy over predictive precision.

In this context, when evaluated in-sample, random forests outperform OLS in classification tasks—even though OLS is more closely aligned with the true data-generating process (Tables 2 and 3). When applied to the same data used for modeling, random forest classifies households with notably higher accuracy. However, this advantage disappears when applied to the entire population (i.e., from where the SRS sample was taken from) where the classification performance of random forests closely matches that of OLS (Tables 4 and 5). This suggests that random forests in this example is overfitting in the training sample, underscoring the importance of cross-validation.[24]

The pattern is also evident in the mean squared residuals: in-sample, random forests achieve a low mean squared residual of 0.075, compared to 0.20 for OLS. When looking at out-of-sample results, however, the random forest residuals increase to 0.19, converging toward the OLS out-of-sample value of 0.20.

4.2.4 Under heteroskedastic errors

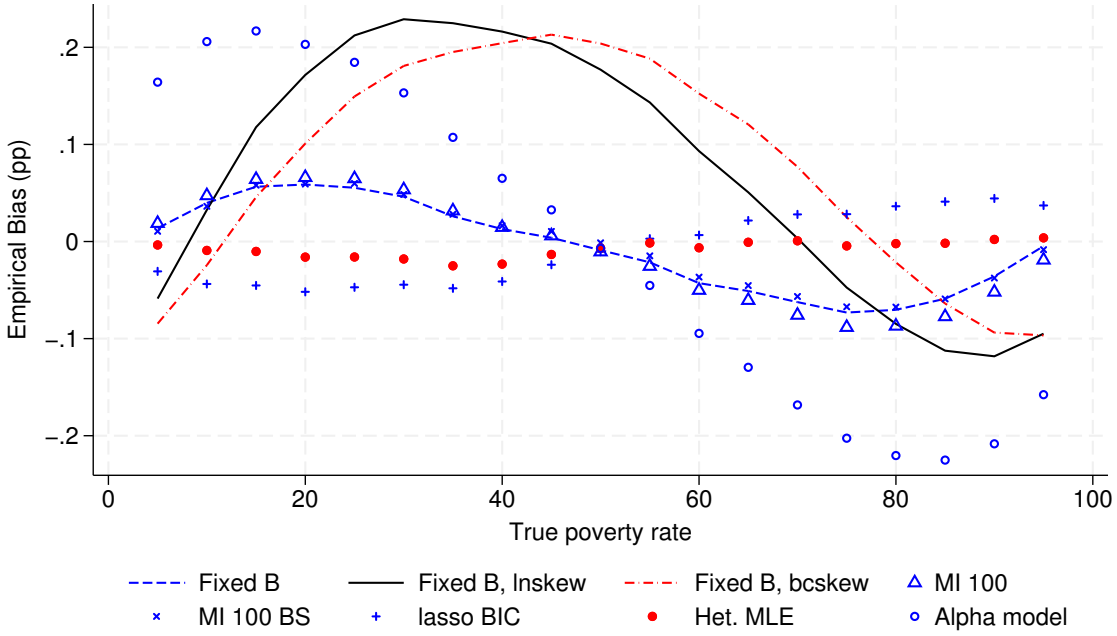
An additional simulation is conducted where the error distribution of the DGP is modified to follow heteroskedastic errors. The errors are designed to mimic a situation where the uncertainty or variability in the dependent variable grows as an independent variable increases.[25] Under this simulation, data transformations are also of little help (Fig.4), and are aligned to the statement made by Betti et al. (2024) who point out that ignoring heteroskedasticity can lead to biased estimates. It seems like the best answer for heteroskedasticity is to address it directly via the model. However, the alpha model from Elbers et al. (2002) does not seem to align as well to the DGP implemented here as the method from Harvey (1976) but fit with MLE ("Alpha model" vs. "Het. MLE").[26] Finally, a lasso model coupled with residuals drawn from their empirical distribution also yields solid results. This last result suggests that drawing from the empirical distribution may be a robust option when one is uncertain about the distribution.

[24] In the simulations conducted here, given that the true DGP is known, model selection and tuning is not undertaken.

[25] The variance of the error term is: $\sigma^2 = \frac{\exp((1/6)x)}{2}$, where $x \sim N(0.5, 0.5)$

[26] The method from Harvey (1976) is applied via Stata's `hetregress` command. While the alpha model from Elbers et al. (2002) is applied using `hetmireg`.

Figure 4: Bias in FGT0 under different methods under heteroskedasticity



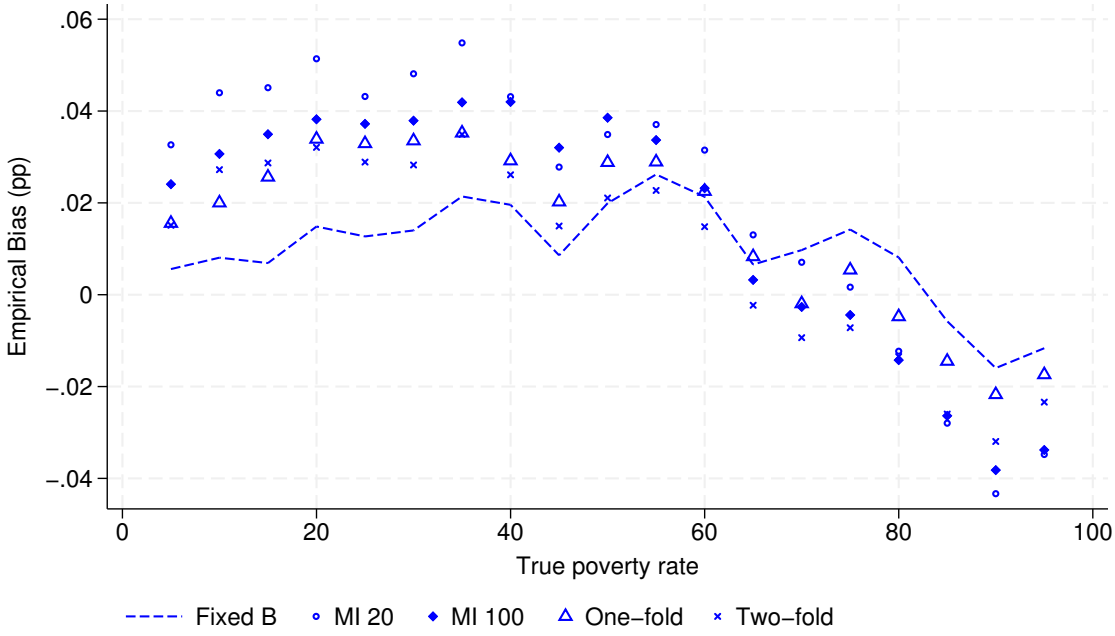
Note: Data are generated as described in 4.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5. **Fixed B**: uses the method described in Eq. 4 to generate predictions. For each of 100 Monte Carlo (MC) simulations, poverty is estimated at each threshold and then averaged across simulations to produce the final estimate. **Fixed B lnskew**: follows the same approach as "Fixed B" but the dependent variable is transformed using a zero-skewness log transformation. **Fixed B bcskew**: follows the same approach as "Fixed B" but the dependent variable is transformed using a Box-Cox transformation. **MI X**: implements X imputations following Eq. 3, poverty is estimated for each threshold under each imputed vector, and the results are averaged across simulations to obtain the final estimate. **MI X BS**: same as "MI X" but uses Stata's bootstrap option. **lasso BIC**: applies parameters from lasso model where λ is selected using the Bayesian information criterion (BIC) function and residuals for each imputation are randomly drawn from the predicted residuals. **Het. MLE**: uses parameters from the Harvey (1976) model for heteroskedasticity and follows Eq. 4 for the predictions where 100 MC simulations are undertaken except that $\sigma^2_{e_i}$ is household specific. **Alpha model**: obtains parameters following Elbers et al. (2002) model for heteroskedasticity and follows Eq. 3, except that $\sigma^2_{e_i}$ is household specific,with 100 imputed vectors.

4.2.5 Under a two-fold double nested normal error DGP

Clustering is unlikely to have an impact on the estimates' bias. In this instance the source and target data differ. This is done to see the impact of applying a model estimated on a given set of clusters and applied to a possibly different set of clusters.

In small area estimation, Marhuenda et al. (2017) suggest that modeling cluster effects at a level misaligned with the level of reporting can bias noise estimates. Corral et al. (2021) reinforce this point using model-based simulations but show that the bias of point estimates remains unaffected. In our case, since we report national-level estimates, ignoring random location effects does not compromise validity. This is evident in the simulation of Figure 5, which uses a two-stage sample drawn from a clustered data-generating process (see Annex 6.4). The figure compares models that account for clustering (one-fold and two-fold) with a model that ignores it ("Fixed-B", "MI 20", and "MI 100") and shows that all three produce similar bias profiles. All result in low bias, at worst it is a little under 0.06 percentage points. The results shown here underscore that accounting for clustering is not critical for point estimates at the national level, though it does matter for the precision of subnational estimates (see Corral et al. (2021) and Marhuenda et al. (2017)).

Figure 5: Bias in FGT0 under a two-fold DGP



Note: Data are generated as described in Annex 6.4. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5. **Fixed B**: uses the method described in Eq. 4 to generate predictions. For each of 100 Monte Carlo (MC) simulations, poverty is estimated at each threshold and then averaged across simulations to produce the final estimate. **MI X**: implements X imputations following Eq. 3, **One-fold** and **Two-fold** follow small area estimation methods accounting for clustering in the data. Poverty is estimated for each threshold under each imputed vector, and the results are averaged across simulations to obtain the final estimate.

4.2.6 PMM - A flexible imputation approach

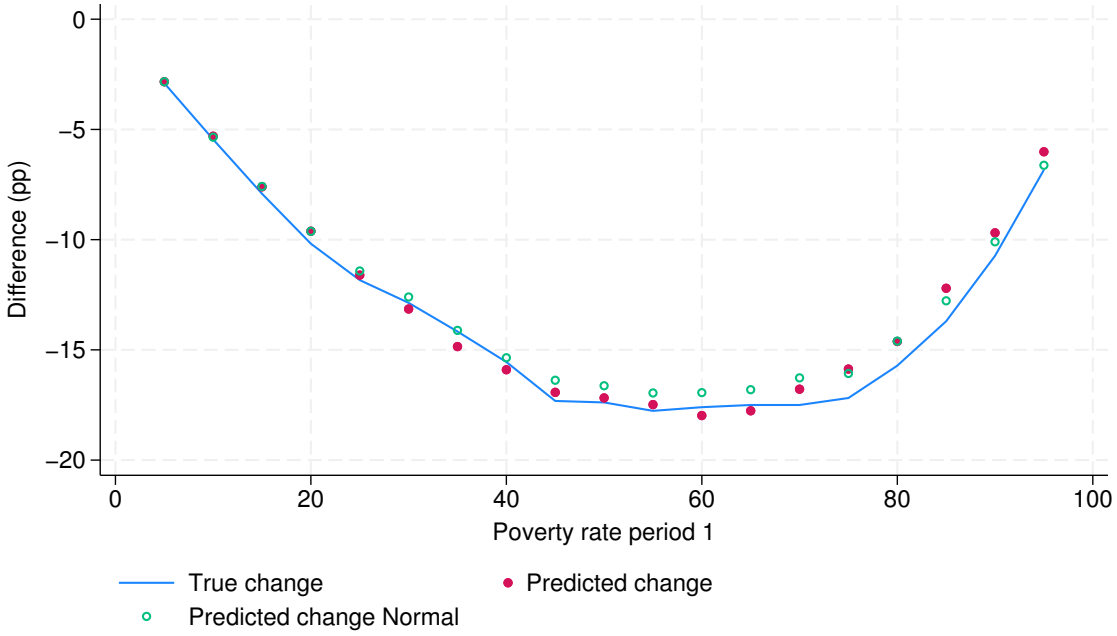
A common alternative imputation approach, predictive mean matching (PMM), differs in its simulation strategy. Rather than drawing residuals or estimating parameters from a posterior distribution, PMM matches observations from the source dataset to those in the target dataset based on predicted values, ensuring that imputed values remain within the observed range of the source data.[27] This method is particularly useful when handling non-normal or discontinuous distributions (Yoshida et al., 2021). However, its reliance on observed values limits its ability to capture shifts in economic conditions between periods, similar to other approaches presented in the section.

To illustrate PMM's limitations, a simulation is conducted where a covariate distribution in the target dataset is altered, creating a mismatch with the source data. Under the simulation, the x_2 variable is adjusted to take a value of 1 when a random uniform number between 0 and 1 is less than 0.8 instead of the original 0.2. This is only done for the target data. The model is fit on the original data, but the imputations are made on the adjusted data. Despite this shift, PMM follows OLS-based predictions closely (Figure 6). However, in a second simulation where economic growth of 20 percent is introduced between the source and target periods,[28] PMM fails to capture the distributional shift (Figure 7). This highlights a critical limitation of PMM as well as other methods illustrated: when substantial economic or structural changes occur between data collection periods, the methods may not be appropriate for imputing poverty estimates.

[27] Predictions typically rely on OLS, see (Lucchetti et al., 2024) for an application to lasso model fitting.

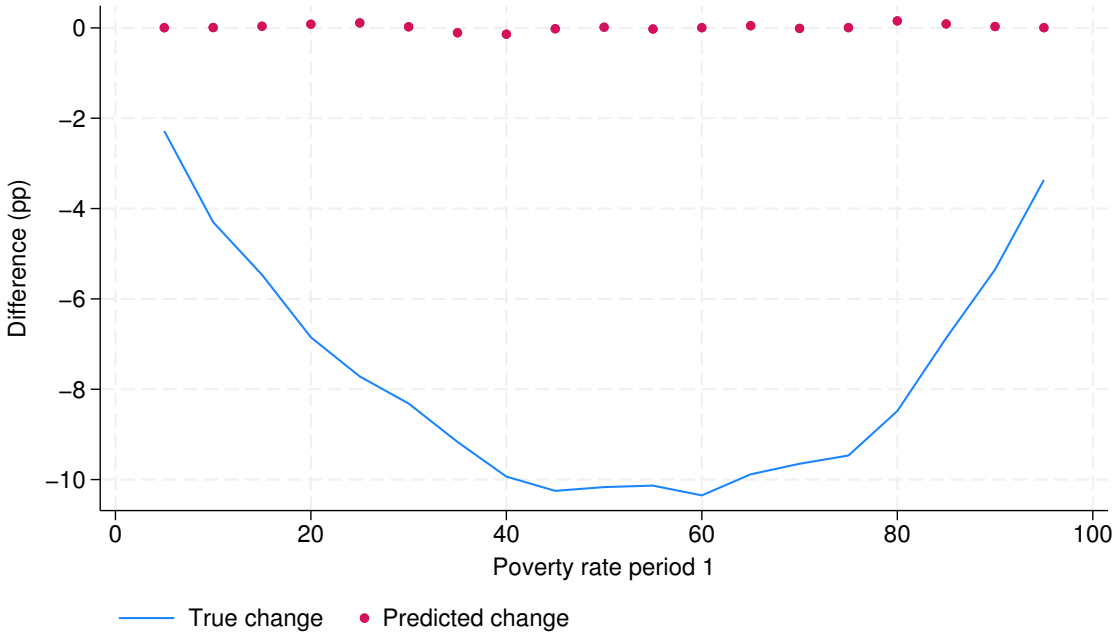
[28] The only change introduced is a change in the constant term, everything else remains equal between source and target data.

Figure 6: Difference in FGT0 under PMM imputations



Note: Data are generated as described in 4.1, with an adjustment for x_2 for the target data. Difference is assessed at various poverty lines across the welfare distribution of the source data. Specifically, these lines correspond to percentiles that are multiples of 5. **Predicted change PMM**: applies PMM imputation using Stata's `impute pmm` command. **Predicted change assuming normal errors:** applies Eq. 4, and follows "Fixed B" from Figure 2.

Figure 7: Difference in FGT0 under PMM imputations



Note: Data are generated as described in 4.1. Difference is assessed at various poverty lines across the welfare distribution of the source data. Specifically, these lines correspond to percentiles that are multiples of 5. **Predicted change:** applies PMM imputation using Stata's `impute pmm` command.

These findings underscore that no single imputation method is universally optimal. The choice of approach should be guided by the data's properties, the error distribution, and the degree of economic

and social change between surveys. Parametric methods can introduce bias when their assumptions are violated, while empirical residual drawing and heteroskedasticity-aware models tend to offer more robust alternatives. Machine learning techniques, particularly when combined with bootstrapping, also show promise for future applications.

The simulation results demonstrate that while many imputation methods perform well under ideal conditions, their effectiveness varies considerably when faced with different error structures and data characteristics. Direct application of parameters ("Fixed B") yields low bias under normal error distributions, but when assumptions are violated—as with Student's t-distributed or heteroskedastic errors—methods using empirical residual drawing perform better. Machine learning approaches, such as lasso regression combined with empirical residuals, show potential but require careful implementation. Conversely, the limitations of Predictive Mean Matching (PMM) become more pronounced when economic conditions change between the training and target periods—though, such instability can affect all models.

Overall, these results highlight the importance of selecting imputation methods that are well-suited to the specific data context, error distribution, and stability of conditions across surveys. Future work could usefully extend bootstrap-based techniques to a wider range of machine learning methods, while ensuring robust error estimation remains central.

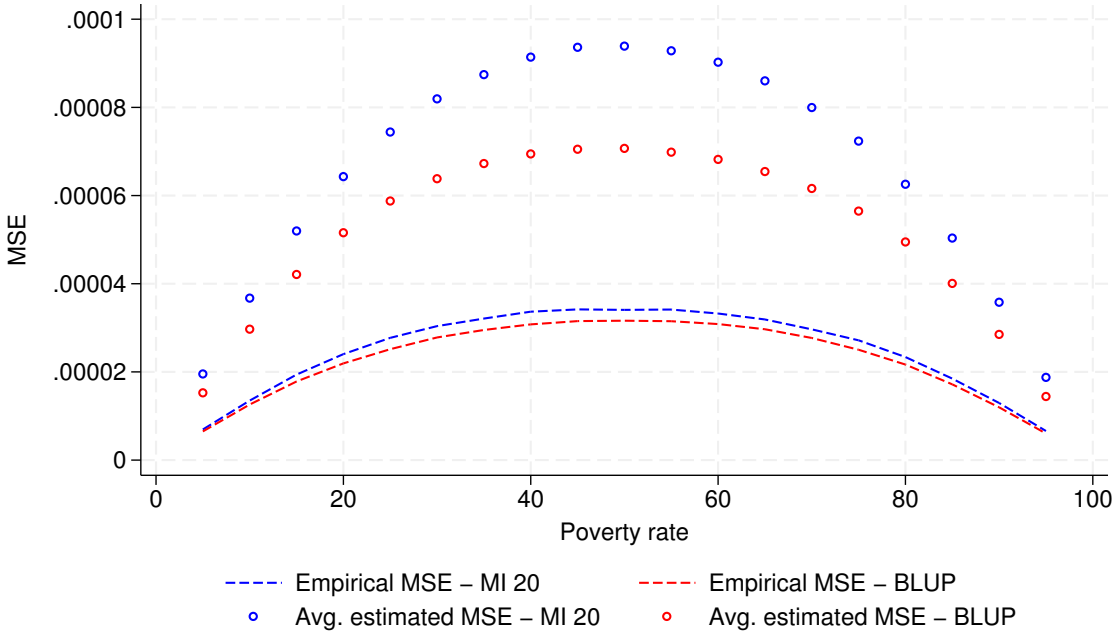
4.3 Noise estimation

A word of caution is warranted here and it is related to how noise is estimated. When applying MI methods the noise is typically estimated using Rubin's rules (Rubin, 1987). Under Rubin's rules the estimated variance is equal to the sum of within-imputation variance, the between imputation variance, and the between imputation variance divided by the number of imputations.[29] It is important to remember that these methods were developed for regression analysis, where a variable with missing observations is imputed and then used in a model. Rubin's rules are applied in those instances to obtain a valid standard error of the regression parameters. Estimating the prediction's noise in a similar manner when parameters are applied directly risks producing improper imputations.[30] When keeping β fixed, as is done by Molina and Rao (2010), MSEs are estimated following a parametric bootstrap presented by González-Manteiga et al. (2008). This bootstrap procedure is aligned to the model's assumptions under small area estimation, but the approach may not necessarily align with multiple imputation.

[29] Under Rubin's rules: $T = W + (1 + \frac{1}{M})B$, where T is the total variance, M is the number of imputations, W is the within survey variance, and B is the between imputation variance. Based on Rubin's rules, the imputed indicator cannot have a variance that is smaller than it would have if it were collected directly in the target survey.

[30] Van Buuren (2018) notes how keeping βs fixed across imputations can lead to improper imputations since the goal of MI is not to minimize the MSE.

Figure 8: Difference in noise estimation under MI and parametric bootstrap



Note: Data are generated as described in 4.1. MSE is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

A simulation is run to test how aligned to the truth are the two methods – Rubin's rules and parametric bootstrap – for noise estimation. Populations of 20,000 observations are generated following the model's assumptions as illustrated in Section 4.1. A random sample of 30 percent is marked as the target survey, and a random sample of 20 percent is marked as the source data for training the model.[31] To compare the noise estimates of each method, 10,000 populations are created. Under each population a new error is drawn from its assumed distribution and predictions are made, as well as noise is estimated – following a parametric bootstrap (González-Manteiga et al., 2008) or following Rubin's rules (Rubin, 1987). Both methods here appear to overestimate the true MSE.[32] The reason for the overestimation is that both methods measure the noise based on the target survey which is smaller than the population. The MSE is equal to the sum of the squared bias and the variance of the indicator of interest. Since both methods are nearly unbiased, it mostly boils down to the variance and given that the target survey is much smaller than the population the variance parameter of the MSE is larger (see Annex 6.2). However, further caution is warranted in instances where the estimates are likely biased since the variance estimated following Rubin's rules will yield a downward biased estimate of the true noise.

4.4 Imputations under biased samples

Survey-to-survey (S2S) imputation methods are often applied to biased samples, such as when specific population segments are systematically underrepresented. To address this issue, practitioners commonly attempt two corrective strategies: re-weighting the sample or standardizing covariates. This section examines these methods and evaluates their effectiveness using simulations.

[31] The samples are selected separately and may have an overlap in observations.

[32] The method's MSE/variance is the average from the 10,000 imputations. The empirical MSE is: $MSE = \sum_{b=1} \frac{(\hat{\tau}_b - \tau_b)^2}{B}$ where τ is the indicator of interest and B is the total number of bootstrap populations.

4.4.1 Extracting samples

Under each of the 1,000 populations created in Section 4.1, the following samples are taken:

1. Random sample: a 20 percent simple random sample (SRS) of the population.
2. Bottom biased sample: here, the poorest quintile is purposely undersampled.
   (a) For the top 80, take a SRS for each centile, c.
   (b) For the bottom 20, take a c percent as a sample. This means that for the 20th centile an SRS sample of 20% is taken, for the 19th centile, 19 percent is sampled, for the 18th centile, 18 percent is sampled, and so forth.
3. Biased sample top and bottom: Create a biased sample, where the poorest quintile and the richest quintile are purposely under sampled.
   (a) For the top 20, sample $100 - c$ percent. This means that for the 80th centile, an SRS sample of 20 percent is taken, for the 81st centile a sample of 19 percent is taken, for the 82nd centile, 18 percent is sampled, and so forth. Note that we do not sample the top 1 percent.
   (b) For the bottom 20, we sample c percent. This means that for the 20th centile an SRS sample of 20 percent is taken, for the 19th centile, 19 percent is sampled, for the 18th centile, 18 percent is sampled, and so forth. Note that the bottom 1 percent is not sampled.
   (c) For all other centiles (21-79), an SRS sample by centile is taken.

4.4.2 Re-weighting the target sample

Re-weighting involves adjusting the sampling weights in the target survey to align its characteristics with those of the overall population. One common approach, minimum cross-entropy,[33] modifies the weights to match constraints like the mean or variance of key covariates. While theoretically sound, re-weighting assumes that adjusted weights can restore the representativeness of a biased sample.

To test the problems with biased samples and re-weighting, two sampling scenarios are used as an illustrative example. The samples are taken from a population created as described in section 4.4.1. For the biased sample exercise, three possible re-weighting scenarios are considered:[34]

1. **Covariate match:** The mean of each covariate from the simple random sample is used as a benchmark to adjust the weights of the biased samples. Re-weighting is performed using minimum cross-entropy, which modifies the prior weights as minimally as possible to ensure that the adjusted sample meets the specified constraints, namely the covariate means.
2. **Linear fit match:** The mean of the model's linear fit from the simple random sample is used to adjust the weights of the biased samples. This calibration is performed using minimum cross-entropy, ensuring that the prior weights are modified as minimally as necessary to satisfy the constraint of matching the linear fit's mean.

[33] When the original sampling weights of the target survey are ignored, it is assumed that every observation had a similar probability of being selected, then the re-weighting method is max-entropy (see Golan, Judge and Miller (1996) for a more thorough exposition of maximum entropy).

[34] Re-weighting is done using Stata's user created command `wentropy` (Corral Rodas & Salcedo DuBois, 2022). The command is used instead of Wittenberg (2010) `maxentropy` command due to `wentropy` adding more flexibility and because it provides solutions to instances where `maxentropy` fails.

3. **Linear fit & variance match:** The linear fit's **mean** and **variance** of the simple random sample are used to calibrate the weights of the biased samples. Weights are calibrated using minimum cross-entropy to adjust prior weights by the smallest amount possible so that the constraints (i.e., the mean and the variance of the linear fit) are satisfied.[35]

4.4.3 Standardizing covariates

Standardizing covariates offers a straightforward approach to addressing disparities between survey samples. Using data from the source survey (or training data), where the model is initially estimated, practitioners can leverage the known means and variances of each covariate to adjust the corresponding variables in the target survey. This alignment ensures that the mean and variance of covariates in the imputed data match those of the source data, enhancing comparability. Dang et al. (2019) advocate this approach as a practical solution for reconciling differences between survey samples. Their application of the method to Jordanian data suggests that it can yield valid results under certain conditions.

The method's efficacy depends on the assumption that covariates follow a normal distribution – an assumption that may not hold universally. To assess the robustness of this approach, the samples described in Section 4.4.1 were used, despite the fact that the covariates in this case deviate from normality. The adjustment process involves scaling and centering the covariates from the target survey so that their means and variances align with those of the source survey, as shown in the following formula:

$$x^{'}_{target} = (x_{target} - \hat{\mu}_{x_{target}}) \frac{\hat{\sigma}_x}{\hat{\sigma}_{x_{target}}} + \hat{\mu}_x$$

Two standardization scenarios are implemented:

1. **Standardize each X**: The **mean** of each covariate from the simple random sample is used as a benchmark to adjust the **mean** and **standard deviation** of the biased samples.

2. **Standardize XB**: The linear fit's ($\hat{y} = x\hat{\beta}$) **mean** and **standard deviation** of the simple random sample are used to standardize the linear fit of the biased target samples.

4.4.4 Results

To isolate the impact of biased sampling in the target data and assess the effectiveness of correction methods, the analysis avoids introducing additional sources of error by using the true values for the coefficients (β) and error distribution (σ_e^2) applied in Section 4.1. Hence, instead of fitting a model, the actual values of the linear fit from Eq. 6 are used. For each sample created in Section 4.4.1, 100 vectors were generated by combining the true β with a randomly drawn error term $e_i \sim N(0, 0.5^2)$ for every observation.[36] Poverty and Gini coefficients were then calculated for each of these 100 vectors using either re-weighted survey weights or standardized covariates. The final estimates were obtained by averaging across the 100 simulations, and bias was determined by comparing these estimates to the true population values across the 1,000 populations, resulting in a robust evaluation of the methods.

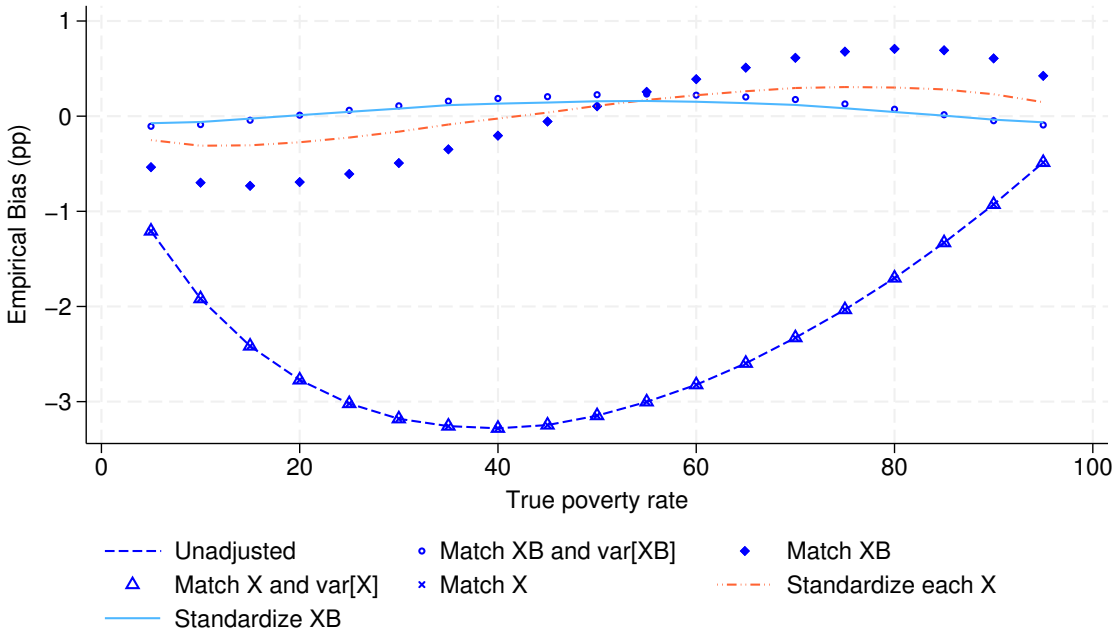
Standard sampling bias-correction techniques, such as re-weighting to match population means, may be insufficient when source and target surveys differ fundamentally—i.e., when they do not appear to

[35] Since the simulated data does not have sampling weights, the priors are equal to 1 and thus minimum cross-entropy is in reality max-entropy in this instance.

[36] Note that the true value is used and not its estimate $\hat{\sigma}_e^2$.

capture the same underlying population, with differing moments of the covariate distributions and altered relationships (covariances) between covariates and outcomes. In such cases, large differences in the covariate distributions suggest that the two surveys do not represent the same underlying population. This limitation becomes evident in bottom-biased samples, where the poorest segments of the population were underrepresented, leading to systematic underestimation of poverty levels across the welfare distribution (Figure 9, "Unadjusted"). Efforts to address this bias by adjusting survey weights to match covariate means had no noticeable effect ("Match X").[37] Incorporating additional constraints to the maximum entropy algorithm, such as aligning both means and variances of covariates, provided marginal improvements but failed to address the core issue of sampling distortion ("Match X and var[X]").

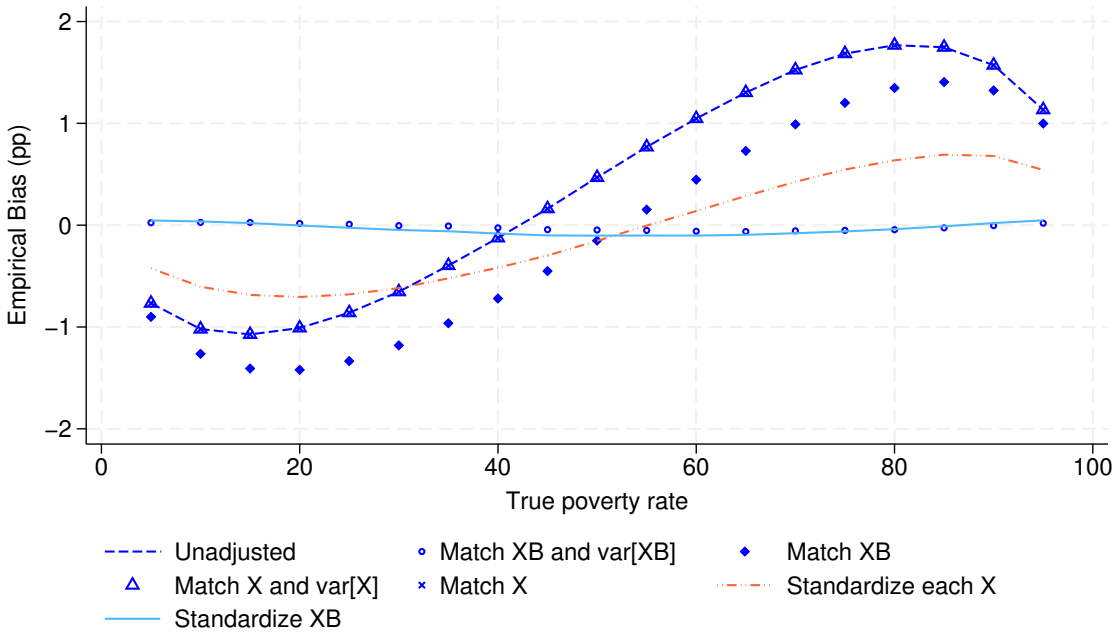
Figure 9: Bias in FGT0 under bottom biased samples (Sec 4.4.1) and different correction measures



Note: Target samples are generated as described in 4.4.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

[37] Since the weights assume no priors, these are in fact calculated using maximum entropy. When prior weights are considered the method adjusts the existing weights by the smallest possible amount to ensure the constraints match, this is known as cross-entropy.

Figure 10: Bias in FGT0 under top and bottom biased samples (Sec 4.4.1) and different correction measures



Note: Target samples are generated as described in 4.4.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

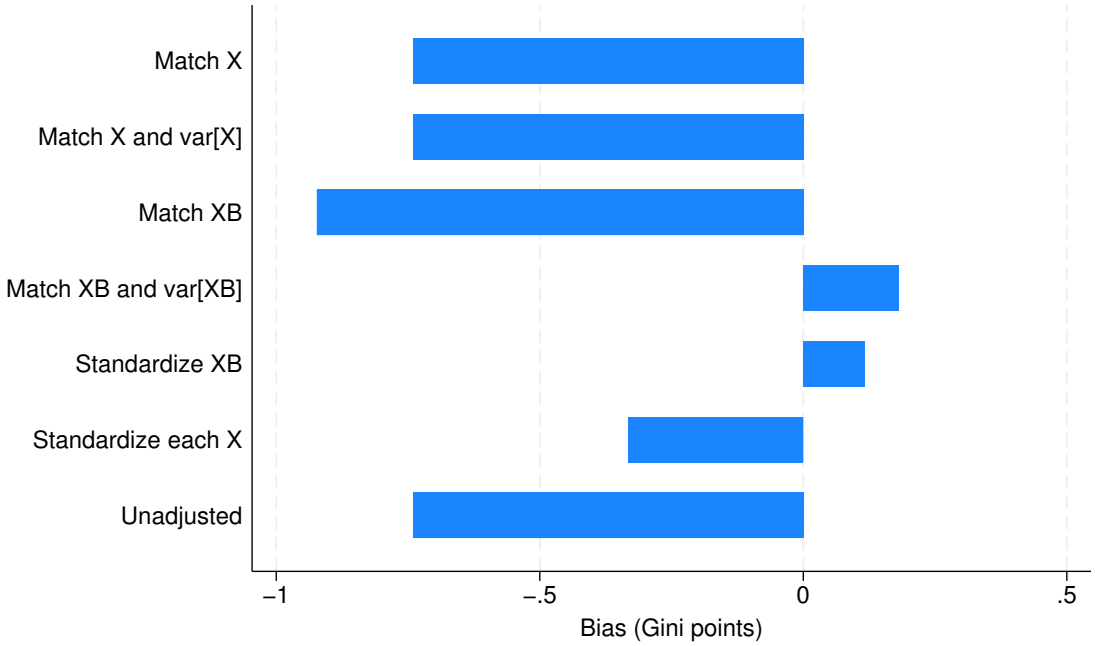
A more effective approach involved aligning the mean of the linear fit of the model ($\hat{y}$) in the target data with the average linear fit of the source data ("Match XB"). This adjustment yielded significantly improved estimates, as did standardizing individual covariates ("Standardize each X"). However, the best performing estimates incorporate the mean of the linear fit and the variance of the linear fit ("Match XB and var[XB]" and "Standardize XB"). These methods effectively address the interaction between mean and variance in determining poverty estimates, as outlined in Equation 5. However, it is important to note that the adjustments made here mimic the moments of the source data, thus when applied to other data would just replicate the source data's welfare distribution. Consequently the approaches would be of use for contemporaneous S2S, but not for S2S over time.

For top-and-bottom-biased samples, where both the poorest and wealthiest households were under-sampled, the performance of the correction methods mirrored the earlier findings (Figure 10). Unadjusted estimates again misrepresented poverty and inequality, while methods that accounted for the mean and variance of the linear fit provided more accurate results. Importantly, results varied significantly across the welfare distribution. Biases that appeared negligible at certain thresholds, such as the 40th percentile, became pronounced elsewhere, highlighting the importance of evaluating imputations across the full distribution.

Implications for Gini Estimation

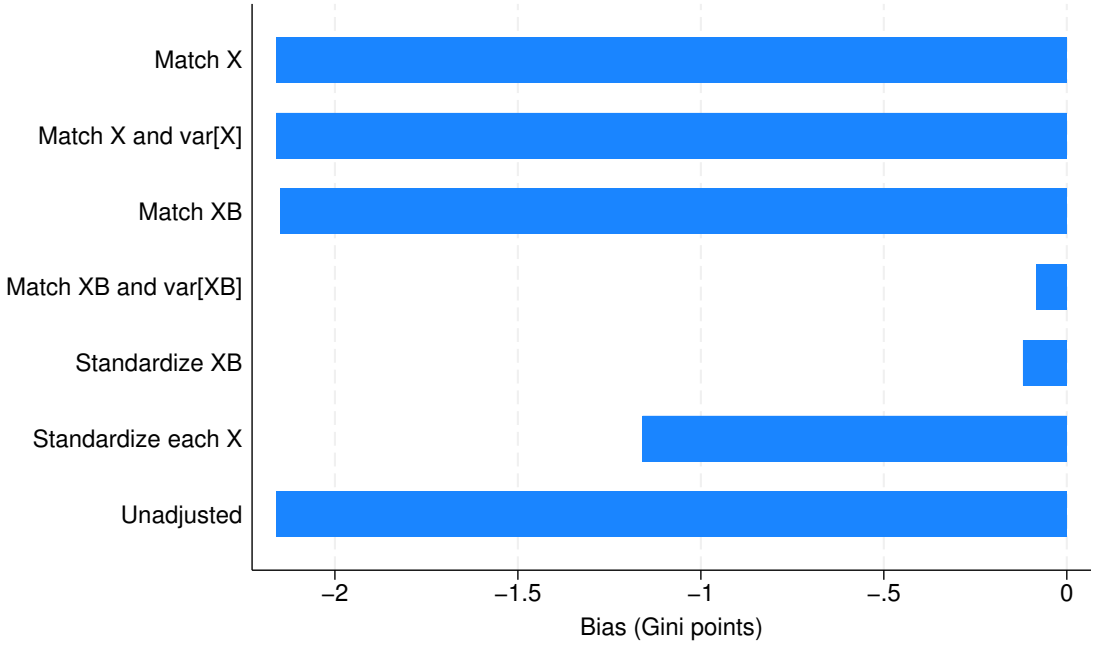
The challenges extend to inequality measurement. Biasing a sample by excluding low-income households reduces var $[x\beta]$, which standard re-weighting methods fail to address (i.e. only addressing differences in the mean). This limitation results in downward-biased Gini estimates (Figure 11). The issue is even more pronounced for top-and-bottom-biased samples, where var $[x\beta]$ is further diminished (Figure 12). Adjusting weights to align both the mean and variance of $x\beta$ with those of the source data significantly improved the accuracy of Gini estimates.

Figure 11: Bias in Gini under bottom biased samples (Sec 4.4.1) and different correction measures



Note: Target samples are generated as described in 4.4.1. Bias is assessed by comparing the mean predicted Gini against the mean Gini of the populations generated (i.e., the truth).

Figure 12: Bias in Gini under top and bottom biased samples (Sec 4.4.1) and different correction measures



Note: Target samples are generated as described in 4.4.1. Bias is assessed by comparing the mean predicted Gini against the mean Gini of the populations generated (i.e., the truth).

4.5 Imputation Over Time

Despite encouraging results for correcting bias in cross-sectional samples, these methods cannot be directly applied to S2S imputation across time. The main limitation is the lack of reliable estimates for the mean and variance of the dependent variable in the target period. Without these benchmarks, standardization or reweighting becomes ineffective, compounding the challenges of imputing poverty estimates in dynamic contexts.

Moreover, when applying S2S to real-world data, the model used is just one of many possible choices. The implicit assumption is that this model represents the true data-generating process (DGP). In practice, there are few clear criteria to select one model over another for poverty imputation. Even a model that fits the source data well offers little guarantee of robust performance when applied across time. Assuming that the estimated model remains valid over time is a particularly strong assumption. As noted by Dang et al. (2025), a high model R^2 does not necessarily translate into accurate poverty predictions. As this section will show, multiple potential sources of bias can arise when applying S2S over time.

Because household living standards surveys are costly and infrequently collected—especially in low-income countries—S2S has become an appealing option in many contexts. Its main applications over time include::

1. Producing comparable poverty estimates: In many countries, surveys may exist, but the welfare aggregate is not comparable across rounds. S2S has been used to bridge this gap by training a model on the original welfare aggregate and applying it to a more recent survey (e.g., Zambia as noted by Yoshida and Aron 2024). This process can also be reversed—if the preferred welfare aggregate is from the newer survey, the model can be trained on that data and applied retrospectively (e.g., Nigeria, Lain et al. 2022).

2. Generating poverty estimates when no living standards data are available: In many cases, a recent survey (e.g., a DHS or labor force survey) does not include an adequate welfare aggregate. Here, a model trained on the last available living standards survey is used to impute poverty in the new survey (For example, see Edochie et al. (2022), Lanjouw, Schirmer, et al. (2024), and Newhouse and Vyas (2019)). The target survey is often assumed to be nationally representative, though this is not always true—sampling bias can distort results (see Roy and Van Der Weide (2022)).

S2S over time introduces additional assumptions beyond those outlined in section 3.1. A key assumption when imputing to a different period than the one used to train the model is that the estimated parameters remain constant over time (Newhouse et al., 2014). This implies that the relationship between covariates and the dependent variable is stable, and the distribution of unobservables does not change. Essentially, this assumes that any shifts in the welfare distribution are entirely driven by changes in the covariates.

Early work on S2S (Stifel & Christiaensen, 2007) often recommended selecting covariates that may vary over time but maintain a stable relationship with welfare – a challenging criterion to meet in practice. Furthermore, restricting the eligible covariates for the model can result in insufficient explanation of the variation in welfare. This, in turn, increases reliance on the distribution of unobservables, which is based on the source data used for the model and may not reflect the target period accurately.

In this section the discussion focuses on how changes in the parameters can affect the poverty predictions. Through simulated data components can be adjusted one at a time to identify how these may impact predictions. In practice, it is possible that any of the parameters used for the imputation (β, σ) has changed over time, however it is impossible to know if changes compound or cancel each other. This

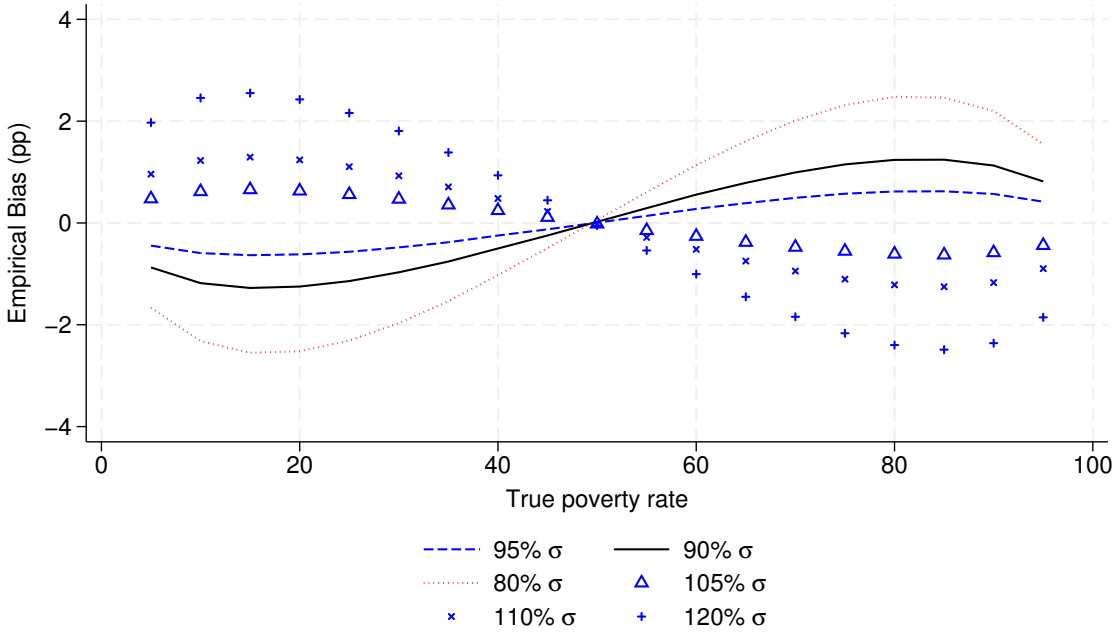
requires careful consideration when applying these methods and transparent communication so potential users, including policy makers, fully comprehend the limitations of the predictions.

4.5.1 Changes in the Error's Distribution

It is possible, though unlikely, for the relationship between covariates and the dependent variable to remain constant over time while poverty levels still change. The change can be driven entirely by changes in the distribution of the unobservables, σ_e^2. Changes in the unobservables also implies a change in inequality. To illustrate this, the same population created in section 4.1 is used. The simulation consists in keeping the linear fit constant and only change the value of σ_e. This is done across the 1,000 populations where predictions are obtained from Eq. 4.

Results for the simulation are presented in Figure 13, and illustrate how poverty rates change under different lines determined at the percentile of the original welfare. Hence, if the original poverty rate was 20 percent, and using that same corresponding poverty line when increasing the value of σ_e by 20 percent, the resulting poverty rate would be biased upwards by over 2 percentage points. At higher thresholds the gap between the original poverty rate and the new rates is less noticeable, although still present. Notably, estimates at the 50th percentile show no bias, reinforcing the importance of evaluating poverty predictions across the full welfare distribution.

Figure 13: Change in poverty prediction if σ changes by x%



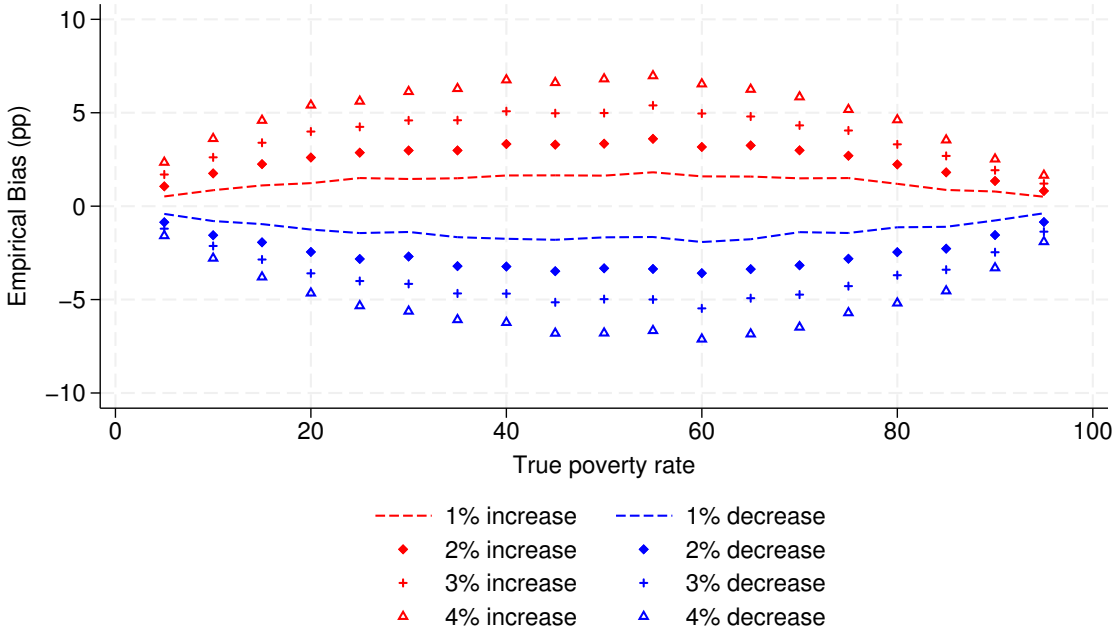
Note: Target samples are generated as described in 4.4.1, only the SRS samples are used for this simulation. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5. Predictions are obtained following Eq. 4, where the value for σ_e^2 is changed.

4.5.2 Changes in the Constant Term

When conducting imputations over time, changes to the constant term are seldom considered. Failing to capture a change in the constant term can lead to relatively stagnant predictions over time. It will also lead to predictions that are considerably off the mark. In Figure 14 the true constant term is adjusted

by $x\%$ in the data generating process for the target data. Even slight changes lead to considerable differences in poverty. Imagine a scenario where transformed welfare for everyone has increased by 1% (Under GDP growth, for example.), yet everything else about the welfare distribution remains the same. This would entail a neutral distribution shift to the right, and thus everyone's transformed welfare is improved by 1%. However, relying on a model fit on data before the increase – even if everything else remains the same – would use a constant term that is lower than what it really is after the increase in welfare. This would lead to imputations that considerably overestimate poverty for the new period.

Figure 14: Resulting poverty prediction if constant term changes by $x\%$



Source: Based on simulated data illustrated in Section 4.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5. Predictions are obtained following Eq. 4.

4.5.3 Omitted Variables

Omitted variables and endogeneity have traditionally not been major concerns in predictions such as S2S or small area estimation. This is because, in the current period, any omitted variables are accounted for by the constant term and coefficients of covariates correlated to the omitted variable, which are adjusted to ensure that OLS predictions of the dependent variable remain unbiased. However, over time, the effects of omitted variables can influence predictions considerably. For a more detailed discussion, see the annex (Annex 6.1).

Consider a plausible application of survey-to-survey (S2S) imputation in a conflict-affected rural setting. Suppose the model predicts a decline in poverty between two survey years. One potential interpretation might be that poverty fell as a result of reduced conflict. However, if the model does not explicitly control for conflict, this interpretation may be misleading. In line with guidance from Yoshida et al. (2022), the model might include fast-changing consumption indicators—such as whether households consumed meat, eggs, or chocolate—as proxies for welfare changes. Yet, if conflict both reduces consumption of these goods and suppresses household expenditure, then omitting conflict from the model could introduce omitted variable bias (OVB). This kind of misspecification can distort poverty estimates, particularly when

key explanatory variables are correlated with both the outcome and unobserved shocks like insecurity (refer to Annex 6.1).

To simulate the potential impact of omitted variable bias on poverty predictions over time the data simulated in section 4.1 is expanded to include 2 new covariates:

1. Conflict. Conflict is assumed to be negatively correlated with welfare with a coefficient equal to -0.3.
   (a) It is simulated as a binary variable taking value 1 when a random uniform number between 0 and 1 is less than $c = 0.4$
2. Eggs purchased. The variable is assumed to be positively correlated to welfare with a coefficient equal to 0.4.
   (a) The variable is simulated as binary, taking the value 1 if a randomly drawn uniform number between 0 and 1 is less than $0.5x - 0.5 \times Conflict$, where $x = 1$ in the baseline. This data generation process (DGP) reflects the negative correlation between conflict and the likelihood of purchasing eggs but also that it is not just a function of conflict but of x as well.

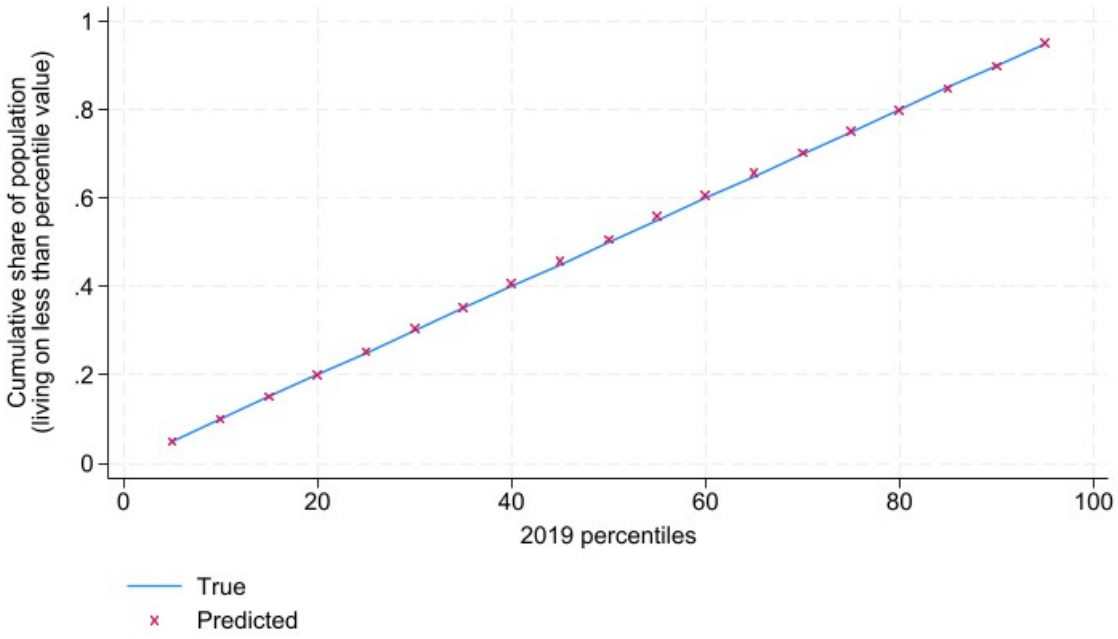
Under the baseline simulation, the value for $x = 1$ and $c = 0.4$. Welfare is simulated assuming the true DGP, and thus includes conflict and eggs with their true coefficients. The prediction model is produced omitting conflict and is executed using the complete data to avoid other potential sources of bias (Table 1). As presented in Annex 6.1, the coefficient on eggs is upward biased. Due to the omission of conflict, the constant term of the regression is downward biased. The adjustment in the constant term ensures that the model's prediction of the dependent variable is unbiased. As can be seen, omitting conflict from the model also leads to a larger RMSE than the truth, which will likely affect poverty predictions over time. Nevertheless, for predictions to the same period OVB has little if any impact (Figure 15).[38]

Table 1: Omitted variable model comparison

	OVB Model	Complete Model
x1	0.0989	0.0989
x2	0.488	0.490
x3	-0.247	-0.244
x4	-0.201	-0.201
x5	-0.160	-0.157
eggs	0.577	0.402
conflict		-0.307
Intercept	2.838	3.012
R2	0.536	0.565
RMSE	0.509	0.492
Observations	20,000	20,000
$\hat{y}$	2.976	2.976

[38]The reason for this is that even in the presence of OVB var $\left[x\hat{\beta}\right] + \hat{\sigma}_e^2$ will be the same to the one where conflict is present and since $\hat{y}$ is also equal poverty predictions in the same period are unaffected.

Figure 15: Resulting poverty prediction under OVB in the same period



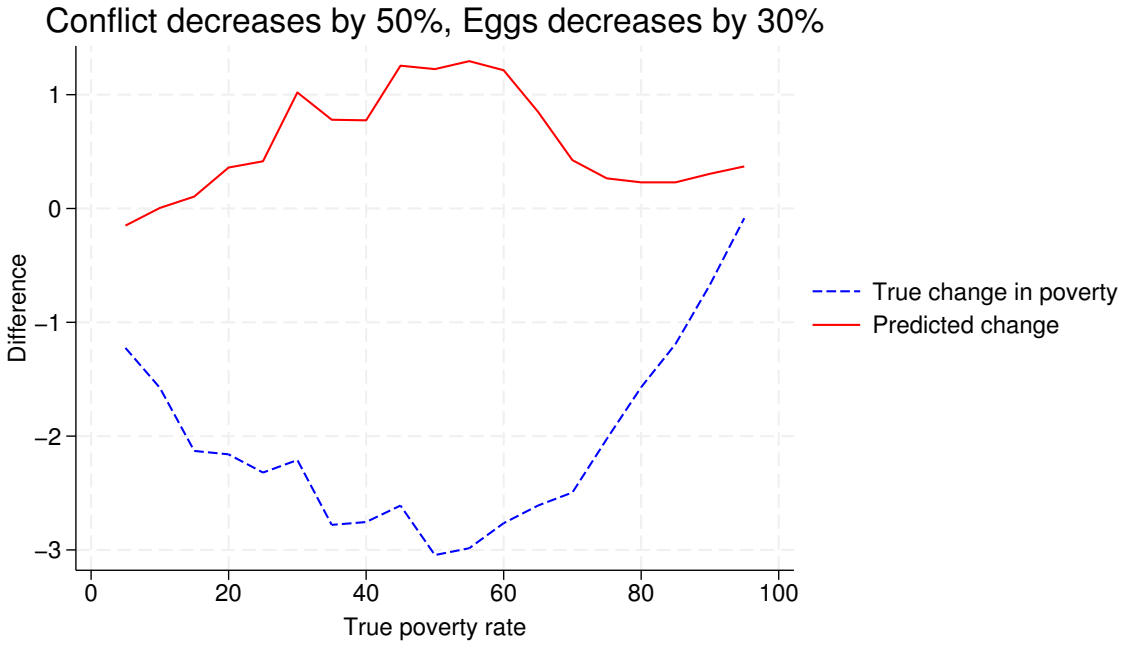
Source: Based on simulated data illustrated in Section 4.1 with added covariates. Predictions are assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

For the simulation, different values of c and x are determined, which yield a change in conflict likelihood and egg purchases. No other covariate is changed. Using the adjusted covariates, the true resulting welfare is calculated – including conflict and the correct coefficients. Then the model parameters obtained from the original data, excluding conflict, is used to obtain predictions of poverty across the welfare distribution. As can be seen in the discussion in the Annex 6.1, the direction of the prediction bias is not immediately clear. In many instances the predictions will suggest an increase in poverty whereas in reality poverty has dropped. In other instances it may underestimate the total change.

Under the assumed data-generating process (DGP), imagine that conflict decreases by 50% ($c = 0.2$), and egg purchases decline by 30% (x =0.7) due to factors unrelated to conflict, for example changes in preferences.[39] In such a case, S2S predictions over time would likely show an increase in poverty based on most thresholds. However, in reality, poverty would have actually decreased (Figure 16). This is just an illustrative example where the purpose is to illustrate the potential problems encountered when imputing over time, particularly in instances that a major shock has occurred in between the year of the actual welfare data and the target data. The inclusion of "fast-changing" consumption variables as covariates is likely to introduce OVB to prediction models for imputations to other periods unless the shock only affects welfare but not individual components of welfare.

[39]Note that because of the relationship to conflict egg purchases do not actually drop by 30% since it is offset by the decrease in conflict.

Figure 16: Resulting poverty prediction under OVB in a different period



Source: Based on simulated data illustrated in Section 4.1 with added covariates. Predictions are assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

The impact of omitted variable bias extends beyond "fast-changing" consumption variables. Consider a model that includes an indicator for subsistence farmers, who are more likely to be poor and thus negatively correlated with welfare. Now imagine a massive drought occurs between the year the model was trained and the target year. This drought could reduce the number of people farming. If those former farmers are unable to meet their needs, the model might underestimate poverty because it omitted the effect of the drought. Another example involves the omission of remittances which are positively related to welfare and likely positively related to "fast-changing" consumption variables, such as eggs or meat. Under this instance the coefficient on eggs would be upward biased, and since the impact of the omitted variable will be absorbed by the intercept it will likely also be upward biased. Thus, changes in remittances that are not captured by the model are likely to lead to biased predictions.

The magnitude and direction of the bias that arises due to OVB is often not clear and could compound other biases just as easily as it could be offset by other biases. Nevertheless, unless actual welfare data are available for the imputed year it is impossible to truly know how the model is performing.

5 Conclusions

This paper has highlighted both the potential and the challenges of survey-to-survey (S2S) imputation for poverty prediction, emphasizing the need for cautious application and robust methodological development. While S2S has several limitations, it may be the best available option in certain scenarios – particularly when real-time poverty estimates are needed, but household surveys are outdated, unavailable, or inconsistent, and when methods relying on real GDP growth are not feasible due to data limitations or reliability concerns. In some cases, S2S may offer more reliable estimates than alternatives such as extrapolations based purely on national accounts, which often diverge from household survey-

based welfare measures. The question is not just whether S2S introduces bias, but how its biases compare to those of alternative methods – or the default of producing no estimate at all.

Based on evidence from model-based simulations that reflect the method's assumed data-generating process, the following recommendations emerge:

1. **Understand and Address Bias Sources:** Practitioners should assess changes in error distributions and parameter instability, as these are key sources of prediction bias. This is especially critical during periods of economic transition. Fast-moving predictors should not be relied upon without thorough validation to mitigate omitted variable bias and other factors affecting the distribution of unobservables.

2. **Enhance Validation Practices:** Validation efforts should go beyond same-period tests and include cross-time validation to assess robustness against economic and structural changes. Practitioners should also evaluate predictions across multiple poverty thresholds to capture distributional nuances. Additionally, external validation – comparing results with broader indicators such as GDP and unemployment, while incorporating country-specific knowledge – enhances reliability. Good examples of these can be seen in Roy and Van Der Weide (2022) as well as Lain et al. (2022) who rely on auxiliary data to lend further credence to their S2S estimates.

3. **Incorporate Structural and Temporal Dynamics:** Methods that allow for time-varying parameters and account for structural changes in predictor-welfare relationships should be considered. For instance, shifting the constant term by observed growth between survey periods can help adjust for evolving dynamics. Combining multiple data sources can further improve the robustness and flexibility of predictions.

4. **Foster Transparency in Methodology and Communication:** Practitioners must pursue clear and transparent reporting of assumptions, validation efforts, and uncertainties in S2S applications. Practitioners should thoroughly document how imputations are implemented—how errors and parameters are applied—and clearly communicate the limitations of the method, especially in contexts with high economic volatility. Uncertainty intervals should go beyond simple noise estimates and incorporate scenario analyses to contextualize findings for stakeholders. This may include sensitivity analyses that adjust key parameters to illustrate how results might shift under different assumptions.

5. **Exercise Caution in Interpretation:** S2S methods should be applied with explicit acknowledgment of their limitations, particularly for tracking poverty over long periods. Transparency is essential – clearly stating that the estimates are predicted rather than derived from an actual household survey ensures proper interpretation and avoids misrepresentation.

6. **Invest in Data and Research:** Greater investment in regular, high-quality data collection is essential for reducing over-reliance on imputation methods and for improving the accuracy of poverty measurement. S2S methods ultimately depend on the quality and representativeness of the data they are built upon—weak or outdated data will naturally compromise the quality of imputed estimates. As the costs of data collection rise, there is a growing temptation to substitute imputation for new data. However, imputation should complement—not replace—regular survey programs. Sustained investment in data is necessary to ensure that models remain valid and are regularly recalibrated to reflect real-world changes. This point is also highlighted in Fujii and van der Weide (2020), who caution against replacing genuine data with modelled estimates in contexts where underlying conditions are evolving.

7. **Support Survey Improvements While Maintaining Comparability**: Statistical Offices should be encouraged to update their surveys in line with best practices – for example, refining the measurement of housing and durable goods in consumption aggregates. However, such improvements can compromise comparability with earlier surveys. To address this, Statistical Offices could introduce "bridging" surveys by fielding the old questionnaire to a sub-sample of the new survey sample. This approach would allow them to assess key S2S assumptions, such as changes in the constant term of prediction models, helping to improve the reliability of imputed estimates.

By recognizing both the challenges and the practical value of S2S, we can better position it as a useful tool for addressing poverty data gaps while maintaining the rigor needed for informed policy decisions. S2S should not be seen as an automatic substitute for household surveys but as a tool that – when carefully applied and validated – can provide meaningful insights when other data sources are unavailable or unreliable. The appropriate use of S2S is not just about improving estimates but also about demonstrating when cruder alternatives may be far more misleading.

References

Allan, F., & Wishart, J. (1930). A method of estimating the yield of a missing plot in field experimental work. *The Journal of Agricultural Science*, *20*(3), 399–406.

Arellano, M., & Meghir, C. (1992). Female labour supply and on-the-job search: An empirical model estimated using complementary data sets. *The Review of Economic Studies*, *59*(3), 537–559.

Beegle, K., De Weerdt, J., Friedman, J., & Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from tanzania. *Journal of development Economics*, *98*(1), 3–18.

Betti, G., Molini, V., & Mori, L. (2024). An attempt to correct the underestimation of inequality measures in cross-survey imputation through generalized additive models for location, scale and shape. *Socio-Economic Planning Sciences*, *91*, 101784.

Bourguignon, F., Ferreira, F. H., & Leite, P. G. (2008). Beyond oaxaca–blinder: Accounting for differences in household income distributions. *The Journal of Economic Inequality*, *6*, 117–148.

Christiaensen, L., Lanjouw, P., Luoto, J., & Stifel, D. (2012). Small area estimation-based prediction methods to track poverty: Validation and applications. *The Journal of Economic Inequality*, *10*(2), 267–297.

Corral, P., Himelein, K., McGee, K., & Molina, I. (2021). A map of the poor or a poor map? *Mathematics*, *9*(21), 2780.

Corral, P., Molina, I., Cojocaru, A., & Segovia, S. (2022). *Guidelines to small area estimation for poverty mapping*. World Bank Washington.

Corral Rodas, P., Molina, I., & Nguyen, M. (2021). Pull your small area estimates up by the bootstraps. *Journal of Statistical Computation and Simulation*, *91*(16), 3304–3357.

Corral Rodas, P., & Salcedo DuBois, R. (2022). *wentropy*.

Crow, E. L., & Shimizu, K. (1987). *Lognormal distributions: Theory and applications*. Marcel Dekker New York.

Dang, H.-A., Jolliffe, D., & Carletto, C. (2017). Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments. https://doi.org/10.1596/1813-9450-8282

Dang, H.-A., Jolliffe, D., & Carletto, C. (2019). *Data gaps, data incomparability, and data imputation: A review of poverty measurement methods for data-scarce environments* (Vol. 33). Wiley Online Library.

Dang, H.-A., Kilic, T., Abanokova, K., & Carletto, C. (2025). Poverty imputation in contexts without consumption data: A revisit with further refinements. *Review of Income and Wealth*, *71*(1), e12714.

Deaton, A., Grosh, M., et al. (1998). Consumption. *Designing Household Survey Questionnaires for Developing Countries: lessons from ten years of LSMS experience. Washington, USA. The World Bank*, 1–78.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Edochie, I. N., Freije-Rodriguez, S., Lakner, C., Moreno Herrera, L., Newhouse, D. L., Sinha Roy, S., & Yonzan, N. (2022). What do we know about poverty in india in 2017/18?

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, *71*(1), 355–364.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2002). Micro-level estimation of welfare. *World Bank Policy Research Working Paper*, (2911).

Fujii, T., & van der Weide, R. (2020). Is predicted data a viable alternative to real data? *The World Bank Economic Review*, *34*(2), 485–508.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, *78*(5), 443–462. https://doi.org/10.1080/10629360600821768

Grosh, M., & Baker, J. L. (1995). Proxy means tests for targeting social programs. *Living standards measurement study working paper*, *118*, 1–49.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica: journal of the Econometric Society*, 461–465.

Hentschel, J., Lanjouw, J. O., Lanjouw, P., & Poggi, J. (1998). Combining census and survey data to study spatial dimensions of poverty. *World Bank Policy Research Working Paper*, (1928).

Lain, J. W., Schoch, M., & Vishwanath, T. (2022). *Estimating a poverty trend for nigeria between 2009 and 2019* (tech. rep.). The World Bank.

Lanjouw, P., Schirmer, P. D., et al. (2024). Imputation-based poverty monitoring in india post-2011.

Lucchetti, L., Corral, P., Ham, A., & Garriga, S. (2024). An application of lasso and multiple imputation techniques to income dynamics with cross-sectional data. *Review of Income and Wealth*.

Madow, W. G., Nisselson, H., Olkin, I., Rubin, D. B., et al. (1983). Incomplete data in sample surveys. *(No Title)*.

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *180*(4), 1111–1136. Retrieved June 6, 2025, from http://www.jstor.org/stable/44682666

Mathiassen, A. (2009). A model based approach for predicting annual poverty rates without expenditure data. *The Journal of Economic Inequality*, *7*, 117–135.

Mathiassen, A. (2013). Testing prediction performance of poverty models: Empirical evidence from uganda. *Review of Income and Wealth*, *59*(1), 91–112. https://doi.org/https://doi.org/10.1111/roiw.12007

Mathiassen, A., & Wold, B. K. G. (2021). Predicting poverty trends by survey-to-survey imputation: The challenge of comparability. *Oxford Economic Papers*, *73*(3), 1153–1174.

Maximum entropy econometrics: Robust estimation with limited data. (1996). Chichester [England] ; New York : Wiley, c1996.

Molina, I., & Rao, J. N. (2010). Small area estimation of poverty indicators. *Canadian Journal of statistics*, *38*(3), 369–385.

Newhouse, D. L., Shivakumaran, S., Takamatsu, S., & Yoshida, N. (2014). How survey-to-survey imputation can fail. *World Bank Policy Research Working Paper*, (6961).

Newhouse, D. L., & Vyas, P. (2019). Estimating poverty in india without expenditure data: A survey-to-survey imputation approach. *World Bank Policy Research Working Paper*, (8878).

Nguyen, M., Corral Rodas, P. A., Azevedo, J. P., & Zhao, Q. (2018). Sae: A stata package for unit level small area estimation. *World Bank Policy Research Working Paper*, (8630).

Rao, J. (2005). *Small area estimation* (Vol. 331). John Wiley & Sons.

Rao, J., & Molina, I. (2015). *Small area estimation* (2nd). John Wiley & Sons.

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(1), 121–148.

Roy, S., & Van Der Weide, R. (2022). Poverty in india has declined over the last decade but not as much as previously thought.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87–94.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rude, B. L., & Robayo, M. (2024). *Statistical matching for combining the european survey on income and living conditions and the household budget surveys: An evaluation of energy expenditures in bulgaria* (tech. rep.). The World Bank.

Simler, K., Harrower, S., & Massingarela, C. (2004). Estimating poverty indices from simple indicator surveys. *conference on âGrowth, poverty reduction and human development in Africa,â Centre for the Study of African Economies, University of Oxford*, 21–21.

StataCorp. (2023). *Stata 18 base reference manual*. College Station, TX, Stata Press.

Stifel, D., & Christiaensen, L. (2007). Tracking poverty over time in the absence of comparable consumption data. *The World Bank Economic Review*, *21*(2), 317–341.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Verme, P. (2024). Predicting poverty. *The World Bank Economic Review*, lhae044.

Wittenberg, M. (2010). An introduction to maximum entropy and minimum cross-entropy estimation using stata. *The Stata Journal*, *10*(3), 315–330.

Wooldridge, J. M. (2009). *Introductory econometrics: A modern approach* (4th). South-Western, Cengage Learning.

World Bank. (2024). *Poverty, prosperity, and planet report 2024*.

Yoshida, N., Chen, X., Takamatsu, S., Yoshimura, K., Malgioglio, S., & Shivakumaran, S. (2021). The concept and empirical evidence of swift methodology. *unpublished manuscript*.

Yoshida, N., & Aron, D. V. (2024). Enabling high-frequency and real-time poverty monitoring in the developing world with swift (survey of wellbeing via instant and frequent tracking).

Yoshida, N., Takamatsu, S., Yoshimura, K., Aron, D. V., Chen, X., Malgioglio, S., Shivakumaran, S., & Zhang, K. (2022). *The concept and empirical evidence of swift methodology*. World Bank.

Zhao, Q. (2006). User manual for povmap. *World Bank. http://siteresources. worldbank. org/INTPGI/Resources/342674-1092157888460/Zhao_ ManualPovMap. pdf*.

6 Annex

6.1 The Case of Omitted Variable Bias

Survey-to-survey imputation across time periods requires careful consideration. For example in a conflict setting, where conflict has significantly subsided between the collection of the training and the target data, it is crucial to control for conflict in the model.[40] In instances where conflict is not included in the model, its effects would only manifest through changes in other variables. This potential omitted variable bias raises concerns about the model's reliability for temporal predictions. When imputing over time, it is possible for multiple issues to arise. Since true expenditure is not available it is impossible to know if the model's biases cancel each other out or compound.

Some may argue that the inclusion of fast changing consumption variables, such as an indicator of whether or not the household consumed eggs over the past week, are sufficient for accurate estimates of poverty . Nevertheless, these variables are likely subject to omitted variable bias which can lead to biased estimates of the coefficients leading to biased poverty measures. In the case of the imputation example given here, the omitted variable is conflict. In others it may be the introduction of a cash transfer or a global pandemic, for example.

Assuming conflict is negatively related to expenditure and negatively related to consumption of certain goods, such as meat and eggs, this would lead to coefficients that are upward biased. Assume the following, simplified, model:

$$Y = \beta_0 + \beta_1 Eggs + \beta_2 Conflict + \varepsilon \tag{7}$$

By omitting conflict the model is now:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 Eggs + u \tag{8}$$

where $u = \beta_2 Conflict + \varepsilon$. In a model where conflict is not included $\tilde{\beta}_1$would be greater than β_1. The OLS estimator for $\tilde{\beta}_1$ is given by:

$$\tilde{\beta}_1 = \frac{\text{Cov}\left(Eggs, Y\right)}{\text{Var}\left[Eggs\right]}$$

$$\tilde{\beta}_1 = \frac{\text{Cov}\left(Eggs, \beta_0 + \beta_1 Eggs + \beta_2 Conflict + \varepsilon\right)}{\text{Var}\left[Eggs\right]}$$

and relying on the linearity of covariance:

$$\tilde{\beta}_1 = \beta_1 + \frac{\beta_2 \text{Cov}\left(Eggs, Conflict\right)}{\text{Var}\left[Eggs\right]}$$

Since we assume that conflict and welfare are negatively correlated, and negatively correlated with the likelihood of buying eggs, then we know that $\beta_2 < 0$ and that $\text{Cov}\left(Eggs, Conflict\right) < 0$. This would lead to $\frac{\beta_2 \text{Cov}(Eggs, Conflict)}{\text{Var}[Eggs]} > 0$ which means that $\tilde{\beta}_1$ is upward biased.

In addition, the intercept will be underestimated. The intercept is equal to:

[40] Note that nothing to control for those shocks was included in the original model.

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \overline{Eggs}$$

$$\tilde{\beta}_0 = \beta_0 + \beta_1 \overline{Eggs} + \beta_2 \overline{Conflict} - \tilde{\beta}_1 \overline{Eggs}$$

$$\tilde{\beta}_0 = \beta_0 + \beta_2 \overline{Conflict} + (\beta_1 - \tilde{\beta}_1) \overline{Eggs} \tag{9}$$

we know that $\beta_2 < 0$, and that $(\beta_1 - \tilde{\beta}_1) < 0$. This suggests that the intercept term is downward biased.

When predicting over time, the direction of the prediction bias is dictated by changes in conflict and the share of households who purchased eggs. Any change in conflict, will lead to a change in egg purchases and expenditure. The direction of the bias in prediction is undetermined:

$$E\left[\bar{\tilde{y}} - \bar{y}\right] = E\left[\tilde{\beta}_0 + \tilde{\beta}_1 \overline{Eggs}' - \beta_0 - \beta_1 \overline{Eggs}' - \beta_2 \overline{Conflict'}\right]$$

replace $\tilde{\beta}_0$ with the value of Eq. 9, and re-arrange:

$$E\left[\bar{\tilde{y}} - \bar{y}\right] = E\left[\beta_2 \left(\overline{Conflict} - \overline{Conflict'}\right) + \left(\beta_1 - \tilde{\beta}_1\right) \left(\overline{Eggs} - \overline{Eggs}'\right)\right]$$

then a prediction will be **upward biased** if:

$$\beta_2 \left(\overline{Conflict} - \overline{Conflict'}\right) > \left(\tilde{\beta}_1 - \beta_1\right) \left(\overline{Eggs} - \overline{Eggs}'\right)$$

and downward biased if:

$$\beta_2 \left(\overline{Conflict} - \overline{Conflict'}\right) < \left(\tilde{\beta}_1 - \beta_1\right) \left(\overline{Eggs} - \overline{Eggs}'\right)$$

and while they could cancel eachother out, it is unlikely.

Note that since poverty also depends on the model's predicted root mean squared error ($RMSE$, $\hat{\sigma}$), which would change over time since it is a function of the omitted variable – conflict – then there are 2 sources of bias. The first is due to the egg coefficient and the intercept, and the second is due to the difference in the distribution of errors which changes due to the omitted variable. A decrease in conflict would potentially lead to a shrinking of the $RMSE$, lowering the poverty likelihood of households but would be ignored in the example provided here.

6.2 Sample Size and MSE

The MSE of an estimator τ is defined as:

$$MSE(\hat{\tau}) = Bias\left(\bar{\tau}\right)^2 + \text{var}\left(\bar{\tau}\right)$$

Under the imputation methods used in the example the bias is assumed to be 0. Thus, the MSE is equal to var $(\bar{\tau})$, which is equal to:

$$\text{var}\left(\bar{\tau}\right) = \text{var}\left(\frac{1}{N}\sum_{i=1}^{N}\tau_i\right)$$

$$\text{var}\left(\bar{\tau}\right) = \frac{1}{N^2}\text{var}\left(\sum_{i=1}^{N}\tau_i\right)$$

Consequently, the variance of an indicator is dependent on the true finite population's size. However, since the methods assess the variance based on the target data (6,000) and not the true population (20,000) the estimated variance will be larger than the true population variance which leads to a larger estimate of the MSE.

6.3 Random forest and OLS classification

Table 2: Random forest confusion matrix - in sample prediction

True\RF Linear fit	Poorest	2	3	4	Richest
Poorest	77.3	20.4	2.2	0.1	0.0
2	20.6	50.3	25.0	4.0	0.1
3	2.2	25.4	45.4	24.9	2.1
4	0.1	3.9	25.3	50.5	20.3
Richest	0.0	0.1	2.1	20.5	77.3

Note: The table reports a transition matrix estimated from simulated data. Each row sums to 100 and indicates the percentage of individuals originally in the given quintile who were classified into each predicted quintile based on the model's linear fit. The values represent averages across 1,000 simulated populations; for each, a sample was drawn, the model was estimated on the sample, and subsequently used to classify households.

Table 3: OLS confusion matrix - in sample prediction

True\OLS Linear fit	Poorest	2	3	4	Richest
Poorest	58.8	25.8	11.2	3.7	0.5
2	25.9	32.1	24.6	13.9	3.6
3	11.3	24.7	28.9	24.5	10.7
4	3.7	13.8	24.6	32.5	25.4
Richest	0.5	3.7	10.8	25.4	59.6

Note: The table reports a transition matrix estimated from simulated data. Each row sums to 100 and indicates the percentage of individuals originally in the given quintile who were classified into each predicted quintile based on the model's linear fit. The values represent averages across 1,000 simulated populations; for each, a sample was drawn, the model was estimated on the sample, and subsequently used to classify households.

Table 4: Random forest confusion matrix - out of sample prediction

True\RF Linear fit	Poorest	2	3	4	Richest
Poorest	57.7	25.6	11.6	4.3	0.8
2	25.7	32.1	24.2	13.8	4.2
3	11.6	24.2	28.9	24.3	11.0
4	4.3	13.9	24.2	32.3	25.2
Richest	0.8	4.2	11.2	25.3	58.6

Note: The table reports a transition matrix estimated from simulated data. Each row sums to 100 and indicates the percentage of individuals originally in the given quintile who were classified into each predicted quintile based on the model's linear fit. The values represent averages across 1,000 simulated populations; for each, a sample was drawn, the model was estimated on the sample, and subsequently used to classify households in a separate sample.

Table 5: OLS confusion matrix - out of sample prediction

True\OLS Linear fit	Poorest	2	3	4	Richest
Poorest	58.7	25.8	11.2	3.8	0.5
2	25.9	32.1	24.6	13.8	3.6
3	11.2	24.6	28.8	24.6	10.7
4	3.8	13.9	24.6	32.4	25.3
Richest	0.6	3.7	10.8	25.4	59.6

Note: The table reports a transition matrix estimated from simulated data. Each row sums to 100 and indicates the percentage of individuals originally in the given quintile who were classified into each predicted quintile based on the model's linear fit. The values represent averages across 1,000 simulated populations; for each, a sample was drawn, the model was estimated on the sample, and subsequently used to classify households in a separate sample.

6.4 Data generation under a two-fold nested error model

Data is generated for the simulations following the assumed double nested error model.[41] The simulated population consists of $N = 20,000$, distributed across $A = 40$ areas ($a = 1, \ldots, A$). Within each area a, observations are uniformly allocated over $c = 10$ clusters ($c_a = 1, \ldots, C_a$). Each cluster c consists of $N_{ac} = 50$ observations and each area consists of $N_a = 500$ observations.

The outcome variable Y is generated is generated from a linear model with additive random effects at the area and PSU level, and household-specific error:

$$\ln y_{aci} = 3 + 0.1x_{1_i} + 0.5x_{2_i} - 0.25x_{3_i} + 0.2x_{4_i} - 0.15x_{5_i} + +\eta_a + \eta_{ac} + \varepsilon_{aci} \tag{10}$$

1. x_1 is a binary variable, taking value 1 when a random uniform number between 0 and 1, at the household-level, is less than or equal to $0.3 + 0.5\frac{a}{40} + 0.2\frac{c}{10}$.
2. x_2 is a binary variable, taking value 1 when a random uniform number between 0 and 1 is less than 0.2.
3. x_3 is a binary variable, taking value 1 when a random uniform number between 0 and 1 is less than 0.5 as long as $x_2 = 1$, otherwise it is equal to 0.
4. x_4 is a binary variable, taking value 1 when a random uniform number between 0 and 1, at the household-level, is less than or equal to $0.5 + 0.3\frac{a}{40} + 0.1\frac{c}{10}$.
5. x_5 is a variable drawn from a Student's t distribution with 5 degrees of freedom and scaled by 0.25.

Random effects are generated as follows:

- Cluster effects are simulated as $\eta_{ac} \overset{iid}{\sim} N(0, 0.1)$
- Area effects as $\eta_a \overset{iid}{\sim} N\left(0, 0.15^2\right)$
- Household specific residuals as $e_{aci} \overset{iid}{\sim} N\left(0, 0.5^2\right)$, where $c = 1, \ldots, C_a$, $a = 1, \ldots, A$.

For each of the 1,000 simulated populations, a two-stage sampling process is used to draw the source sample (used to train the model):

1. In the first stage, 4 clusters are selected at random (without replacement) within each area.

[41] Data is simulated following the same approach as Corral et al. (2022) which adapts the DGP from Marhuenda et al. (2017). The write-up here is also borrowed from Corral et al. (2022).

2. In the second stage, 10 observations are randomly selected from each selected cluster.

This results in a sample of 1,600 observations per population. The target data is drawn using the same sampling procedure, so the target sample may include entirely different clusters and observations than the source sample.