The why, the how, and the when to impute: a practitioners' guide to survey-to-survey imputation of poverty

January 3, 2025

1 Introduction

The measurement and monitoring of poverty are central to assessing global development progress. For the World Bank, whose twin goals include ending extreme poverty, the ability to track poverty reduction is fundamental to 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. This challenge is particularly acute in countries where poverty is concentrated (H.-A. Dang et al., 2017). For instance, India, home to a significant share of the global poor, did not release official consumption survey data between 2011 and 2022. Similarly, Nigeria lacked household survey data for poverty monitoring between 2009 and 2018. These data gaps not only affect our understanding of poverty at the country level but can significantly impact global poverty estimates and our assessment of progress toward poverty reduction targets.

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 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. 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). These methodological challenges add another layer of complexity to the already demanding task of poverty measurement.

The challenges of traditional 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 imputation, often combined with rapid phone surveys or other alternative data collection methods, has emerged as a key tool for

maintaining poverty monitoring during these challenging periods. However, these approaches come with their own limitations and potential biases that must be carefully considered.[1]

Survey-to-survey (S2S) imputation has emerged as a methodological response to these data limitations. The approach builds 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). S2S 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. Small area estimation, as developed by Elbers et al. (2003) and advanced by Molina and Rao (2010), typically applies nationally estimated models to census data to generate precise poverty estimates for small geographic areas. In contrast, S2S focuses on predicting welfare in a separate survey, 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.

This handbook first examines the limitations and potential pitfalls of survey-to-survey imputation through rigorous analysis of its fundamental assumptions. 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 makes it unreliable for measuring changes in inequality. 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, 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 handbook provides practitioners with a framework for understanding when and why S2S methods may produce misleading results.

Because things with real world data tend to work somewhat differently than expected, the handbook then proceeds to validate imputation techniques leveraging the Peruvian data.

[1] For global monitoring see Mahler et al. (2021) Global projections of poverty rely on GDP projections during the pandemic. The GDP projection for each country is used to shift the welfare distribution of a given country by the GDP per capita growth rate observed between year corresponding to the household survey collection and the desired year in the future.

2 Survey-to-Survey Imputation in Action

Survey-to-survey imputation has become a widely used tool for poverty estimation when consumption or income data are unavailable. The World Bank's Survey of Well-being via Instant and Frequent Tracking (SWIFT) program represents a systematic application of this approach, combining S2S methods with rapid surveys to generate poverty estimates.

Evidence from country applications has revealed both the potential and limitations of these methods. In Afghanistan, a 2015 implementation failed to capture significant changes in poverty rates when using a model trained on 2011 data (Yoshida et al. 2021). Similar challenges have emerged in other contexts experiencing rapid economic transitions.

The fundamental challenge lies in the method's core assumptions. S2S imputation 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 2014). This assumption becomes particularly tenuous in dynamic economic contexts. As Christiaensen et al. (2012) note, such assumptions may be especially problematic in rapidly growing economies like India, where structural economic changes can alter the relationship between poverty and its predictors.

Researchers and practitioners have proposed various approaches to address these limitations. Stifel and Christiaensen (2007) advocate for including time-varying variables such as rainfall and prices to capture temporal changes. In response to experiences like Afghanistan, the SWIFT program has adapted its application of S2S by incorporating variables that track economic conditions more directly (Yoshida et al., 2021). 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 variables.

The effectiveness of S2S imputation across time has been subject to empirical validation across various contexts, revealing both its potential and limitations. While some studies report estimates within acceptable margins of error, others highlight significant discrepancies between predicted and observed poverty rates. These mixed results underscore the importance of careful consideration when applying S2S imputation across different temporal and economic contexts.

The bias of estimates obtained with data that correspond to very different time periods or data where the covariates are considerably different has mostly been studied using real world data. For example, Dang, Lanjouw and Serajuddin (2014) conduct experiments using the Household Expenditure and Income Survey and the Unemployment and Employment Survey in Jordan; Stifel and Christiaensen (2007) rely on survey data for Kenya to conduct experiments; Dang et al. (2021) applies the method to several countries and note that estimates are well within margins of error.[2]

Christiaensen et al. (2012) undertakes an empirical validation of survey-to-survey imputation methods over time. The authors perform survey-to-survey imputation over time in scenarios where there is a comparable expenditure data which provides a "true" estimate of poverty. The authors validate their approach using 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

[2]The countries are: Ethiopia, Malawi, Nigeria, Tanzania, and Vietnam

the methods work relatively well. However, depending on the model used differences between predicted and observed rates were considerable.

Applications of survey to survey imputation have also made their way to global poverty monitoring.[3] This is mostly due to India's lack of recent survey data. The following case studies illustrate specific instances where S2S applications have found their way to the World Bank's global poverty numbers.

2.1 S2S in India

India's importance in global poverty measurement cannot be overstated. Its large population implies that even a slight shift in poverty in the country can make or break the World Bank's pledge on ending extreme poverty. Consequently, since 2011, the last time data was collected before 2022, there are at least 3 different studies which attempt to estimate a poverty rate for the country making use of survey-to-survey imputation (see Edochie et al. (2022), Newhouse and Vyas (2019), and Roy and Van Der Weide (2022)). These papers obtain a welfare model constructed on the 2011 data and apply it to several more recent data which lack a welfare measure to obtain poverty predictions.

In 2017, when the next expenditure survey for the country was supposed to be released it was scrapped due to concerns regarding its validity. There were leaks of the report, however, and these suggested that between 2011 and 2017/18 consumption in the country had fallen by 3.7 percent.[4] While rural consumption fell by 8.8 percent, urban areas fared somewhat better and grew by 2 percent over the period. Given the lack of actual data to validate the leaked report and the fact that the survey was never "official" the leaked report never fed into the World Bank's global poverty monitoring efforts.

Various studies utilizing survey-to-survey (S2S) imputation have produced divergent estimates of poverty in India, raising questions about the robustness of such methods. Newhouse and Vyas (2019) estimate a dramatic reduction in poverty, from 22.5 percent in 2011/12 to 12.7 percent in 2014/15, suggesting poverty was nearly halved during this period. In contrast, Roy and Van Der Weide (2022) provide a higher poverty estimate for 2015, ranging between 18.6 percent and 20.6 percent, while an earlier version of the same paper placed the figure at around 15 percent. Similarly, estimates for 2017/18 differ: the World Bank's Poverty and Shared Prosperity Report (PSPR) (2020), following Edochie et al. (2022), provides estimates of poverty at 9.9 percent. However, Roy and Van Der Weide (2022) suggest the poverty rate for 2017 lies between 12.2 and 15.3 percent, with a previous version of their work estimating 13 to 14 percent.

A common feature across these imputation exercises is their reliance on parameters derived from the 2011 National Sample Survey (NSS), despite significant changes in the Indian economy since then. This reliance has produced conflicting trends. For example, the imputation-based models suggest declining poverty and inequality, while data from a leaked, retracted 2017-18 report indicates a fall in rural consumption across all deciles of the welfare distribution, with higher declines among wealthier groups. This discrepancy underscores the limitations of imputation-based approaches when direct data is unavailable, leaving critical questions about the accuracy and reliability of poverty estimates unanswered.

Other authors have attempted to arrive at a poverty rate for India leading to a wide range of predictions, leaving many to wonder what in fact is the true poverty rate for the country.[5] Additionally, a key question in these contexts is what meaning do the estimated standard errors for the imputations carry?

[3] As of this writing, every poverty number for India after 2011 is an imputation. These numbers are reported in the World Bank's Poverty and Inequality Platform with no warning to visitors on the origins of the numbers.

[4] https://www.thehinducentre.com/the-arena/current-issues/article30265409.ece

[5] See for example: Bhalla et al. (2022) and Lanjouw, Schirmer, et al. (2024)

In some instances, through some very elaborate statistical applications poverty numbers are presented with statements of precision that may be misleading.

2.2 S2S in Afghanistan

Afghanistan's 2023 poverty estimates, derived through survey-to-survey (S2S) imputation, illustrate the complexities of applying this methodology during periods of significant economic transition. Barriga-Cabanillas et al. (2023) trained their model using the 2019-2020 Expenditure and Labor Force Survey (IE-LFS), which reported a national poverty rate of 52.3 percent, and applied it to 2023 phone survey data to estimate updated poverty rates.

The estimates indicate a decline in poverty to 48.3 percent in 2023, driven by reductions in rural poverty despite slight increases in urban areas. While these results are noteworthy, several methodological and contextual factors warrant caution. First, the reliance on phone survey data introduces challenges related to representativeness. The 2023 estimates were based on responses from households in the 2019-20 IE-LFS with access to phones and who participated in the follow-up survey. Although weights were adjusted for socioeconomic characteristics such as region, urban/rural status, electricity access, and household assets, such adjustments may not fully address potential biases inherent in phone-based surveys. Experimental evidence (see section 4.3.3) demonstrates the risks of bias when using S2S imputation with re-weighted data.

Second, the model's performance metrics raise questions about its predictive accuracy. Even on the original 2019-20 data used for estimation, model predictions deviated by nearly 1 percentage point, indicating potential violations of key assumptions (Table 4 of Barriga-Cabanillas et al. (2023)). When applied to 2021 data, the model overestimated urban poverty by 1.7 percentage points and rural poverty by 2.1 points, suggesting that the reported 4-percentage-point reduction in 2023 could fall within a margin of error rather than representing a clear trend.

Third, broader economic indicators appear inconsistent with the reported decline in poverty. Between 2019 and 2023, Afghanistan faced significant economic challenges, including severe GDP contraction after administrative changes and U.S. military withdrawal, droughts exacerbating food insecurity, reductions in emergency food assistance, a locust outbreak affecting key agricultural areas,[6] and a doubling of unemployment rates. These factors collectively point to heightened vulnerabilities rather than improvements in welfare.

The authors attribute the poverty reduction primarily to reduced conflict and lower food prices. However, the model does not directly account for conflict, leaving this explanation speculative. Any effects would need to operate indirectly through the model's existing covariates, raising concerns about omitted variable bias that could compromise its predictions over time (see Annex 5.1 for further discussion and section 4.4.3 for simulation results).

Additionally, the model's reliance on consumption indicators, such as meat and egg consumption, introduces further uncertainty. Much of the welfare gains in rural areas stem from variables capturing these specific consumption patterns, which may be closely tied to conflict and other contextual factors. For instance, reported meat consumption increased by 30 percent and egg consumption doubled, with these variables contributing significantly to the estimated rise in real per capita consumption. Yet such trends seem incongruous with broader economic realities, underscoring the challenges in interpreting these results.

[6] World Food Programme (WFP) (2023)

In sum, Afghanistan's 2023 poverty estimates highlight the difficulties of using S2S imputation during periods of economic upheaval. The case underscores the need for cautious interpretation of imputed results, particularly when data collection limitations, methodological assumptions, and contextual factors may influence the findings. Recognizing these complexities is critical to ensuring the reliability of poverty estimates in challenging contexts.

2.3 The Case of Zambia

Zambia's gap in survey data between 2015 and 2022 presented a challenge for monitoring poverty trends. Compounding this challenge, the 2022 consumption data is not directly comparable to 2015 data due to differences in survey design. While the 2015 survey used a fixed recall period for food consumption, the 2022 survey allowed respondents to choose different reference periods for reporting quantities and values of consumption. This variation in reporting likely undermines the comparability of the two datasets, complicating the construction of poverty trends (Beegle et al., 2012).

To address this issue, the Zambia team employed survey-to-survey (S2S) imputation to project the 2015 welfare aggregate onto the 2022 data. Using this approach, they estimate that international poverty ($2.15 2017 USD PPP) increased by nearly 4 percentage points between 2015 and 2022, rising from 60.8 to 64.4 percent. At the same time, inequality, as measured by the Gini index, is predicted to have decreased from 55.9 to 51.5, and average consumption fell by 15 percent, from $2.97 (2017 PPP) to $2.53 (2017 PPP). While these results suggest a worsening of economic conditions, they also raise questions about the broader macroeconomic context.

Although the imputed estimates indicate declining welfare, Zambia's GDP per capita in constant terms barely changed between 2015 and 2022, recovering to its pre-pandemic levels by 2022 and exceeding its 2018 peak by 2023. This divergence between household survey-based consumption estimates and national accounts data highlights a growing gap to national accounts that is not easily explained. Literature suggests such gaps may stem from underreporting of incomes in surveys (Ravallion, 2003), although they tend to narrow as countries become wealthier (Prydz et al., 2022).

The comparable components of expenditure correspond to 33.7 percent of the 2015 survey expenditure, but does not include food, and frequent non-food components. The comparable component consists mainly of health, a sub-set of education, clothing, financial services, durables, and housing. The last item, housing, corresponds to imputed rent. In urban areas, the comparable component suggests consumption has decreased in real terms. In rural areas there is no discernible change. The authors validate their results applying a method from Deaton (2003) that relies on a comparable subset of the welfare aggregate and is aligned to that based on the imputation model. The authors also validate their model by imputing on the same data as the one used for the model. Their validations already point toward a slight upward bias in their model for urban areas, same for their Gini predictions, both likely driven by the residuals not meeting the model's assumptions (Table 6).

Additional concerns about the imputation models remain. First, the models fit have a surprisingly high R^2 value – 0.8 for rural, and 0.91 for urban areas. Such a high R^2 value may be suggestive of overfitting. The rural model includes as a covariate the natural logarithm of comparable meat consumption per adult equivalent, and its square – the coefficients for both are positive.[7] The rural model also includes number of items purchased in logarithms, where presumably $\ln(0)$ is treated as 0,[8] which can lead to

[7] Includes health, a subset of education, clothing, financial services, durables, and housing. It has a correlation of 0.987 with total consumption and corresponds to 33.7 percent of consumption as noted by the authors.

[8] Not explicitly stated in the report.

biased coefficients particularly when the proportion of 0 is large (Battese, 1997).[9] Finally, the model likely includes many covariates that are potentially highly correlated. For example, number of tubers consumed from own consumption and purchased.

The model for urban areas has an R^2 value of 0.91, one of the higher values observed in such an exercise. The model includes the number of inactive household members, which a priori one would expect to be negatively related to consumption, but is positive in this case. Moreover, the model suggests that between 2015 and 2022 the number of inactive members increased by nearly 1 person, from 1.6 to 2.4. Beyond the issue of the inactive members, the urban and rural models share many of the same limitations. Mainly, the high $R^2$is suggestive of overfitting that would limit its predictive out-of-sample capacity, particularly when applying the model to data that is 7 years ahead.

In summary, Zambia's 2023 poverty estimates underscore the challenges of applying S2S imputation in the context of survey design differences and extended time gaps. While the results provide useful insights, issues related to model assumptions, parameter stability, and data representativeness warrant careful consideration. This case emphasizes the importance of robust validation and sensitivity analyses to ensure reliable estimates when addressing data gaps.

To properly understand the limitations of S2S the following section goes in-depth into the basics of S2S.

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,[10] and the first widespread use of the term "imputation" comes from Madow et al. (1983).[11] 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 the multiple imputation approach is not to create a single imputation, but multiple imputations to reflect the uncertainty around the actual value. Multiple imputation's goal is to create imputations for observations with missing data to obtain a valid estimand with adequate 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.

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 the World Bank, 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. Nevertheless, earlier examples exist where data from different sources are combined to predict a variable of interest at the household level (for example, Arellano and Meghir 1992). Nevertheless, what is innovative of the work from Hentschel et al. (1998) is that the authors apply models fit on survey data to the census with the goal of replicating

[9] The authors include the nat. log. of hoes owned as well a fishing and hunting gear which likely have multiple 0.
[10] Allan and Wishart (1930)
[11] As noted by Van Buuren (2018).

the entire consumption distribution and from that distribution obtain estimates of poverty and other welfare indicators as if one had welfare in the census. 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 small area estimation 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 welfare distribution can be captured by a linear model. Hence, the assumed data generating process (DGP) for transformed welfare $\ln y_i$ is:[12]

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

where x_i is a vector of independent variables common to the source survey and the target survey,[13] 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. These estimated parameters are then applied to the target data to predict poverty or other welfare related indicators.

Because normally distributed errors are assumed, for any given household, i, the probability of being poor is entirely dependent on its expected welfare, $x_i\hat{\beta}$, and its error, e_i, which is assumed to follow $e_i \sim N(0, \sigma_e^2)$.

$$\hat{FGT}_{0i} = \Phi\left(\frac{\ln z - x_i\hat{\beta}}{\hat{\sigma}_e}\right) \tag{2}$$

where $\ln z$ is the natural log of the poverty line,[14] 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. Additionally, 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.[15]

The implementation of welfare predictions in the target data can follow different approaches. The traditional method, grounded in the multiple imputation literature, generates several welfare vectors to reflect prediction uncertainty. This approach follows 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 three-step process:

[12] 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.

[13] the source survey is the one used to fit equation 1, and the target survey is where the parameters estimated in the source survey are applied to impute $\ln y_i$.

[14] 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.

[15] Under a PMT the threshold is chosen at a given percentile of $x\beta$ so that by construction it yields a desidred proportion of people eligible.

$$\sigma_e^{2*} \sim \hat{\sigma}_{e0}^2 \frac{(n-K)}{\chi_{n-K}^{2*}},$$

$$\beta^* \sim MVN\left(\hat{\beta}_0, \widehat{\text{vcov}}(\hat{\beta}_0)\right),$$

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

Hence, for each imputed vector, a new σ_e^{2*} is drawn and that is 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 then used for regression analysis.

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 method, 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).

Recent methodological advances have expanded estimation 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. 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,[16] and could be extended to other ML methods such as gradient boosting, random forests, or Bayesian additive regression trees.

3.1 Imputing to a contemporaneous survey

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.

Covariate comparability

The variables used to predict welfare must be present and measured consistently in both source and target surveys. This requirement extends beyond simple presence 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 present:

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

1. Similar distributions of predictor variables
2. Comparable moments (means, variances) across key characteristics
3. Consistent patterns in the relationships between variables

For example, a commonly used predictor like 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, the resulting distributions may differ significantly, potentially leading to biased poverty estimates.

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 discussed in the previous section
2. Homoscedasticity of error terms, although this can be relaxed by modeling for heteroskedasticity following the methods from Elbers et al. (2002) or Harvey (1976)
3. Linear relationships between predictors and welfare

When these 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.

2.2 Imputing across time

While survey-to-survey (S2S) imputation offers a potential solution for missing welfare data, its application across different time periods requires particular scrutiny. The method's fundamental assumption—that changes in welfare are driven solely by changes in observable characteristics while their relationship with welfare remains constant—becomes increasingly tenuous as the temporal gap widens.

The Variance Decomposition Challenge

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

- The explicable variance captured by the model: $\text{var}\left[x\beta\right]$, 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, implicitly assuming that households with similar predicted welfare face the same probability of being poor across different time periods.

Population-Level Implications

For the overall population, poverty estimates 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 log-normality, for the population, poverty is given by:

$$FGT_0 = \Phi\left(\frac{\ln z - \bar{X}_{new}\hat{\beta}}{\sqrt{\text{var}\left[X_{new}\hat{\beta}\right] + \hat{\sigma}_e^2}}\right) \tag{3}$$

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

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)

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,[17] 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 5.1).

Implications for Inequality Measurement

The challenges 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.

[17] H.-A. H. Dang et al. 2021 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.

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

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

where σ is the 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 since:

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

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

These limitations suggest 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.

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):

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

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

1. x_i 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 $\ln y_i$.

4.1.1 Extracting samples

Under each of the 1,000 populations, take 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 an 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% 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% 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.

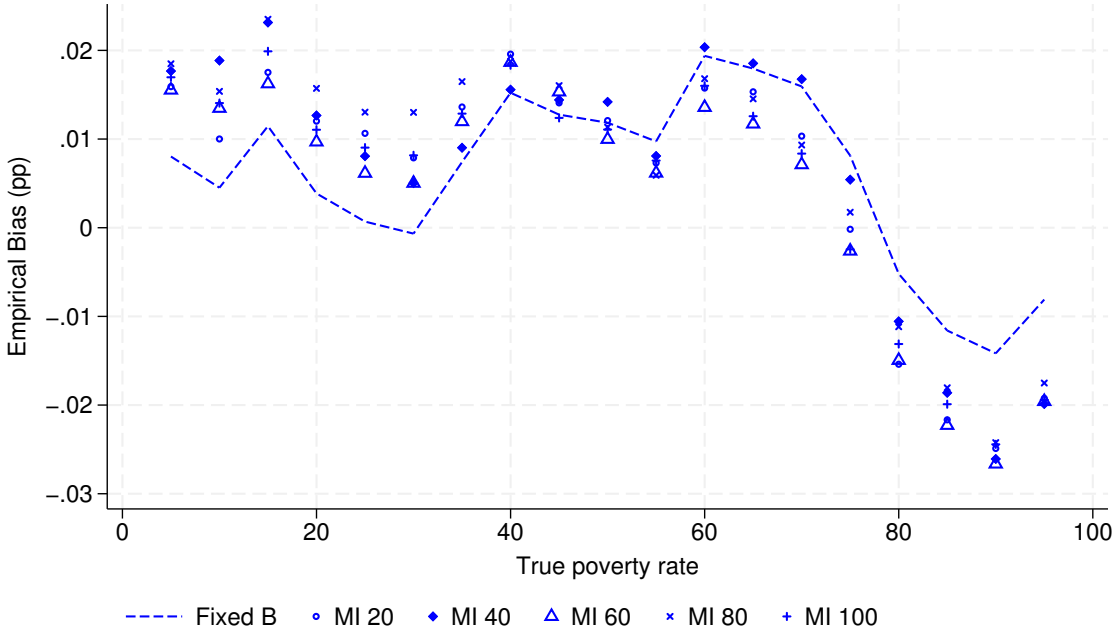
4.2 How to Impute?

The typical imputation approach undertaken, at least under the S2S work done to predict poverty, follows the literature on multiple imputation. A similar approach was followed in the original software implementation of small area estimation method 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 approach from Elbers et al. (2003) has been updated to follow the approach from Molina and Rao (2010). 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)).

Using a SRS of the data created in Section 4.1 as training and target data allows for comparison of different imputation methods. The underlying assumption of the baseline imputation methods is that welfare is linearly related to a set of characterisitics and that the errors are normally distributed. The DGP of the simulated data follows this. Under this simulation, the prediction bias between approaches is minimal across different poverty lines, suggesting they replicate the welfare distribution well (Figure 1). Although applying parameters directly ("Fixed B") results in the lowest bias, the differences between

all methods are minimal. In the most biased case from the simulation presented here, the difference is less than 0.025 percentage points. Additionally, the number of imputations performed under MI appears to have little, if any, impact on bias.

Figure 1: Bias in FGT0 under MI and direct application of parameters

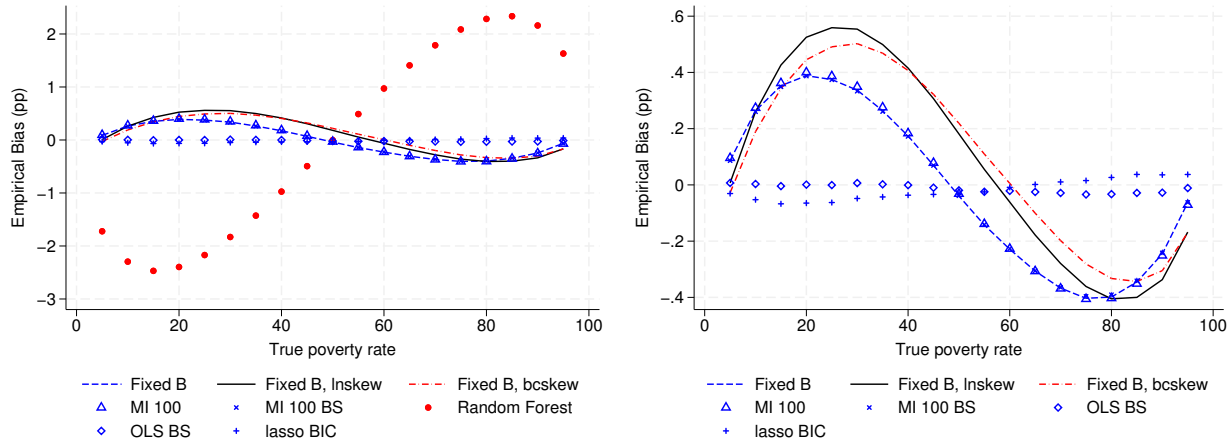


Note: Data are generated as described in 4.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

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. 2022), but there are instances where transformations may not help. If the residuals do not follow a normal distribution, an alternative is to bootstrap 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, inspired from PovMap methods (Zhao 2006), supports heteroskedasticity following the alpha model described in Elbers et al. (2002) and generates bootstrap samples with errors drawn from the empirical distribution, though it omits the area random effect required for area-level estimates.

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.

Figure 2: Bias in FGT0 of different methods under non-normal errors



Note: Data are generated as described in 4.1 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.

Results from a simulation where 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 2. 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. 2, left).[18] Under this type of residuals, the applied data transformations are of limited use (Fig. 2, right - "Fixed B, lnskew" and "Fixed B, bcskew").[19] Additionally, Stata's `mi regress` with the bootstrap option does not yield considerable improvement under this simulation.[20] Drawing residuals from their empirical distribution seems to yield the best results under this simulation. 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.[21]

An additional simulation is conducted where the error term is heteroskedastic. The errors are designed to mimic a situation where the uncertainty or variability in the dependent variable grows as an independent variable increases.[22] Under this simulation, data transformations are also of little help (Fig.3). It seems like the best alternative 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"). Finally, a lasso model coupled with residuals drawn from their empirical distribution also yields solid results.

[18] 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.

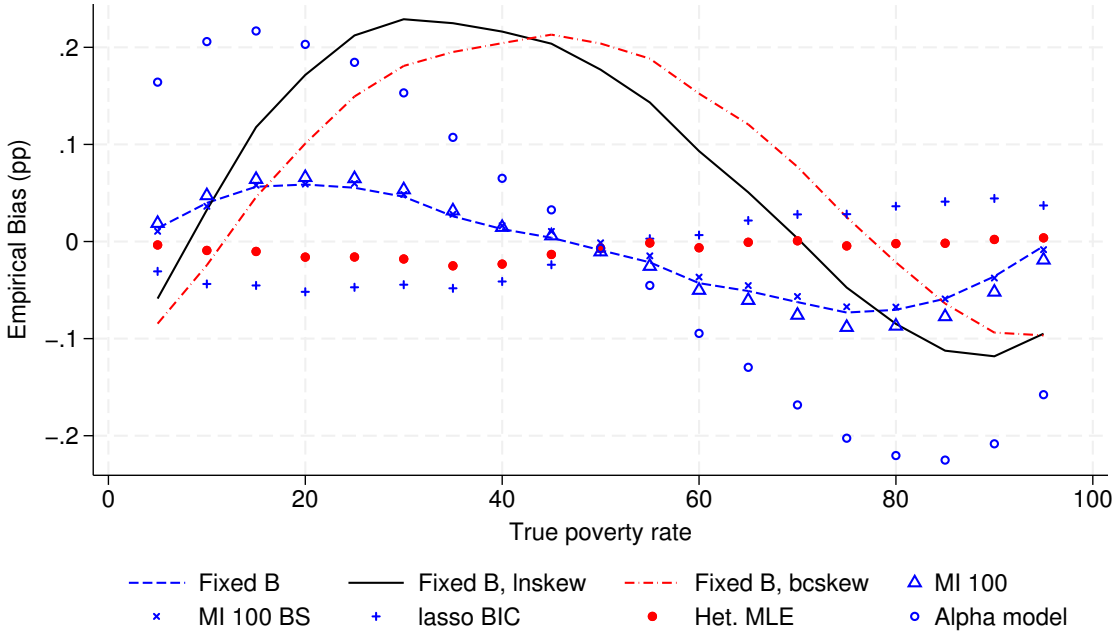
[19] Zero-skewness log (lnskew) or Box-Cox transformation (bcskew). Both can be easily implemented in Stata by using the commands: lnskew0 and bcskew0. If weights are needed, users can use lnskew0w, a modified version of lnskew0 within Stata's SAE package.

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

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

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

Figure 3: Bias in FGT0 under different methods under heteroskedasticity



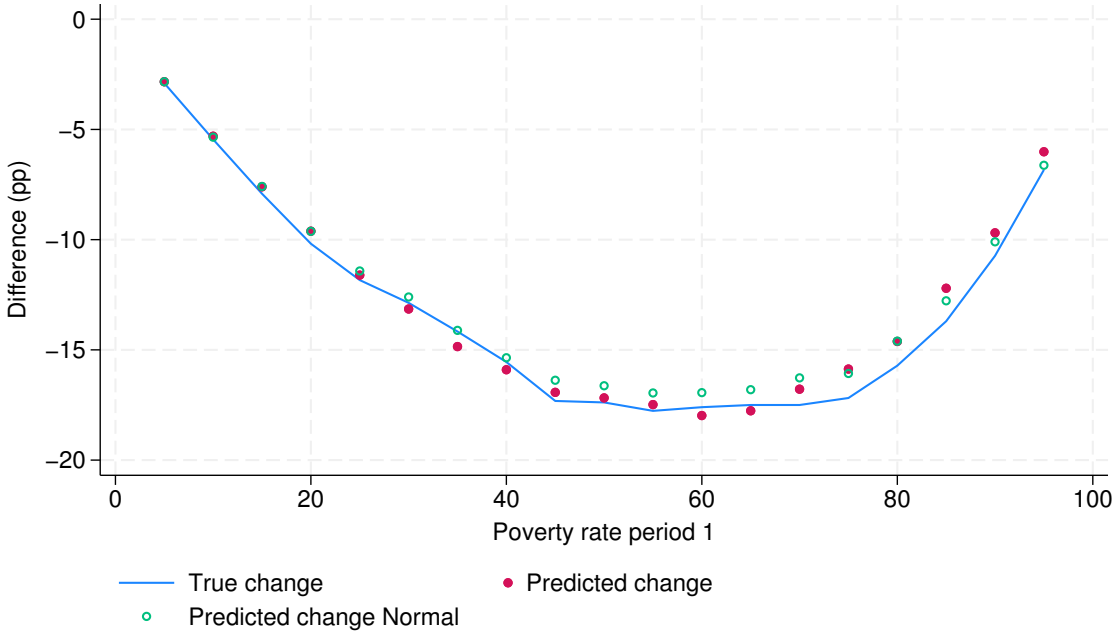
Note: Data are generated as described in 4.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

A common approach for conducting imputations involves predictive mean matching (PMM). The method combines linear regression with nearest-neighbor imputation to provide plausible imputed values. It first runs an linear regression on the training dataset[23] to estimate regression coefficients and calculate predicted values for both the training and imputation datasets. For each missing value, PMM identifies donor observations from the training data with predicted values closest to that of the case requiring imputation, then randomly selects one and uses its actual observed value as the imputed value. This approach ensures imputed values are realistic, reflecting actual data points, and is particularly suited for handling non-normal or discontinuous distributions (Yoshida et al., 2021). However, PMM's reliance on existing data means it cannot impute values outside the range observed in the training dataset, which may limit its applicability in certain contexts (ibid).

A simulation is run to test the performance of PMM. 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. This is only done for the target survey. The model is fit on the original data, but the imputations are made on the adjusted data. Under this scenario, PMM does a decent job of predicting and follows OLS predictions (Figure 4). Nevertheless, as noted by Yoshida et al. (2021), PMM is not able to impute values outside the oserved range and should be used with caution when there are large economic changes between modeling and imputation periods. An additional simulation is run where there has been a 20 percent growth between the source and target data periods (Figure 5). Under this scenario, PMM fails to capture the change.

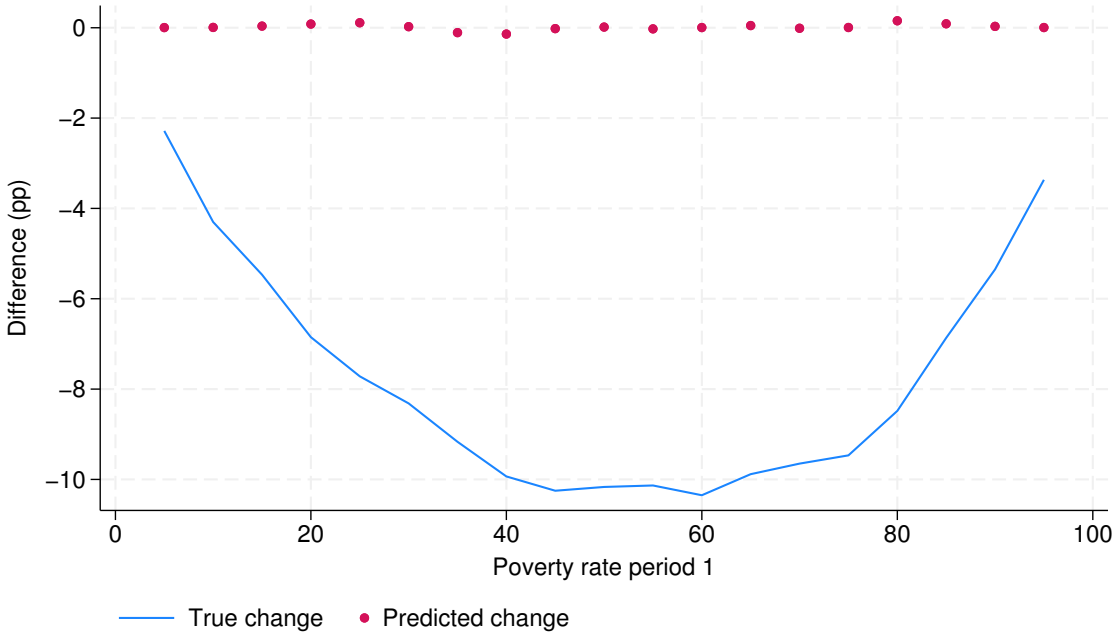
[23] Usually OLS, see (Lucchetti et al., 2024) for an application to lasso model fitting.

Figure 4: 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.

Figure 5: Difference in FGT0 under PMM imputations

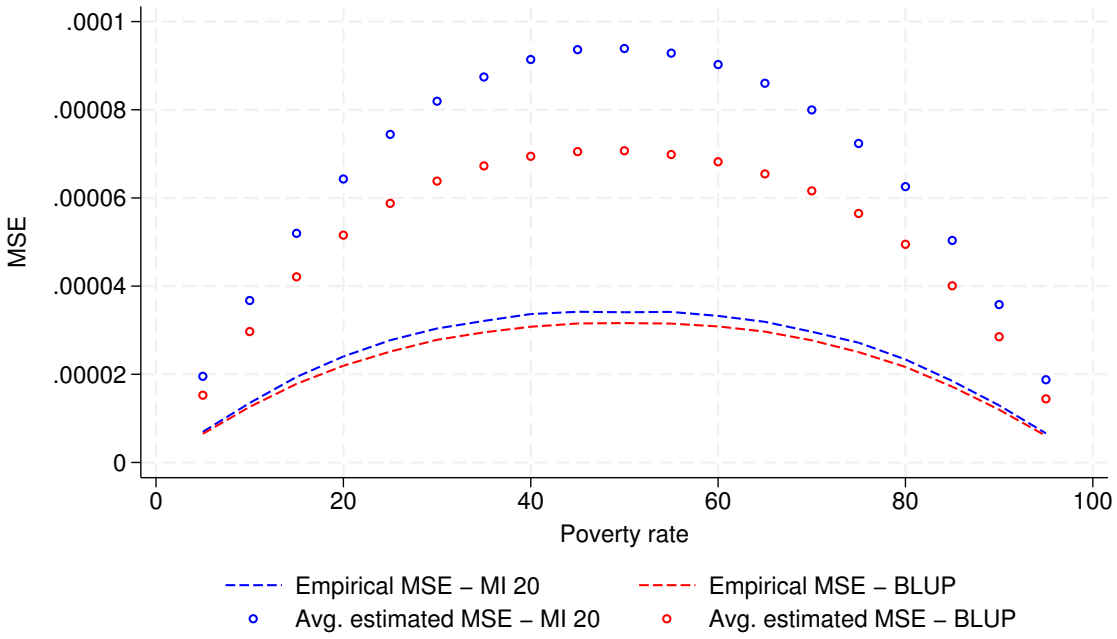


Note: Data are generated as described in 4.1. Difference is assessed at various poverty lines across the welfare distribution of the source data. Specifically, these lines correspond to percentiles that are multiples of 5.

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.[24] 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 here to obtain a valid standard error. Estimating the prediction's noise in a similar manner when parameters are applied directly risks producing improper imputations.[25] 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 is not necessarily aligned for S2S.

Figure 6: 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 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 training data. To compare the noise estimates of each method, 10,000 populations are created. Under each population a new error is drawn form 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.[26] The reason for the overestimation is that both methods measure the noise based on the target survey. The MSE is equal to the sum of the squared bias and the variance of the indicator of interest. Since both methods are unbiased, it all 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.

[24] 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.

[25] 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.

[26] 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.

4.3 Imputations under biased sample

The purpose of this section is to test the implications of under-sampling segments of the population and then adjusting the target data to correct for this bias. Two approaches are tested. The first is creating or adjusting survey weights so that totals or moments of the distribution match those of the underlying population. The second approach involves standardizing covariates across surveys so that the mean the standard deviation matches that of the unbiased sample.

4.3.1 Re-weighting the sample

Re-weighting is applied by Roy and Van Der Weide (2022) for the case of India where there were concerns regarding the sampling of the target survey. Weights are also frequently recalculated to adjust phone surveys like is done in the case of Afghanistan (Barriga-Cabanillas et al., 2023).[27] Although referred to as a max-entropy approach by Roy and Van Der Weide (2022) and Zhang et al. (2023), the method used to adjust the target sample weights is in reality minimum cross-entropy where the sampling weights in the target survey are adjusted by the smallest extent possible given the constraints (Wittenberg 2010).[28] The constraints, in the case of India's imputation, correspond to the mean of a set of variables that are found in other nationally representative surveys.

To test the problems with under-sampling and re-weighting, two sampling scenarios are used as an illustrative example. The samples are taken from a population created as described in section 4.1.1. For the re-weighting exercise, three possible re-weighting scenarios are considered:[29]

1. Covariate match: The simple random sample's mean of each covariate is used to calibrate the weights of the biased samples. For re-weighting, I use minimum cross-entropy to adjust prior weights by the smallest amount possible so that the constraints (i.e., the means) are satisfied.

2. Linear fit match: The linear fit's mean of the simple random sample is 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 means) are satisfied.

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.[30]

4.3.2 Standardizing covariates

The basics of standardizing covariates is simple. In principle, from the data where the model is fit (source or training data) the mean and variance of each covariate is known. Consequently, covariates in the sample to be imputed (target data) can be adjusted so that the mean and variance of the covariates

[27] Zhang et al. (2023) present how sampling weights were calculated to correct the sampling and nonresponse biases of phone surveys.

[28] 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).

[29] Reweighting 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` providing solutions to instances where `maxentropy` fails.

[30] 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.

are aligned to that of the source data. This is what H.-A. H. Dang et al. (2014) recommend for making different survey samples more comparable. However, a key assumption here is that the covariates follow a normal distribution. In tests done by H.-A. H. Dang et al. (2014) with data from Jordan the method is judged to be valid.

To test the validity of covariate standardization, the same samples detailed in section 4.1.1 are used. Note, however, that the covariates here do not follow a normal distribution. Covariate standardization is done by adjusting the covariate from the *new* survey in the following manner:

$$x_{new} = (x_{new} - \hat{\mu}_{x_{new}}) \frac{\hat{\sigma}_x}{\hat{\sigma}_{x_{new}}} + \hat{\mu}_x$$

4.3.3 Results

To remove the potential for other sources of bias, instead of refitting a model on the sample, the actual values for the coefficients β and the error distribution σ_e^2 are used. For each sample, under each population, 100 vectors in the sample are created by adding the linear fit to a randomly drawn error term $e_i \sim N(0, 0.5^2)$ to every observation. Then, poverty and Gini estimates are obtained for each of the 100 created vectors using the newly created survey weights or the standardized covariates to calculate poverty and Gini for the samples. The average value of the 100 Gini and poverty estimates is used as the final estimate. Bias is calculated for each measure of interest by comparing the estimate with the actual value from the population. In total, 1,000 populations are created and 1,000 imputations are made.

If left unadjusted, poverty estimates under the bottom biased samples will underestimate poverty across the entire welfare distribution (Figure 7). Attempts to correct the bias via the creation of weights that match the means of the model covariates ("Match X") do not have any impact on the bias.[31] Adjusting weights so that the covariate means and variances match also seems to have little impact on the bias ("Match X and var[X]"). However, adjusting weights so that the mean linear fit of the target data matches the average of the linear fit of the model ($\hat{y}$) leads to considerably better results ("Match XB and var[XB]"). Similarly, much better results are achieved by standardizing each covariate "Standardize each X", as well as "Standardizing XB". The latter achieves a similar result to adjusting weights so that the mean of the linear fit and its variance matches that of the training data. In principle, both methods are attempting to achieve the same thing. The results presented here are related to Eq. 3, thus what matters for poverty is not just the mean of the linear fit, it is also dependent on the variance of the linear fit.

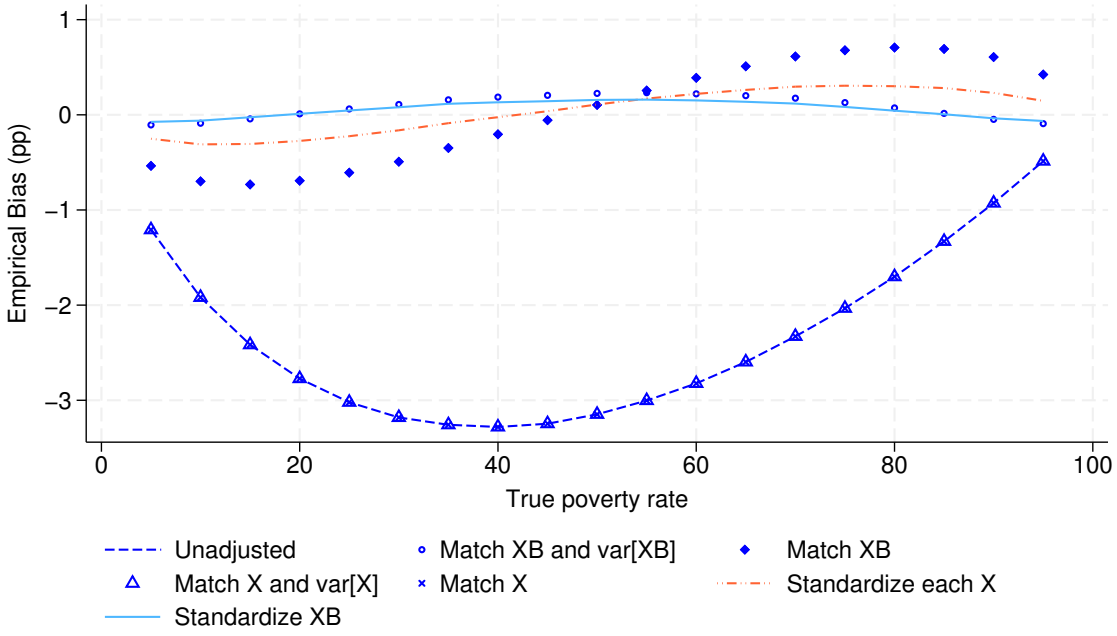
Similar performance of the different adjustment methods is observed when applying these to top and bottom biased samples (Figure 8). An interesting outcome in these results is that if the model were only judged at the 40th percentile (i.e. at a threshold that yields a poverty rate of 40 percent) it would ignore the biased nature of the samples. For this reason, it is always recommended to check results across the entire welfare distribution, like is done here.

The results show that standardizing covariates as suggested by H.-A. H. Dang et al. (2014) may be just as good or a better approach than re-weighting covariates, of course this result may not hold under different covariates or DGPs. However, even better results are obtained by standardizing the linear fit ($X\beta$). Standardizing $X\beta$ yields phenomenal results under both biased samples tested. Given that the standardizing relies on normally distributed data, perhaps a better option may be achieved from re-weighting in other scenarios.

[31] 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.

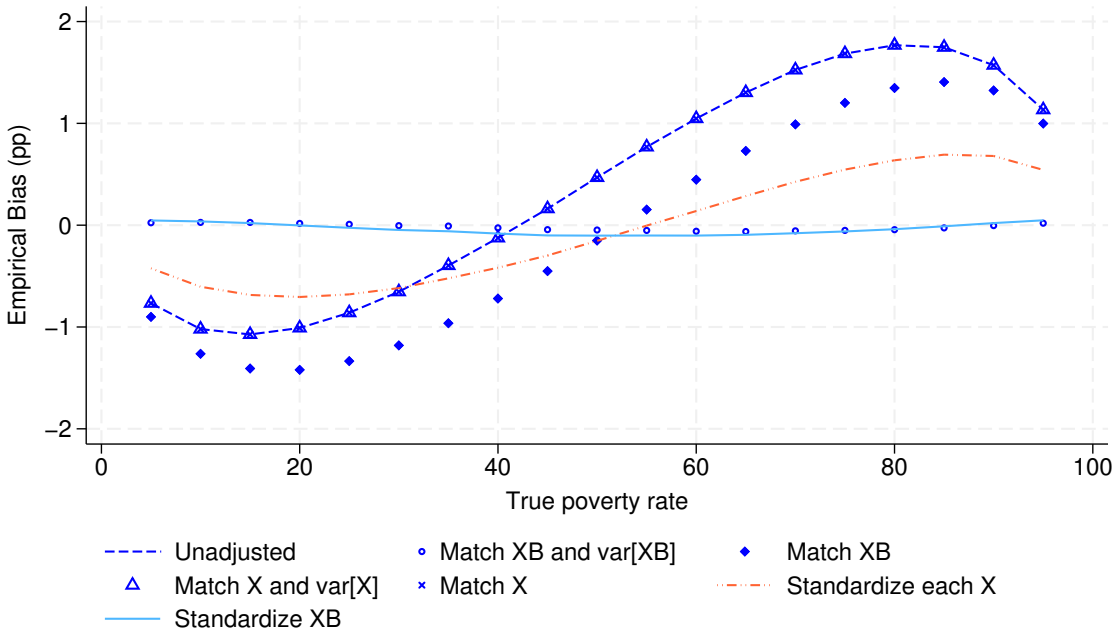
Results presented in Figures 7 and 8 suggest that not all types of re-weighting work well. Re-weighting to match means, as done in an application in India by Roy and Van Der Weide (2022), still yields biased results. Since the β and σ_e^2 used are always the true values, one can surmise that the bias from the sampling cannot be overcome by simply re-weighting the data. This also holds true for cases where the weights are adjusted so that the linear fit in the biased sample matches the mean of the SRS data. This sampling bias will be in addition to any other potential sources of bias, for example, non-normally distributed residuals or using a σ_e that does not correspond to the same period. The bias is in addition to all the other potential sources of bias of the imputation.

Figure 7: Bias in FGT0 under bottom biased samples (Sec 4.1.1) and different correction measures



Note: Target samples are generated as described in 4.1.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

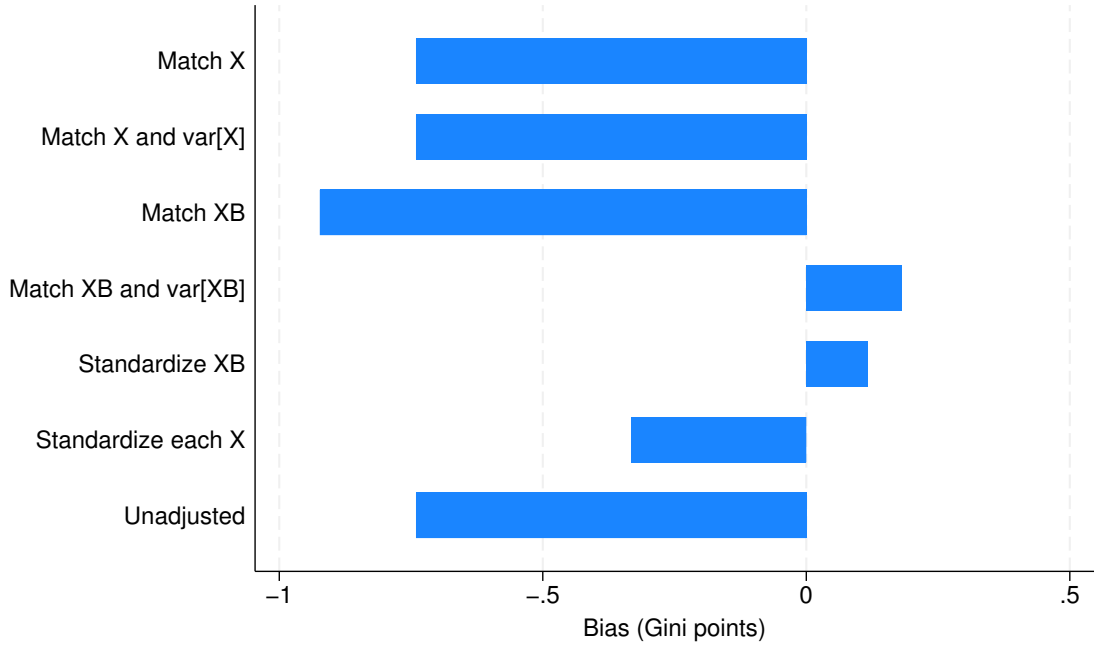
Figure 8: Bias in FGT0 under top and bottom biased samples (Sec 4.1.1) and different correction measures



Note: Target samples are generated as described in 4.1.1. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

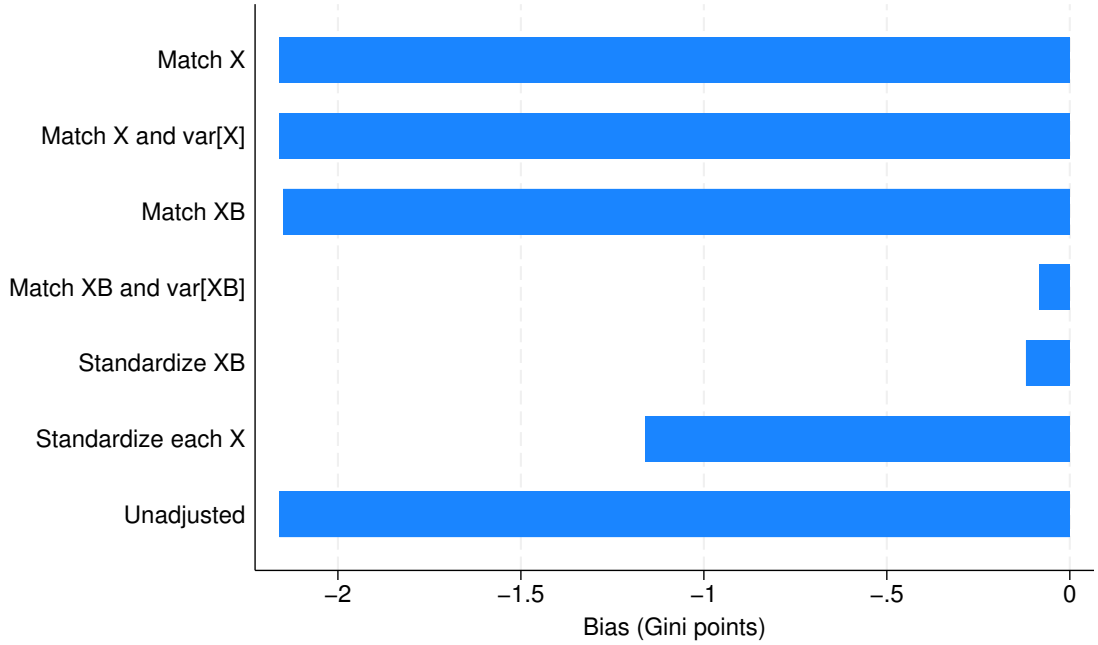
For Gini, a similar issue is at play. Because biasing a sample by excluding households at the bottom of the distribution will likely lead to lower var$[x\beta]$, which is not addressed by the re-weighting procedure used, keeping everything else equal leads to a downward biased estimate of Gini (Figure 9). This is worse under the top and bottom biased samples because the value of var$[x\beta]$ is considerably smaller than under the bottom biased sample (Figure 10). Consequently, adjusting the weights so that not just the mean of $X\beta$ matches the value observed in the SRS but also var$[x\beta]$ yields much better imputed values for Gini.

Figure 9: Bias in Gini under bottom biased samples (Sec 4.1.1) and different correction measures



Note: Target samples are generated as described in 4.1.1. Bias is assessed by comparing the mean predicted Gini against the mean Gini of the populations generated (i.e., the truth).

Figure 10: Bias in Gini under top and bottom biased samples (Sec 4.1.1) and different correction measures



Note: Target samples are generated as described in 4.1.1. Bias is assessed by comparing the mean predicted Gini against the mean Gini of the populations generated (i.e., the truth).

As encouraging as these results are, it is not possible to apply these to S2S imputation across time. The reason for this is because the true or a valid estimate of the mean for the dependent variable (which would be equal to the mean linear fit) or the variance of the linear fit would be unknown.

4.4 Imputation Over Time

For accurate poverty measurement, countries frequently resort to collect surveys on the living standards of its population. These surveys are costly and their implementation is complex. Thus, data is infrequently collected for the world's poorest economies. It is for these instances where S2S's appeal is heightened. The main applications of S2S over time attempt to:

1. Produce a poverty estimate that is comparable. There are many instances where a country may collect data on living standards, however due to any number of issues, the welfare aggregate is not comparable to a previous one. S2S has been applied in these scenarios by training a model on the original welfare aggregate and applying this to more recent survey (see the imputation of Zambia – Yoshida and Aron (2024)).

    (a) Note that this could also be done backwards. If the preferred welfare aggregate corresponds to the new survey then the model could be trained on the new data and its parameters applied to the old data to predict poverty (see the imputation of Nigeria – Lain et al. (2022))

2. Produce a poverty estimate when data on living standards is unavailable. In most instances where S2S is applied the country has recently collected data that does not include a comparable welfare aggregate, e.g., a recent Demographic Health Survey or a Labor Force Survey. In these instances, the model is trained on the last existing living standards survey with an adequate welfare aggregate and its estimated parameters are applied to the more recent survey – i.e. the target survey. The target survey is assumed to be nationally representative, but this is not always the case – for example, due to sampling bias (see the work of 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 training data and may not reflect the target period accurately.

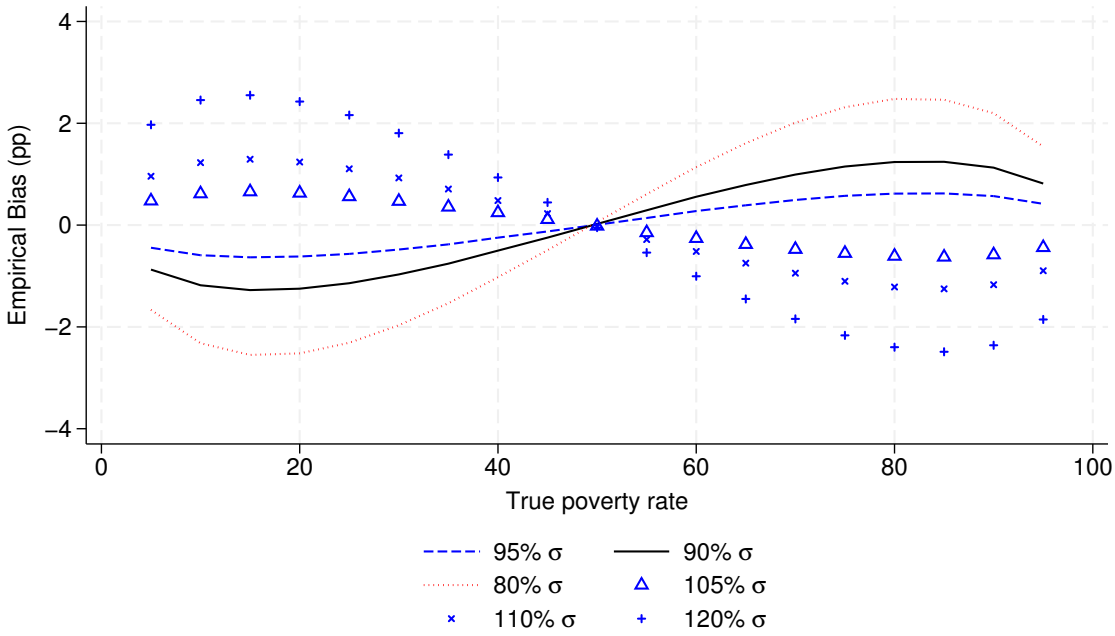
In this section the discussion focuses on how changes in the parameters can affect the poverty predictions. Through simulated data components can be adjusted one at a time to identify how these may impact predictions. In practice, it is possible that any of the parameters used for the imputation (β, 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.

4.4.1 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 unobservables. 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. Results are presented in Figure 11, 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 threshold 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.

Figure 11: Change in poverty prediction if σ changes by $x\%$

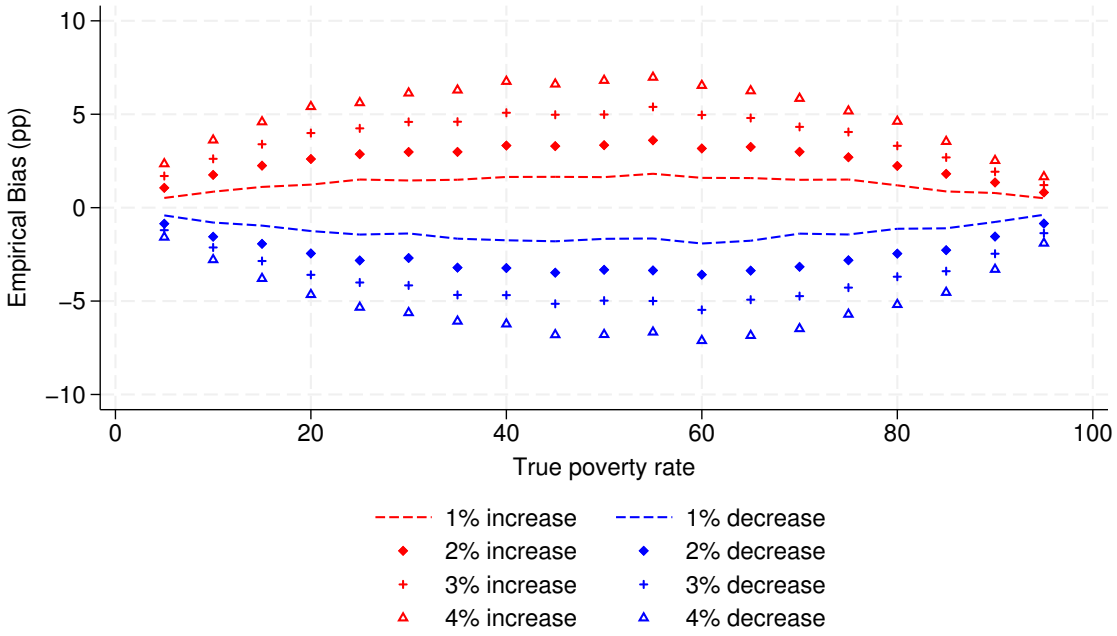


Note: Target samples are generated as described in 4.1.1, only the SRS samples are used for this simulation. Bias is assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

4.4.2 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. In Figure 12 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 would predict 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 12: 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.

4.4.3 Omitted Variables

Omitted variables and endogeneity have traditionally not been major concerns in S2S or small area estimation. This is because, in the current period, any omitted variables are accounted for by the constant term, which adjusts 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 5.1).

The simulations presented here take as an example the S2S application of Afghanistan (Barriga-Cabanillas et al., 2023). The authors presume that the decrease in poverty predicted in rural areas is driven by a reduction in conflict. The models do not control for conflict. Additionally, the authors follow recommendations of the SWIFT program and include "fast-changing consumption variables to better capture welfare changes during shocks" (Barriga-Cabanillas et al. 2023, p6). Among the variables the authors include to the model are a list of food consumption dummies that include meat, eggs, and chocolate. However, it is quite likely that the model used by the authors suffers from omitted variable bias. Mainly, conflict is likely negatively correlated to expenditure and is also likely to be negatively correlated to these consumption dummies as it may affect the availability of these goods. The discussion in the annex (Annex 5.1) illustrates that in this case, the omission of conflict from the model is likely to lead to biased estimates of poverty due to omitted variable bias (OVB).

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) It is simulated as a binary variable taking value 1 when a random uniform number between 0 and 1 is less than $(0.5x + 0.5Conflict)$, with $x = 1$ in the baseline. The DGP for the covariate illustrates the negative correlation of conflict and the likelihood of purchasing eggs.

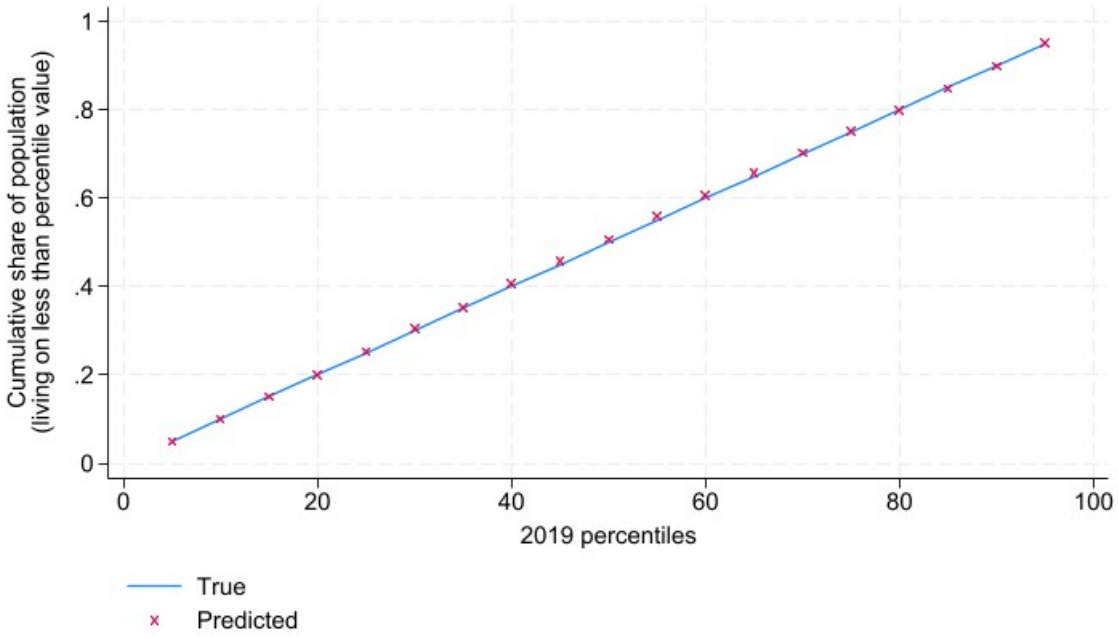
Under the baseline simulation, the value for $x = 1$ and $c = 0.4$. Welfare is simulated assuming the true DGP, and thus includes conflict and eggs with their true coefficients. The prediction model is produced omitting conflict and is executed using the complete data to avoid other potential sources of bias (Table 1). As presented in Annex 5.1, 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, which will likely affect poverty predictions over time. Nevertheless, for predictions to the same period OVB has little if any impact (13).[32]

Table 1: Omitted variable model comparison

	OVB Model	True Model
x1	0.1022	0.1017
x2	0.4965	0.4970
x3	-0.2492	-0.2481
x4	-0.1997	-0.1995
x5	-0.1423	-0.1462
eggs	0.5745	0.3927
conflict		-0.3182
Intercept	2.8264	3.0092
R2	0.5327	0.5634
RMSE	0.5120	0.4949
Observations	20,000	20,000
$\hat{y}$	2.9826	2.9826

[32]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 13: Resulting poverty prediction under OVB in the same period



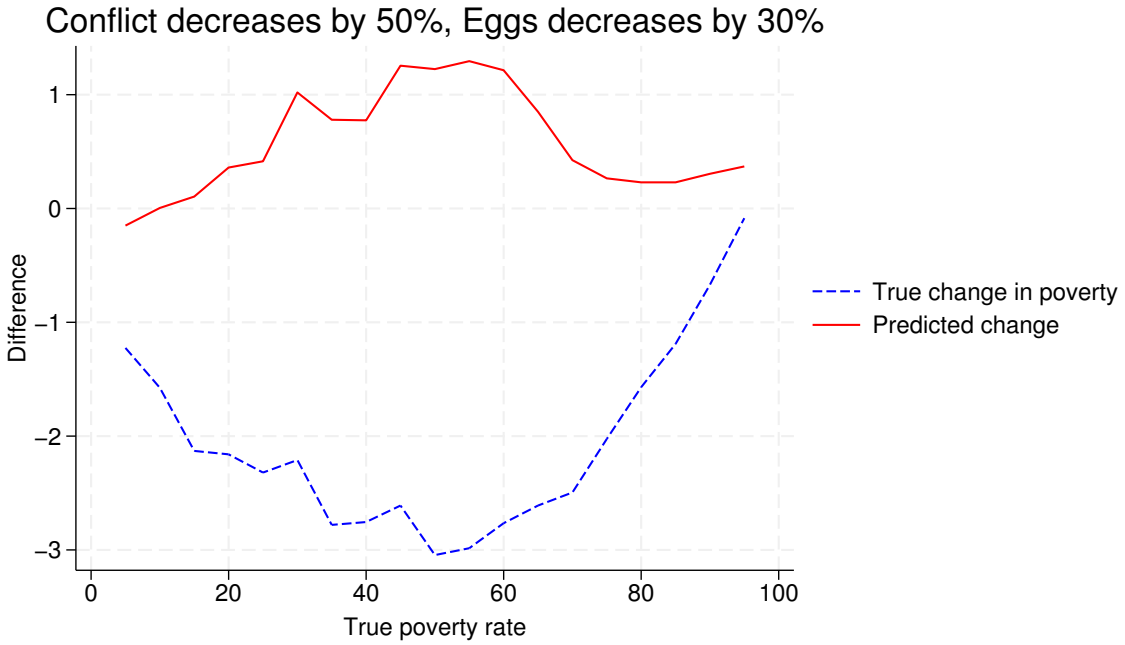
Source: Based on simulated data illustrated in Section 4.1 with added covariates. Predictions are assessed at various poverty lines across the welfare distribution. Specifically, these lines correspond to percentiles that are multiples of 5.

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 5.1, 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.[33] 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 14). 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.

[33]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 14: 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.

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.

Bank, W. (2020). *Poverty and shared prosperity 2020: Reversals of fortune*. https://doi.org/10.1596/978-1-4648-1602-4

Barriga-Cabanillas, O., Chawla, P., Redaelli, S., & Yoshida, N. (2023). *Updating poverty in afghanistan using the swift-plus methodology* (tech. rep.). The World Bank.

Battese, G. E. (1997). A note on the estimation of cobb-douglas production functions when some explanatory variables have zero values. *Journal of agricultural Economics*, *48*(1-3), 250–252.

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.

Bhalla, S., Bhasin, K., & Virmani, M. A. (2022). *Pandemic, poverty, and inequality: Evidence from india*. International Monetary Fund.

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

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*. The World Bank. https://doi.org/10.1596/1813-9450-8282

Dang, H.-A. H., Kilic, T., Carletto, C., & Abanokova, K. (2021). Poverty imputation in contexts without consumption data: A revisit with further refinements.

Dang, H.-A. H., Lanjouw, P. F., & Serajuddin, U. (2014). *Updating poverty estimates at frequent intervals in the absence of consumption data: Methods and illustration with reference to a middle-income country*. The World Bank. https://doi.org/10.1596/1813-9450-7043

Deaton, A. (2003). Adjusted indian poverty estimates for 1999-2000. *Economic and political Weekly*, 322–326.

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).

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

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)*.

Mahler, D., Yonzan, N., Lakner, C., & Castaneda Aguilar, H., R.A. abd Wu. (2021, June). Updated estimates of the impact of covid-19 on global poverty: Turning the corner on the pandemic in 2021? https://blogs.worldbank.org/opendata/updated-estimates-impact-covid-19-global-poverty-turning-corner-pandemic-2021

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).

Prydz, E. B., Jolliffe, D., & Serajuddin, U. (2022). Disparities in assessments of living standards using national accounts and household surveys. *Review of Income and Wealth*, *68*, S385–S420.

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

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

Ravallion, M. (2003). Measuring aggregate welfare in developing countries: How well do national accounts and surveys agree? *Review of Economics and Statistics*, *85*(3), 645–652.

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. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

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.

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 Food Programme (WFP). (2023, May). *Wfp afghanistan: Situation report, 24 may 2023* (Accessed: 2024-12-06). https://reliefweb.int/report/afghanistan/wfp-afghanistan-situation-report-24-may-2023

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.

Zhang, K., Takamatsu, S., & Yoshida, N. (2023). Correcting sampling and nonresponse bias in phone survey poverty estimation using reweighting and poverty projection models.

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

5 Annex

5.1 The Case of Afghanistan - Omitted Variable Bias

The Afghanistan team's discussion inadvertently demonstrates why survey-to-survey imputation across time periods requires careful consideration. While Barriga-Cabanillas et al. (2023) report overall poverty reduction driven by rural improvements—despite GDP contraction, reduced aid, and a locust outbreak—they attribute this to decreased conflict.[34] However, since conflict is not explicitly included in their 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 eachother out or compound.

A key aspect of the team's imputation is the inclusion of fast changing consumption variables, such as an indicator of whether or not the household consumed eggs over the past week. Barriga-Cabanillas et al. (2023, p2) follow the proposed approach from Yoshida et al. (2022) and argue that these dummies are included "to better capture welfare changes in a context where large economic shocks have occurred". This is what the authors refer to as the "SWIFT Plus methodology". 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 Afghanistan imputation the omitted variable is conflict. In others it may be the introduction of a cash transfer, for example.

If 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:

$$Y = \beta_0 + \beta_1 Eggs + \beta_2 Conflict + \varepsilon \tag{4}$$

By omitting conflict the model is now:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 Eggs + u \tag{5}$$

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:

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

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

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 $\text{Cov}(Eggs, Conflict) < 0$. 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:

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

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

$$\tilde{\beta}_0 = \beta_0 + \beta_2 \overline{Conflict} + (\beta_1 - \tilde{\beta}_1) \overline{Eggs} \tag{6}$$

we know that $\beta_2 \overset{`}{<} 0$, and that $(\beta_1 - \tilde{\beta}_1) < 0$. This suggest 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. 6, 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 ($RMSE$), 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 Afghanistan exercise.

5.2 Parametric Bootstrap to Estimate Noise

<IN PROGRESS - Need to update with simulations comparing MI noise and PBS noise estimation. I was just testing this for EU SAE team>

Many teams engaged in poverty measurement and welfare analysis face challenges when performing Small Area Estimation (SAE) using datasets that are not derived from a full population census. These challenges often require the use of alternative data sources, such as a 10 percent sample of the census or a larger survey representative at the desired geographic or administrative level. Historically, such tasks were approached as a survey-to-survey imputation exercise, which accounted for both within-survey variation (arising from sampling) and between-survey variation accounting for the imputation.

However, recent updates to the censusEB methodology have introduced changes in noise estimation that warrant careful reconsideration of this approach. To understand the potential bias in noise estimation, simulations are to evaluate the extent to which noise is over- or underestimated when imputing to a sample of the population. Key findings emerged from two distinct experiments:

Parametric Bootstrap Noise Estimation

Using multiple samples from a simulated census as the target data, welfare is imputed to these and poverty rates are estimated. The source data, i.e. the training data, is a 5% sample. The multiple samples taken from the census are then used as the target data used to simulate income and estimate poverty. The estimated MSE(x10000) is presented below, the column represent the sample size of the target survey.

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

sample_size	p25	p50	p75	Mean	Min	Max
10	4.021938	4.343824	4.70552	4.379047	3.119347	6.185248
20	2.871218	3.079045	3.335209	3.092172	2.256884	3.970806
30	2.49273	2.65838	2.823314	2.672387	2.013348	3.329488
40	2.253806	2.424606	2.598804	2.447872	1.911704	3.76836
50	2.146552	2.349193	2.470929	2.334104	1.730694	3.024491
60	2.062104	2.206561	2.374666	2.217627	1.67711	2.99695
70	2.007197	2.156962	2.306258	2.162625	1.506513	3.050986
80	1.963778	2.115128	2.293958	2.12345	1.429528	2.739508
90	1.933341	2.071389	2.229076	2.093298	1.629131	2.777996
100	1.935832	2.075467	2.212728	2.074826	1.452233	2.516723

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.

Model-Based Noise Estimation

A more comprehensive simulation was conducted to derive "true" noise estimates: Step 1: Generated 5,000 populations with fixed covariates. Step 2: Extracted a 5% sample from each population to represent the survey. Step 3: Drew additional smaller samples (e.g., 10%, 20%, etc.) as target datasets, imputed poverty rates using the SAE model, and compared these to the true census-derived poverty rates. Findings revealed that the true Mean Squared Error (MSE) was consistently smaller than the MSE estimated using parametric bootstrap methods. <The key to understand what is driving this is the MSE formula which is a function of the variance and the bias. Under the true census, the variance is smaller since the N is larger, and when using a smaller data as the target data for the imputation the variance is overestimated while bias is unaffected.>

Figure 16: Resulting poverty prediction under OVB in the same period

sample	p25	p50	p75	Mean	Min	Max
10	2.175637	2.587618	3.492377	3.198957	1.934621	11.73034
20	2.101218	2.245922	2.603549	2.491091	1.919861	5.550759
30	2.088099	2.226638	2.473215	2.346288	1.901716	3.919386
40	2.082568	2.177601	2.329175	2.236822	1.903696	3.433073
50	2.049648	2.112758	2.190216	2.147656	1.916898	2.945403
60	2.039034	2.097432	2.157243	2.113996	1.91293	2.756133
70	2.03332	2.083001	2.13195	2.091111	1.930297	2.508838
80	2.024282	2.066572	2.120934	2.067828	1.922601	2.255446
90	2.021329	2.053147	2.106505	2.06184	1.903525	2.246271
100	2.011459	2.045698	2.086494	2.047698	1.899568	2.244449

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.

Implications for Practice

The consistent overestimation of noise using parametric bootstrap is significant. While overestimating noise may appear conservative, it introduces inefficiencies that could undermine the precision of SAE outcomes. Encouragingly, the model-based approach provides more accurate noise estimates, aligning predictions more closely with true values derived from the census.

Practical Considerations

For practitioners leveraging Stata SAE, it is worth noting that the tool allows incorporation of sample weights (e.g., for labor force surveys) through the pwcensus() option. Ensuring proper specification of weights can further improve the reliability of SAE results.

Why This Matters

Accurate estimation of welfare and poverty indicators at granular levels is critical for effective policymaking and resource allocation. Misestimated noise—whether underestimated or overestimated—can distort the interpretation of SAE outputs, potentially leading to misguided interventions. By refining noise estimation approaches and leveraging robust methodologies, we can ensure that small area estimation delivers actionable insights that align with the realities of vulnerable populations.