



Centro de Estudios®
Espinosa Yglesias

PROMOVEMOS LA IGUALDAD
DE OPORTUNIDADES

Survey to Survey Imputation when External Covariates Matter: Estimating Inequality of Opportunity in Mexico

Autores:


Pedro J. Torres L.
*Department of Social Policy and III -
LSE, London*

Luis A. Monroy-Gómez-Franco
*Department of Economics -
University of Massachusetts, Amherst*

Roberto Vélez-Grajales
*Centro de Estudios Espinosa Yglesias,
CDMX*

Documento de trabajo núm.

01/2025

Centro auspiciado por:  ESRU
FUNDACION ESPINOSA YGLESIAS

Survey to Survey Imputation when External Covariates Matter: Estimating Inequality of Opportunity in Mexico¹

Pedro J. Torres L.²

Luis Á. Monroy-Gómez-Franco³

Roberto Vélez-Grajales⁴

March 2025

Accurate measurement of income and consumption is essential for understanding socioeconomic disparities and informing effective policy interventions. However, survey data often lack comprehensive income or consumption data. Researchers typically employ asset-based indices or survey-to-survey imputation techniques to address this limitation. While these methods provide valuable insights, they can introduce biases, mainly imputations, when the analysis involves covariates not used in the prediction process. This paper proposes a novel imputation methodology that aims to mitigate these biases. Our approach focuses on preserving individuals' relative positions within the income distribution while introducing variance. This ensures that the underlying rank order of individuals is maintained, addressing a common shortcoming of existing imputation techniques. Through rigorous validation and comparison, we demonstrate the robustness and effectiveness of our proposed method using the ENIGH survey in Mexico. To illustrate one practical application of our approach, we analyze inequality of opportunity, imputing income from the ENIGH into the ESRU-EMOVI survey. Our findings underscore the importance of carefully considering imputation methods in socioeconomic research. We demonstrate that traditional imputation procedures can lead to downward biased estimates of inequality of opportunity compared to our proposed method.

Keywords: distribution analysis, imputation, data

JEL Classification: C40, C53, D63

----- **Documento de Trabajo CEEY núm. 01/2025** -----

Los resultados, interpretaciones y opiniones expresadas en este documento son responsabilidad de sus autores y no reflejan necesariamente la postura del CEEY y sus entidades afiliadas.

Publicado bajo una Licencia Creative Commons Atribución-No Comercial 4.0 Internacional ([CC BY-NC 4.0](https://creativecommons.org/licenses/by-nc/4.0/)).



¹ This document is the product of a collaboration between the Centro de Estudios Espinosa Yglesias (CEEY) and the London School of Economics (LSE) in the framework of the Global Estimates of Opportunity and Mobility (GEOM) project.

² Department of Social Policy and III - LSE, London. p.torres-lopez@lse.ac.uk

³ Department of Economics - University of Massachusetts, Amherst. lmonroygomez@umass.edu

⁴ Centro de Estudios Espinosa Yglesias, CDMX. rvelezg@ccey.org.mx

Introduction

A key challenge to distributional analyses in developing countries is that information on income and/or consumption is either completely unavailable or not present in surveys. In the first case, that of complete lack of information about income or consumption, alternatives such as household asset indices have been developed to fill that gap (Filmer and Pritchett 2001; Filmer and Scott 2012; Poirier *et al.* 2020). In the second case, several imputation methods have been developed to use auxiliary information to reconstruct the distribution of income or consumption implicit in surveys that lack information about such variables (Dang 2021).

Among the imputation methods developed, a substantial part of the literature has focused on developing regression-based imputations. These methods have their origins in the seminal contribution of Elbers *et al.* (2003) with the Small Area Estimation (SAE) procedure and can be broadly divided into two groups (Corral *et al.* 2022).

Design-based methods rely on the sampling design and survey weights to produce direct or indirect estimates for small areas without making assumptions about the underlying population distribution (Lehtonen and Veijanen 2009; Pfeffermann 2013). In political science, multilevel regression and post-stratification (MRP) has been widely used to obtain reliable percentages of subnational point estimates from nationally representative surveys (Park *et al.* 2017; Kiewiet de Jonge *et al.* 2018); or to produce accurate estimates from non-representative surveys (Wang *et al.* 2015; Downes *et al.* 2018). The idea is to estimate a multilevel model to predict an outcome and then weight the predictions based on the observed frequencies of a subgroup taken from the census in a post-stratification manner.

Model-based methods assume a statistical model that relates the variable of interest to auxiliary variables and accounts for random variation between and within small areas in a two-sample two-stage (TSTS) manner. A commonly used approach is the Random Effects Linear Model (Elbers *et al.* 2003; Pfeffermann 2013). Following the model proposed by Elbers *et al.* (2003), these methods have been applied to the analysis of poverty dynamics through synthetic panels (Dang and Lanjouw 2023a), the evolution of poverty in contexts where there is no data across time (Dang and Lanjouw 2023b; Sinha Roy and Weide 2023)

and poverty mapping (Elbers *et al.* 2003).

Three key assumptions must be met for model-based methods to produce reliable imputations. *i)* The two surveys have the same set of questions that explain our outcome of interest. *ii)* The functional form, specifically the coefficients, must be stable and consistent across surveys (or over time). *iii)* The available predictors are sufficiently correlated with our outcome.

Predictions of income and consumption typically concentrate around the conditional expected mean, often exhibiting lower variance than observed data due to the uncertainty introduced by the model's error term. This can lead to bias when estimating inequality and poverty measures. The downward bias arises for inequality measures because the predicted distribution displays lower variance than the actual distribution. For poverty measures, bias can occur if the predicted values cluster above or below the poverty threshold, resulting in a downward or upward bias.

Several techniques have been developed to address the challenge of preserving variance when imputing missing values. Two widely used methods for managing item non-response in surveys are Hot Deck imputation and Predictive Mean Matching (PMM) (Campion and Rubin 1989). Hot Deck imputation addresses missing values by copying data from a similar, fully observed case. PMM extends this approach by using a statistical model to predict missing values, identifying similar cases based on these predictions, and randomly selecting a donor to impute the missing information.

Hot Deck imputation and PMM are generally effective for handling item non-response, especially when the missing data is at random, as they help preserve the observed mean and variance (Suárez-Arbesú *et al.* 2024). However, their effectiveness diminishes significantly when entire variables are missing for a subset of observations (unit non-response) where the randomness of the missing values does not hold.

When a variable of interest is missing from a survey, a two-sample two-stage procedure is often employed. In the first stage, a model is estimated using data from one sample and then used to predict values in a second sample with the same set of covariates (model-based). The second stage involves analysing the second sample using the predicted values.

In poverty dynamics, for example, cross-sectional data do not allow individuals to be

tracked over time to estimate the probability of moving in and out of poverty. Researchers address this using the TSTS approach, first estimating a model for income in the second latter survey and then predicting income in the first survey using the model to assess the probability of individuals moving into or out of poverty in a pseudo panel manner (Dang and Lanjouw 2023b; Sinha Roy and Weide 2023). In the context of intergenerational mobility, most lower-income countries lack surveys linking parental income with their offspring, making it challenging to estimate intergenerational elasticities (Björklund and Jäntti 1997; Bloise *et al.* 2021). In these cases, the TSTS approach is commonly used to predict parental income in a survey, allowing for the estimation of intergenerational mobility.

To more accurately reflect variance, inequality, and poverty measures, a random error drawn from the empirical distribution of the model is added to each predicted value (McKenzie 2005; Dang and Lanjouw 2023a,b). This approach helps recover variance lost during estimation by incorporating the uncertainty associated with the error term. The process is repeated using a bootstrap approach (Rodas *et al.* 2021), which enables the calculation of a mean and a confidence interval for key variables of interest, such as the Gini coefficient and poverty rates, this procedure is typically referred to as the Multiple Imputation Procedure (MIP).

The Problem

Often, the analysis in the second stage requires information not used in the imputation process— in other words, external covariates— which are used in the second stage after the imputation. When imputing income (or consumption) y from one survey into another, in the first stage, we model y as a function of available predictors X , such that $y = f(X) + \varepsilon$.

Predicting y and incorporating variance through the error term results in $\hat{y} = \hat{f}(X) + \tilde{\varepsilon}$, where $\tilde{\varepsilon}$ is a random error drawn from the empirical distribution of the model’s residuals. The second stage is assessed by examining $cov(y, Z)$, where Z represents covariates of interest that are not observed in the source survey and, therefore, not included in X ($Z \notin X$). The true relationship can be expressed as:

$$cov(y, Z) = cov(\hat{y}, Z) + cov(\tilde{\varepsilon}, Z)$$

Where the error terms in the first stage are highly likely correlated with Z ($cov(\hat{\varepsilon}, Z) \neq 0$). However, since $\tilde{\varepsilon}$ is a random error drawn from the model, it is constructed to be orthogonal to Z due to its random nature. Thus, we have $cov(\tilde{\varepsilon}, Z) = 0$, which reduces the true relationship between y and Z as the second part of the relation is lost in the imputation process. This results in an underestimation of the impact of covariates on the outcome y , effectively downward biasing our estimations.

For the analysis in the second stage, we need to predict y in such a way that:

$$cov(y, Z) = cov(\hat{y}, Z)$$

This equality holds when the error term $\hat{\varepsilon}$ is purely random and uncorrelated with z , meaning it does not affect the covariance structure in the second stage. Therefore, we aim to ensure that the imputation process renders the error term irrelevant for the relationship between y and Z . One way to formalise this is to ensure that the correlation between y and its imputed counterpart \hat{y} is close to 1, signifying that \hat{y} closely approximates y .

Considering the Spearman rank correlation, which measures the rank-order correlation between y and \hat{y} , we can look at the association between the predicted and actual values. Rather than focusing on the exact values, the Spearman correlation compares their ranks, which allows us to ignore variance differences that may arise due to the random error term. By ranking the data, we focus on the monotonic relationship between y and \hat{y} as:

$$\rho_{y,\hat{y}} = 1 - \frac{6 \sum (R(y_i) - R(\hat{y}_i))^2}{n(n^2 - 1)}$$

where $R(y_i)$ and $R(\hat{y}_i)$ denote the ranks of y_i and \hat{y}_i respectively. The term $\sum (R(y_i) - R(\hat{y}_i))^2$ captures the degree of rank mismatch between y and \hat{y} allowing for differences in the variance of both distributions stemming from a truly random error term. As this mismatch diminishes, the rank correlation approaches 1.

If we express y_i as $\hat{y}_i + \hat{\varepsilon}_i$ where $\hat{\varepsilon}_i$ represents the residual error term from the imputation

model, the Spearman correlation becomes:

$$\rho_{y,\hat{y}} = 1 - \frac{6 \sum (R(\hat{y}_i + \hat{\epsilon}_i) - R(\hat{y}_i))^2}{n(n^2 - 1)}$$

where $R(\hat{y}_i + \hat{\epsilon}_i) - R(\hat{y}_i)$ reflects the distortion in ranks due to the error term. As this difference approaches 0 (i.e., as the rank difference induced by the error term vanishes), the Spearman correlation approaches 1. This implies that the imputed values \hat{y}_i become a near-perfect rank-preserving transformation of the actual values y_i while still allowing for the variance to differ between y and \hat{y} when the error term is truly random.

Consequently, the extent to which we can preserve the rank in our predictions is directly related to the bias introduced in our second-stage estimates. The preservation of rank ensures that the relationship between y and \hat{y} remains robust, thereby minimising any distortion in the covariance estimates concerning the external covariates Z .

Given that the prediction may inherently involve a variance loss due to a truly random error term, our primary objective shifts to matching the observed variance. By achieving this, we can confidently assert that $cov(y, Z) \approx cov(\hat{y}, Z)$, thereby enhancing the reliability of our second-stage analyses.

This paper proposes to address this problem by imputing income (or consumption) using a novel methodology based on a combination of model and design-based approaches. We provide validations to our approach by first applying it to two surveys where income is present. We impute income from the ENIGH 2016 to the ENIGH 2018 and vice versa to confirm the validity of the three assumptions. The results confirm that assumptions *i*, *ii* and *iii* are met.

We apply a log-linear specification for prediction— the model-based part of our procedure— as proposed by Elbers *et al.* (2003), McKenzie (2005), Ferreira, Gignoux, and Aran (2011), and Dang and Lanjouw (2023b). We then assess the bias introduced in the estimation process by comparing our predictions with the observed data across various population subgroups. To correct these biases, we adjust our predictions— the design-based part— similar to how post-stratification is used in MRP.

We evaluate four different imputations for predicting income: *i*) a simple prediction

without variance adjustment; *ii*) a cluster-specific (CS) adjustment, which accounts for the mean of each specific cluster; *iii*) a cluster-rank (CR) adjustment, which adjusts the mean at each percentile rank within each cluster; and *iv*) a final adjustment that combines the CS and CR adjustments using a convex combination of both approaches.

The original imputation predicts a Gini coefficient more than 10 points below the observed values for both surveys, consistent with the findings of Vélez-Grajales *et al.* (2019) who impute household income from the ENIGH 2010 into the ESRU-EMOVI 2011 without adding any variance. The first adjustment (CS) offers a slight improvement in the Gini coefficient estimation, with its main advantage being the reduction of RMSE by adjusting for cluster means. In contrast, the second adjustment (CR) incorporates variance and overestimates the Gini coefficient in both surveys, increasing the error in our predicted income. The linear combination of the CS and CR adjustments significantly improves the estimation of income distribution and inequality measures. This enhancement is evident in a Gini coefficient that closely aligns with the one observed in the ENIGH 2016 and 2018 survey estimates and in the improved accuracy of the observed distribution of household income.

Lastly, in a practical exercise, we impute household income from the ENIGH 2016 into the ESRU-EMOVI 2017 survey and adjust our predictions using imputation method *iv*. We compare inequality of opportunity (IOp) estimates using our imputed measure, the multiple imputation procedure, and an asset-based index. In this example, circumstances—our external covariates—are not observed in the source survey, therefore it works as a perfect example to demonstrate our methodology..

The two economic well-being measures— the asset-based index and our imputation— show a positive correlation of 0.45. Our adjustment yields a Gini coefficient of 0.51, significantly higher than that reported by Vélez-Grajales *et al.* (2019). The multiple imputation method produces a Gini coefficient of 0.47, closest to the ENIGH observations.

IOp estimates using the asset index and the multiple imputation method show a reduction of more than 10 points compared to our adjusted measure. The decrease in IOp when using the asset index is attributed to its closer relationship with consumption (Filmer and Scott 2012), which tends to be less volatile than income. The reduction in IOp when applying the multiple imputation method is likely due to the systematic weakening of the association

between circumstances not included in the imputation procedure and outcomes, suggesting that the random error term introduces a downward bias.

The rest of the article is structured as follows: Section 1 presents the framework on which we build our imputation approach; Section 2 discusses the data and the transformations made; Section 3 validates our method, followed by robustness checks in Section 4; In Section 5 we estimate IOp using the ESRU-EMOVI 2017 and our imputed measure; Finally, Section 6 concludes.

1 Methods

Consider two samples drawn from the same population, each from different surveys. The first survey, termed the "source survey" (denoted by superscript 1), includes income information but lacks data on covariates of interest, such as parental background. The second survey termed the "destination survey" (denoted by superscript 2), contains information on said covariates but does not include income data. Both surveys share a common set of predictors X .

Let y denote income or consumption for each observation i . Our objective is to impute y from the source survey (1) to the destination survey (2) using a model of the form:

$$y_i^1 = f^1(X_i^1) + \varepsilon_i^1 \quad (1)$$

where y is modelled as a function $f(*)$ of covariates X , which have predictive power over y and are specific to each observation. The error term ε is assumed to be normally distributed with a mean of 0.

Note that $f(*)$ is not intended to capture the effect of X on y , but rather to approximate (or predict) y as closely as possible using observable factors. Therefore, it is possible to specify $f(*)$ as a general function to be approximated. Once we have approximated our functional form, we estimate our predicted y in the destination survey as

$$\hat{y}_i^2 = \hat{f}^1(X_i^2) \quad (2)$$

The variance of the predicted values will typically be less than the variance of the true values due to the inherent uncertainty and imprecision introduced during the prediction process associated with the error term ε_i . This leads to a prediction that concentrates values around the conditional expected mean, implying a downward bias in poverty measures when predictions are concentrated above the poverty threshold and an upward bias when they are concentrated below it. Concentrating predictions will, by definition, lead to a reduction in inequality measures.

To recover variance, we propose imputing data from the source survey to the destination survey using a regression model (model-based) and employing adjustment ratios based on the empirical distribution of the source survey (design-based). Our imputation process consists of the following five steps:

1. **Adjusting Survey Weights:** We adjust the source survey weights to achieve a similar representativeness to that observed in the destination survey (Cowell *et al.* 2018).

Following DiNardo *et al.* (1996), we use a semi-parametric decomposition method to estimate the proportion of the difference between the source and destination surveys due to differences in the distribution of characteristics.

Using a logistic regression, we estimate the probability π_i that observation i with predictors X_i in the source survey is present in the destination survey. We adjust the source survey weights by multiplying them by the estimated probability $w_i^* = (\frac{w_i}{N} * \pi_i) * N$, where w_i is the original weight and N is the total number of observations.¹ Thus, observations highly likely to be in both surveys retain a weight close to their original weight, while those highly likely not to be in the destination survey have smaller weights.

These weights w^* are used in step 3 of our procedure to estimate weighted means.

2. **Estimating $f^1(*)$:** We estimate Equation 1 by using a log-linear specification following Elbers *et al.* (2003), McKenzie (2005), Ferreira, Gignoux, and Aran (2011), Dang and

¹We normalise the weights w^* so that the probabilities sum up to 1.

Lanjouw (2023a,b), and Sinha Roy and Weide (2023):

$$\log(y_i^1) = \alpha^1 + X_i^1\beta^1 + \varepsilon_i^1 \quad (3)$$

We then predict y on the source survey as $\hat{y}_i = \exp(\hat{\alpha} + X_i\hat{\beta})$. If we assume $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\log(Y) \sim \mathcal{N}(\mu, \sigma_y^2)$, it is possible to use the smearing retransformation $\exp(\hat{f}(X_i) + \frac{\hat{\sigma}^2}{2})$ to predict \hat{y}_{ic} (Duan 1983). We do not use this to follow the literature on imputation as closely as possible (see for example: Elbers *et al.* 2003; McKenzie 2005; Ferreira, Gignoux, and Aran 2011; Dang and Lanjouw 2023a; Sinha Roy and Weide 2023). An alternative to this is presented in the next step.

3. **Design Base Adjustment:** We categorise the population into clusters c based on subgroups of interest. We use regional clusters in this example, although this method can also be applied to other classifications.

We account for both within-cluster and between-cluster effects to evaluate our model. We do this by estimating two sources of bias in our predictions. For each cluster c , we measure the deviation of our estimates from the actual data using the cluster-specific ratio:

$$CS_{ratio}^c = \frac{\mu_c^1}{\hat{\mu}_c^1} \quad (4)$$

This metric quantifies our model’s average overestimation or underestimation of the true value for each cluster. Given that our prediction comes from a log-linear specification and is then rescaled, the cluster-specific ratio adjusts for the difference between the mean of the logarithm and the logarithm of the mean in a similar manner to the smearing re-transformation without the need to assume normally distributed errors within each cluster. For instance, if the predicted mean for a cluster is higher than the actual survey mean, the cluster-specific ratio will be less than one, indicating that the predictions for that cluster in the destination survey should be reduced.

Beyond matching the cluster means, we aim to match the shape of their specific distributions. Let $F_c(y)$ represent the observed cumulative density function of cluster c , and $F_c^{-1}(q)$ denote the corresponding quantile function. For each quantile q , we assess

the deviation between the observed and predicted values by calculating:

$$\frac{F_c^{-1}(q^1)}{\hat{F}_c^{-1}(q^1)} \quad (5)$$

To approximate this, we sort individuals into percentile ranks based on observed y and evaluate the accuracy of our predictions for each rank r within each cluster c using the cluster-rank ratio:

$$CR_{ratio}^{cr} = \frac{\mu_c^{r1}}{\hat{\mu}_c^{r1}} \quad (6)$$

This ratio indicates the average degree of over- or underestimation of the target variable by our model for each rank within each cluster. Since we approximate this by discretizing quantiles into percentile ranks, in the limit, we have that:

$$\lim_{r \rightarrow q} CR_{ratio}^{cr} = \frac{\mu_c^{r1}}{\hat{\mu}_c^{r1}} = \frac{y_c^1}{\hat{y}_c^1} = \frac{F_c^{-1}(q^1)}{\hat{F}_c^{-1}(q^1)} \quad (7)$$

We estimate the cluster and rank averages using the adjusted weights w^* from step 1 for both the observed μ and the predicted $\hat{\mu}$ values.

4. **Predicting over the destination:** We apply our model to predict y on the destination survey following Equation 2 as $\hat{y}_h^2 = \exp(\hat{\alpha}^1 + X_h^2 \hat{\beta}^1)$. Within each cluster, individuals are ranked according to their predicted outcomes (\hat{y}_{hcr}^2).
5. **Adjusting predictions:** We then adjust our predictions for the destination survey using the ratios calculated in the previous steps.

In the context of design-based Small Area Estimations, it is common to employ a composite estimate as $Y_n^{\hat{COM}} = \gamma \hat{Y}_n^{syn} + (1 - \gamma) \hat{Y}_n^{s-r}$, where the mean of area n is estimated as a linear combination of a synthetic (*syn*) estimator and a survey regression ($s - r$) (Pfeffermann 2013; Schaible 2014). The synthetic estimator refers to an estimation obtained from a linear regression of the area mean on the available mean of the covariates and tends to have large bias. The survey regression estimator is the probability-weighted estimator of the mean on the covariates using the Horvitz-Thompson estimators to reduce bias, but incorporating survey weights increases the

variance of the estimation. The convex combination of both serves as a compromise between large bias and large variance. γ is selected to minimise the MSE of the area means.

We adapt this methodology and estimate a composite prediction as a compromise between reducing bias (CS) in our prediction and incorporating variance (CR) to match our inequality measure as:

$$\tilde{y}_{hcr} = \gamma(CS_{ratio}^c * \hat{y}_{hcr}) + (1 - \gamma)(CR_{ratio}^{cr} * \hat{y}_{hcr}); \quad 0 < \gamma < 1 \quad (8)$$

where the adjustment ratios are specific to each rank within each cluster, by combining both ratios in a convex manner, we balance these effects. The cluster-specific ratio helps correct the mean values at the cluster level, while the within-cluster ratio fine-tunes the distribution within each cluster.

Using a cross-validation procedure, the algorithm selects γ to minimise the deviation of our preferred inequality or poverty measure from an out-of-bag sample of the source survey. Unlike conventional approaches that minimise Mean Squared Error (MSE), we prioritise minimising the deviation from our inequality measure to incorporate some variance in the estimation process, acknowledging that a degree of noise may improve the accuracy of the overall distribution.

To select the optimal γ , we divide our source survey into a training sample to estimate the linear regression and the ratios. Then, for different values of γ we estimate the deviation between the observed and predicted Gini over the remaining source survey that was not used to estimate the model and ratios. To prevent over-fitting, we opt for a γ , one standard deviation away from the value minimising the deviation of our inequality measure (Hastie *et al.* 2009; Chen and Yang 2021).

For this procedure to produce reliable predictions, we assume that the set of predictors shared across surveys is the same and measures the same concept. Second, we assume stability between surveys, meaning that $f(*)$ holds between the two surveys. We assume that X predicts y , but this prediction may have significantly less variance than the original

variable. We correct for systematic errors in the prediction using the ratios estimated with the source survey, which approximate the observed distribution.

To estimate confidence intervals for the Gini coefficient, we estimate R bootstrapped estimates of the Gini. Starting from step 2, we draw R bootstrapped samples from our source survey and estimate equation 3. We evaluate our model and calculate the bias ratios based on these samples R times. Finally, we predict over the entire target survey at each draw and adjust the predictions accordingly.

2 Data & Setting

To validate our methodology, we take advantage of the periodicity of the Mexican National Survey on Household Income and Spending (ENIGH),² which is conducted every two years by the National Institute for Statistics Geography and Information (INEGI). We impute income from the ENIGH 2016 into the ENIGH 2018 and vice versa. This setting allows us to examine the performance of our methodology in an environment where we have income information available and over which the model was not approximated to conduct several robustness tests.

Lastly, we apply our methodology to the ESRU Survey on Social Mobility in Mexico 2017 (ESRU-EMOVI), which does not contain any income information, and is conducted by the Espinosa Yglesias Research Centre (CEEY for its acronym in Spanish),³ and estimate inequality of opportunity comparing our approach to an asset-based index and the multiple imputation procedure.

2.1 Data

The ENIGH survey is representative of men and women at the national and state levels in urban and rural areas. It is conducted biennially and collects information about the household’s composition, income, and spending. The ESRU-EMOVI is representative of men

²Encuesta Nacional de Ingresos y Gastos del Hogar.

³ESRU-EMOVI.

and women at the national level and for five big regions of Mexico.⁴ It collects information about the current household and that of the respondent when she was 14. The recollection of childhood information allows for intergenerational analysis such as inequality of opportunity. However, the ESRU-EMOVI does not recollect income information.

Our analysis centres on household disposable income (HHI), which we assess at the per-capita level by dividing by household size. To account for inflation, we adjust HHI using the Consumer Price Index (CPI) provided by the [World Bank](#). Specifically, we rebase the CPI to a value of 1 in 2016. Summary statistics of HHI can be found in [Table 1](#) and the distribution of both measures in [Figure 1](#).

Table 1: Summary Statistics

	ENIGH 2016	ENIGH 2018	ESRU-EMOVI 2017
HHI	14,934.79 (23,349.11)	16,863.13 (22,756.53)	-
HHI*	14,934.79 (23,349.11)	15,159.54 (20,457.55)	-
log HHI*	9.19 (0.83)	9.22 (0.82)	-
Gini	0.492 (0.004)	0.484 (0.004)	
Deflator	1	1.11	1.06

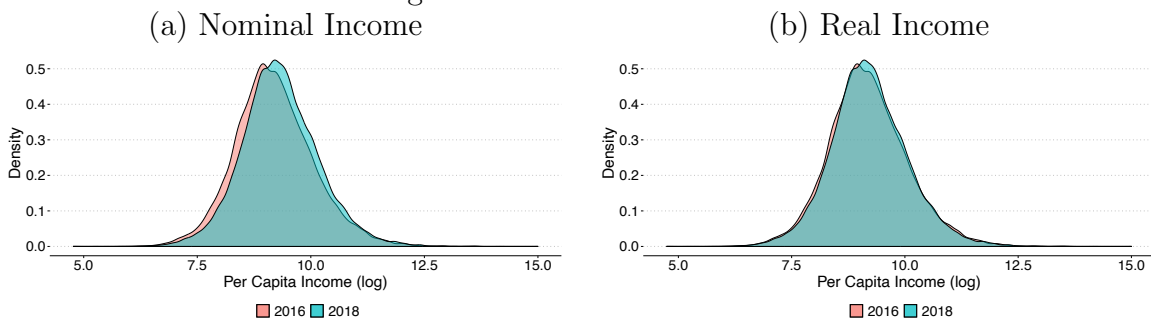
Notes: Means are estimated using the original survey weights. HHI represents per-capita disposable household income. * Denotes CPI adjusted to represent real terms. Standard deviations are presented in parentheses.

Mean income increased by around 13% in nominal terms between 2016 and 2018. In real terms, however, this amounts to an increase of less than 2%. Inequality, as measured through the Gini coefficient, is relatively stable between the two years, with both values being inside the 95% confidence interval of the other year at around 0.49.

Variables common in the ENIGH 2016, ENIGH 2018, and ESRU-EMOVI 2017 datasets are our predictors to ensure the validity of assumption *i*. These include household assets, composition, social security status, characteristics of the household head, and regional variables. For a detailed list of these variables, refer to [Table 2](#).

⁴The regions are divided into North (Baja California, Coahuila, Chihuahua, Monterrey, Sonora, Tamaulipas), North-West (Baja California Sur, Sinaloa, Zacatecas, Nayarit, Durango), Center-West (Aguascalientes, Colima, Jalisco, Michoacán, San Luis Potosí), South (Campeche, Chiapas, Guerrero, Oaxaca, Quintana Roo, Tabasco, Veracruz, Yucatán), Center (Guanajuato, Hidalgo, Mexico, Morelos, Puebla, Queretaro, Tlaxcala), and Mexico City.

Figure 1: Distribution of HHI



Notes: Density distribution of logged per-capita HHI using survey weights. Panel (a) shows the nominal values. Panel (b) shows real values. HHI is scaled using the World Banks CPI (2016 = 1).

The descriptive statistics of the predictors in the three samples suggest that they are representative of the same population. Some variables, such as the percentage of households where the floor material is other than cement or dirt, differ between ENIGH and ESRU-EMOVI. To consider this, we correct the sample weights as explained in step 1 of the methodology. Furthermore, to ensure that the estimated means accurately reflect the target population, we align our cluster selection with the state level, as the ENIGH is representative at this level. Since the ESRU-EMOVI is only representative at the regional level, we adjust the clusters accordingly as a robustness check and in our final imputation exercise.

Setting

We test to see if assumption *ii* holds by accounting for different scenarios as done by Newhouse *et al.* (2014). We impute forwards from 2016 to 2018 and backwards from 2018 to 2016. In each scenario, we approximate Equation 3 using a bootstrapped sample of the source survey. We estimate our ratios over 80% of this source survey sample and select γ using the remaining 20%. Once this is done, we test how well our methodology is doing by looking at the imputation in the full target survey.

We set $R = 100$ and estimate 100 bootstrapped Gini coefficients. From these, we estimate a mean expected value for the Gini coefficient as well as 95% confidence intervals. Finally, we average our predictions over the bootstrapped samples for a final imputation.

Table 2: Variables Used as Predictors (Household Level)

	ENIGH 2016	ENIGH 2018	ESRU-EMOVI 2017
Telephone	0.37 (0.46)	0.35 (0.45)	0.38 (0.49)
Cellphone	0.86 (0.35)	0.89 (0.32)	0.85 (0.36)
TV	0.48 (0.50)	0.43 (0.50)	0.50 (0.50)
Internet Connection	0.38 (0.46)	0.42 (0.48)	0.41 (0.49)
Water	0.74 (0.45)	0.75 (0.45)	0.77 (0.41)
Electricity	0.99 (0.12)	0.99 (0.13)	0.91 (0.27)
Gender (HH)	0.73 (0.43)	0.72 (0.44)	0.52 (0.49)
Car	0.45 (0.50)	0.45 (0.50)	0.58 (0.77)
Owns Property	0.69 (0.45)	0.68 (0.45)	0.58 (0.49)
Dirt Floor	0.03 (0.17)	0.03 (0.17)	0.03 (0.16)
Cement Floor	0.51 (0.50)	0.51 (0.50)	0.58 (0.50)
Other Floors	0.46 (0.49)	0.46 (0.49)	0.39 (0.49)
Share Men	0.49 (0.23)	0.49 (0.23)	0.50 (0.23)
Share Occupied	0.52 (0.28)	0.53 (0.28)	0.44 (0.28)
IMSS	0.39 (0.49)	0.39 (0.49)	0.38 (0.49)
IMSS Prospera	0.01 (0.08)	0.00 (0.05)	0.01 (0.11)
ISSSTE	0.02 (0.11)	0.02 (0.11)	0.05 (0.24)
Other (Social Security)	0.58 (0.49)	0.58 (0.49)	0.29 (0.44)
PEMEX (Social Security)	0.01 (0.09)	0.01 (0.09)	0.01 (0.08)
Prospera	0.18 (0.41)	0.18 (0.41)	0.13 (0.33)
Elderly	0.07 (0.27)	0.07 (0.26)	0.06 (0.24)
Big Town	0.51 (0.49)	0.49 (0.48)	0.09 (0.28)
Median Town	0.14 (0.34)	0.15 (0.33)	0.18 (0.38)
Small/Median Town	0.14 (0.34)	0.14 (0.34)	0.21 (0.40)
Small Town	0.21 (0.48)	0.23 (0.48)	0.52 (0.50)
No Education (HH)	0.06 (0.24)	0.05 (0.24)	0.03 (0.19)
Kindergarden (HH)	0.15 (0.37)	0.13 (0.36)	0.00 (0.04)
Primary School (HH)	0.21 (0.41)	0.20 (0.41)	0.24 (0.43)
Secondary School (HH)	0.30 (0.46)	0.30 (0.46)	0.31 (0.46)
High School (HH)	0.15 (0.34)	0.16 (0.35)	0.21 (0.40)
Graduate Degree (HH)	0.11 (0.29)	0.12 (0.30)	0.16 (0.35)
Postgraduate Degree (HH)	0.03 (0.13)	0.03 (0.13)	0.01 (0.11)
State of Residency	X	X	X

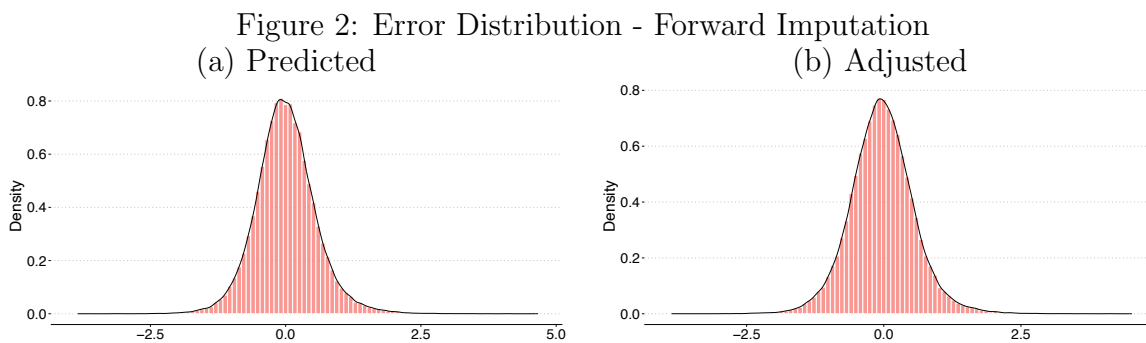
Notes: Values correspond to the weighted mean in each survey. Standard deviations are presented in parentheses. HH stands for the household head.

3 Validation

We validate our methodology by applying it to the ENIGH 2016 and 2018. We first present results for the forward imputation, estimating a model over the ENIGH 2016 and imputing on the ENIGH 2018. We then present the results for the backward imputation. All results in this section are estimated using a linear regression following the methodologies of Elbers *et al.* (2003), McKenzie (2005), and Dang and Lanjouw (2023a). Future work remains to optimise the procedure to incorporate predictions based on machine learning algorithms.

3.1 Forward Imputation

In the forward imputation setting, we employ the ENIGH 2016 as our source survey and the ENIGH 2018 as the target survey. Initially, when applied to the source survey, our model yields an out-of-sample R^2 of 0.53. The distribution of our error terms are shown in Figure 2. Panel (a) shows the original error term on the source survey, while Panel (b) shows the distribution once the adjustment in step 5 is applied to the source survey.



Notes: Density distribution of error terms. Panel (a) shows the original error distribution. Panel (b) shows the error distribution after applying the adjustment.

The ENIGH 2018 reveals a Gini coefficient of 0.48. Our primary objective is to impute household income (HHI) to mirror this inequality estimate. The outcomes of our imputed income measure are presented in Table 3. We present results for one imputation of the bootstrap replications to showcase what each adjustment is doing.

As anticipated, our predicted Gini coefficient falls more than 12 basis points below the observed value for 2018. While adjusting solely for between-cluster bias does reduce our

Table 3: Results: Forward Imputation

	Gini	RMSE
Observed	0.484	
Predicted	0.362	20,326
Adjusted (CS)	0.365	19,737
Adjusted (CR)	0.566	28,176
Adjusted	0.493	23,375

Notes: The Gini is computed using survey weights. RMSE is presented in real Mexican pesos ($y - \tilde{y}$).

estimation error, it does so marginally. The resultant Gini coefficient remains notably lower than the expected one, still trailing by around 12 points. Conversely, adjusting exclusively for within-cluster bias exacerbates our prediction’s error, resulting in an overestimation of the Gini coefficient by 8 points. The composite estimate, with a larger error than the between-cluster adjusted values, exhibits a smaller deviation than the within-cluster adjustment. Correspondingly, the Gini coefficient closely aligns, differing by one point from the observed value.

We use a bootstrap procedure to estimate confidence intervals and assess the variability of our methodology. Table 3 presents these results. It is important to note that the values in Table 3 and Table 4 differ because the first table displays results from a single bootstrap iteration, while the second table reports the mean and standard errors derived from the full bootstrap process.

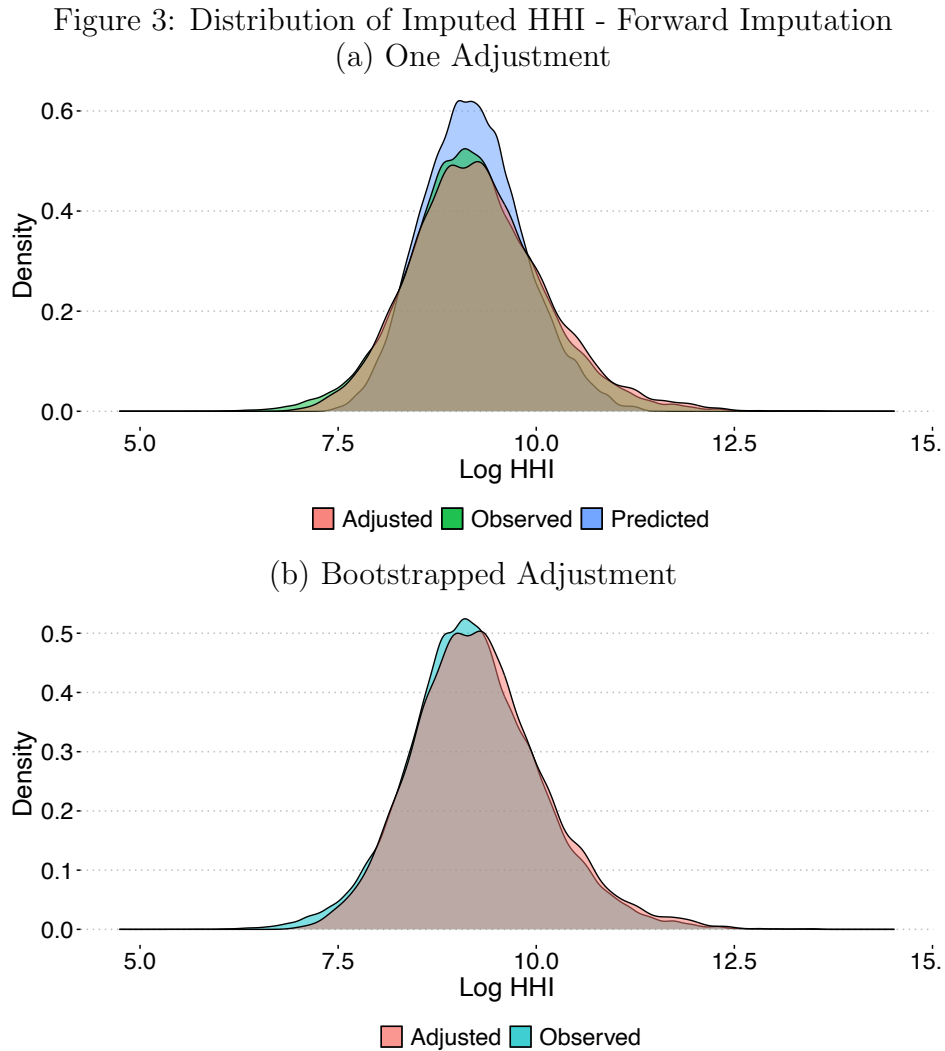
Table 4: Confidence Intervals: Forward Imputation

	Gini
Observed	0.484 (0.475, 0.493)
Adjusted	0.491 (0.471, 0.511)

Notes: Observed values are computed using survey weights following Horvitz and Thompson (1952). We estimate confidence intervals by bootstrapping from step 2 in our methodology for the adjusted value.

Despite our expected Gini coefficient surpassing the observed value for 2018, it remains within the confidence intervals of the observed Gini. Furthermore, our estimated Gini coefficient exhibits a marginally higher variability than the observed counterpart. Nevertheless, the confidence interval remains within a 5% error margin of the observed Gini coefficient.

Figure 3 shows the density estimations of our imputed value.



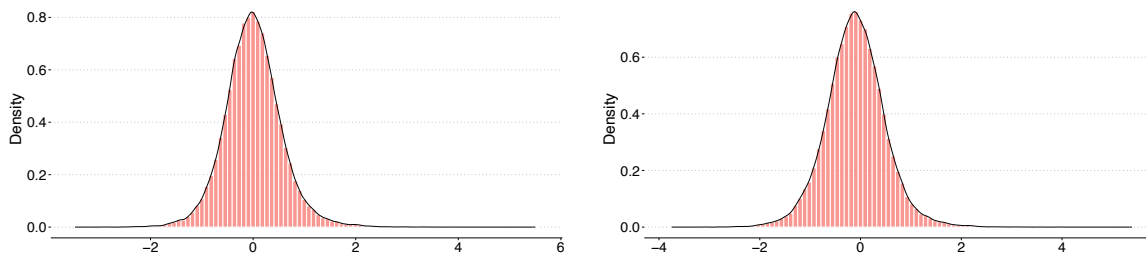
Notes: Density distribution of logged HHI using survey weights. Panel (a) shows the distribution when imputing using the whole survey. Panel (b) shows the results of the bootstrapped imputation.

3.2 Backward Imputation

For the backward imputation, we employ the ENIGH 2018 as our source survey and the ENIGH 2016 as the target survey. When applied to the source survey, our model yields an out-of-sample R^2 of 0.53. Figure 4 shows the distribution of our error terms.

The ENIGH 2016 shows a Gini coefficient of 0.49, slightly higher than the one observed in 2018. The results of our imputation process are shown in Table 5.

Figure 4: Error Distribution - Backward Imputation
(a) Predicted (b) Adjusted



Notes: Density distribution of error terms. Panel (a) shows the original error distribution. Panel (b) shows the error distribution after applying the adjustment.

Table 5: Results: Backward Imputation

	Gini	RMSE
Observed	0.492	
Predicted	0.353	21,199
Adjusted (CS)	0.362	20,978
Adjusted (CR)	0.552	28,726
Adjusted	0.494	24,957

Notes: The Gini is computed using survey weights. RMSE is presented in real Mexican pesos (\tilde{y}).

Similar to the forward imputation, our original prediction of the Gini is more than 13 points below the observed one for 2016. While adjusting for between-cluster bias reduces the estimation error, the Gini coefficient still remains more than 12 points below the observed value. On the other hand, adjusting for within-cluster bias exacerbates the error in our prediction, leading to an overestimation of the Gini coefficient by more than 6 points. Despite the composite estimate exhibiting a larger RMSE than the between-cluster adjusted values, it shows a smaller deviation than the within-cluster adjustment. Consequently, the Gini coefficient closely aligns with the observed value.

We present confidence intervals estimated through the bootstrap procedure to estimate the variability of our methodology. Table 6 shows the results.

In this setting, our imputation yields the observed value of the Gini on average. As in the forward imputation, our estimated Gini coefficient exhibits higher variability than the observed counterpart. In this case, the confidence intervals are higher than the ones observed in the forward imputation. Figure 5 shows the density estimations of our imputed value.

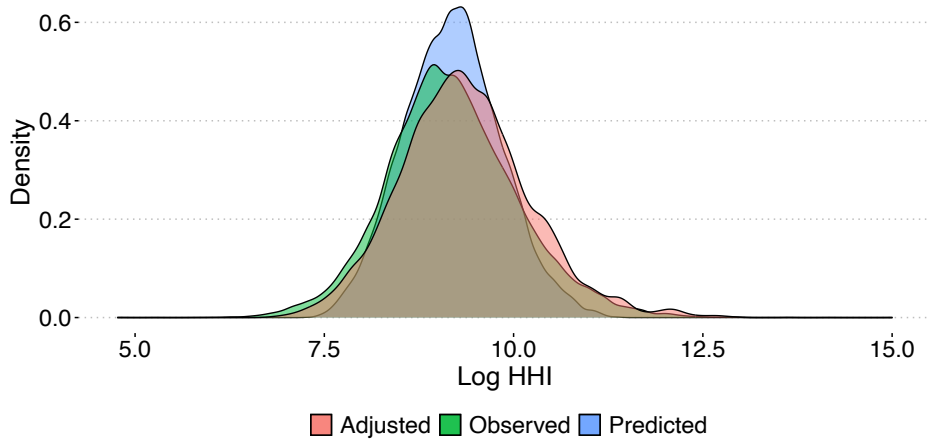
Table 6: Confidence Intervals: Backward Imputation

	Gini
Observed	0.492 (0.485, 0.500)
Adjusted	0.492 (0.468, 0.514)

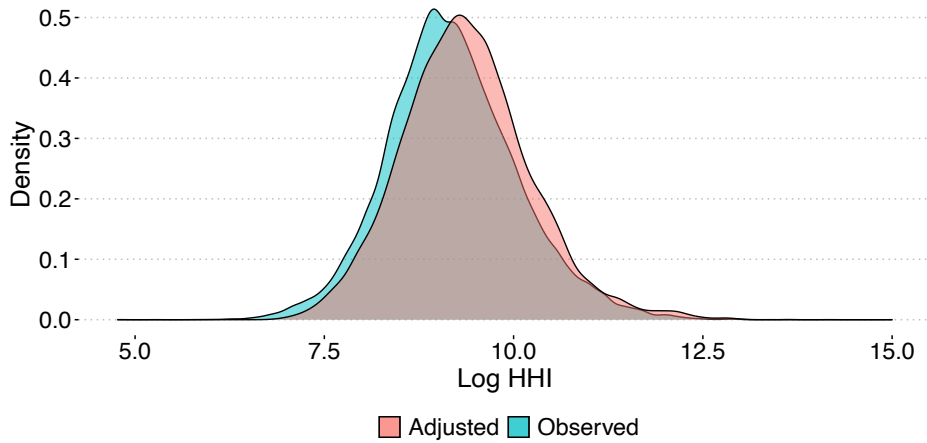
Notes: Observed values are computed using survey weights following Horvitz and Thompson (1952). We estimate confidence intervals by bootstrapping from step 2 in our methodology for the adjusted value.

Figure 5: Distribution of Imputed HHI - Forward Imputation

(a) One Adjustment



(b) Bootstrapped Adjustment



Notes: Density distribution of logged HHI using survey weights. Panel (a) shows the distribution when imputing using the whole survey. Panel (b) shows the results of the bootstrapped imputation.

4 Robustness Checks

We consider various scenarios to assess the robustness of our imputation methodology. We recognise the similarity in the sampling designs of both ENIGH surveys. To address this, we subset the ENIGH and compute confidence intervals across different sample sizes and changing the cluster level.

A crucial aspect of imputation methods is maintaining consistency in income and consumption profiles throughout the process. This means maintaining the relative position of households in the conditional distribution. Our approach involves introducing variance to the imputation by leveraging predicted individual positions at the cluster level.

To address this concern, we undertake two exercises. First, we analyse profiles by segmenting the population into quantile groups and comparing the distribution of covariates across observed and predicted quantile groups. This analysis offers insights into the alignment between the original and imputed profiles.

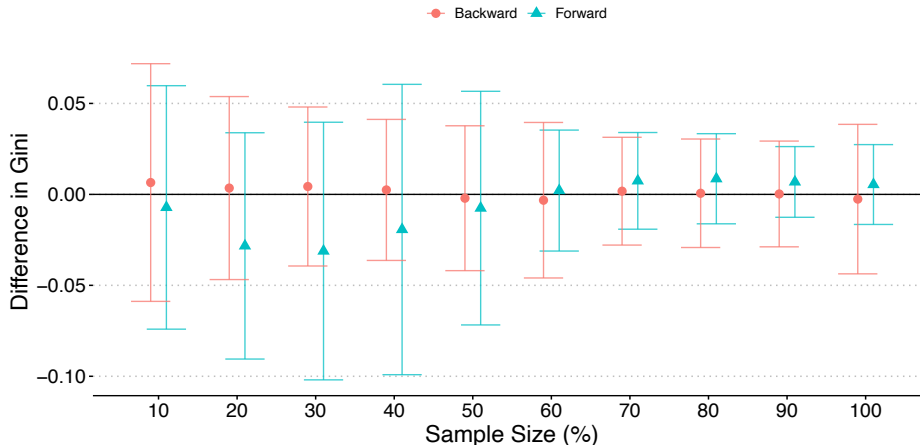
Additionally, we conduct a comparative analysis of the regression coefficients derived from the two ENIGH surveys, particularly emphasising their direction and magnitudes. This comparative analysis further validates the stability assumption and reinforces the reliability of our imputation methodology for assessing IOp.

Sample Size

To account for differences in the sample of the target survey, we first subset the target survey and estimate the difference between the predicted and observed Gini coefficients within this subset. We perform 100 bootstrap iterations to ensure robustness at varying percentages of the target survey. This involves increments of 10 per cent, ranging from 10 to 100 percent coverage. Figure 6 shows the results.

Our results show a convergence trend. The average point estimate of the Gini coefficient tends to align more closely with the observed values for the backward imputation method. However, as the sample size approaches 100%, both imputations converge towards the observed values. This convergence is accompanied by a reduction in the confidence intervals

Figure 6: Difference in Ginis - Sample Size



Notes: Points show the expected difference between the observed and the estimated Gini coefficients for different sample sizes of the target survey. Confidence intervals are derived using bootstrapped replications from step 1 in our methodology.

and a tightening of the point estimates, reflecting increased precision.

Additionally, the forward imputation exhibits lower levels of volatility, particularly evident when the sample size nears 100%. This suggests higher stability in the forward imputation method under such conditions.

Cluster Level

We modify our procedure to select clusters at the regional level as a further robustness check to see how our methodology performs when changing the clusters. The results of this adjustment are presented in Table 7.

Table 7: Regional Cluster Level

	Forward	Backward
Observed	0.484 (0.475, 0.493)	0.492 (0.485, 0.500)
Adjusted	0.491 (0.473, 0.507)	0.486 (0.444, 0.527)

Notes: Values are computed using survey weights following Horvitz and Thompson (1952). We estimate confidence intervals by bootstrapping from step 2 in our methodology for the adjusted value.

The outcomes derived from the regional cluster-level analysis are consistent with our earlier findings. Both forward and backward imputation methods exhibit slightly higher

variance in the imputed Gini coefficients than the observed values.

In the case of forward imputation, the average Gini coefficient remains identical to that obtained when clustering was conducted at the state level. Conversely, with the backward imputation, we observe a Gini coefficient lower than the observed and previously predicted values but still within the observed confidence intervals. Additionally, the confidence intervals for the backward imputation are wider than those of the forward imputation, aligning with our previous results.

Profile Compositions

A critical aspect of imputation in assessing income and consumption regarding external covariates is preserving individuals' relative positions based on their socioeconomic characteristics. When imputing data from ENIGH to ENIGH, it is challenging to ascertain whether the original ranking according to external covariates is maintained since these have the same set of covariates. To address this, we select predictors closely linked to socioeconomic characteristics and evaluate whether our imputation yields reliable estimates.

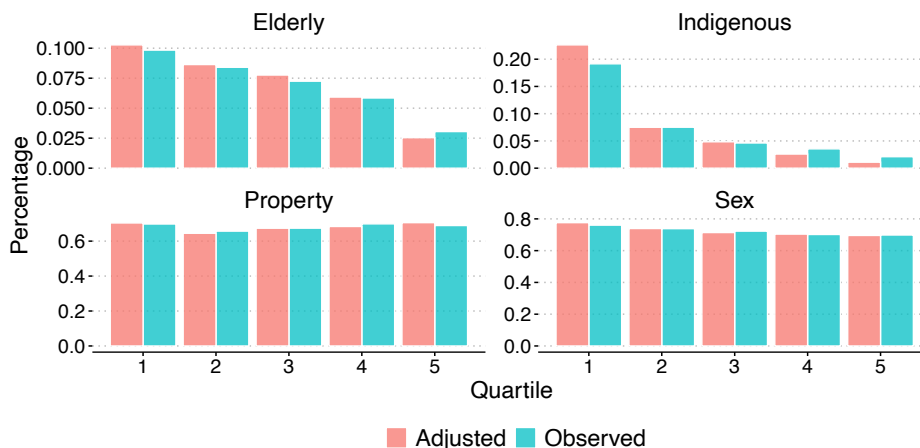
In particular, we include variables such as the presence of an elderly person in the household, home-ownership status, and the gender of the household head from our pool of predictors. Additionally, we incorporate the ability of at least one household member to speak an indigenous language as an external socio-demographic characteristic, which was not used during the imputation process.

We first divide the population into observed and predicted income quartiles and compare the share of subgroups across the two to analyse if individuals' relative is preserved. Figure 7 shows the results.

We observe consistent patterns across imputation profiles. Specifically, when examining socioeconomic characteristics such as the ability to speak an indigenous language and the presence of elderly individuals in the household, there is a clear downward trend across income quartiles. As we move from lower to higher income quartiles, the prevalence of these characteristics gradually decreases.

Beyond examining quartile composition, it is crucial that the imputation preserves individuals' ranks within the income distribution. To further validate our approach, we analyse

Figure 7: Household Income Profiles



Notes: Observed percentages are estimated using survey weights. Adjusted percentages are estimated from bootstrapped samples. We bootstrap from step 2 in our methodology.

the correlation between our imputed estimates and the observed values and compare it with the correlation from a multiple imputation procedure incorporating an added error term. This comparison allows us to assess whether our imputation method preserves the original distribution’s structure more effectively.

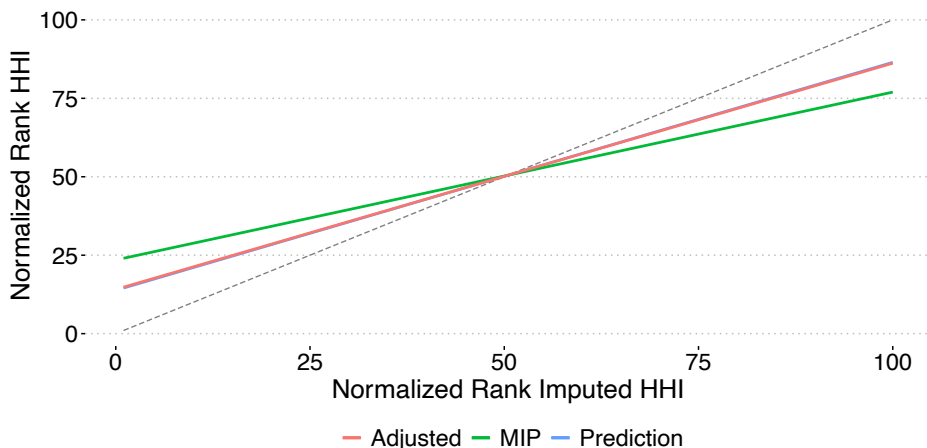
Figure 8 presents the estimates from regressing observed (log) income values against the imputed ones and the results from regressing normalised ranks. Our findings indicate a Spearman-rank correlation of 0.72 for our adjusted imputation, compared to 0.52 for the multiple imputation method. These results suggest that adding the error term in the multiple imputation process disrupts individuals’ positions within the conditional income distribution.

Stability

Finally, we check for the stability of our estimations by comparing the coefficients derived from both surveys. We aim to discern whether both surveys exhibit similar patterns in terms of magnitude and direction. The results of this analysis are depicted in Figure 9.

We observe consistency across most of our predictors, with coefficients displaying significant overlap. Even in cases where there is a lack of overlap, we find that the direction and magnitude of the coefficients remain similar between the two surveys.

Figure 8: Household Income Correlation



Notes: Results for the forward imputation, data points correspond to the ENIGH 2018. Panel (a) shows the distribution of logged values across specifications. Panel (b) shows the results of a rank-rank regression of imputed and observed household income. The dotted line represents the 45-degree line.

5 IOp in Mexico: Assets vs. Income

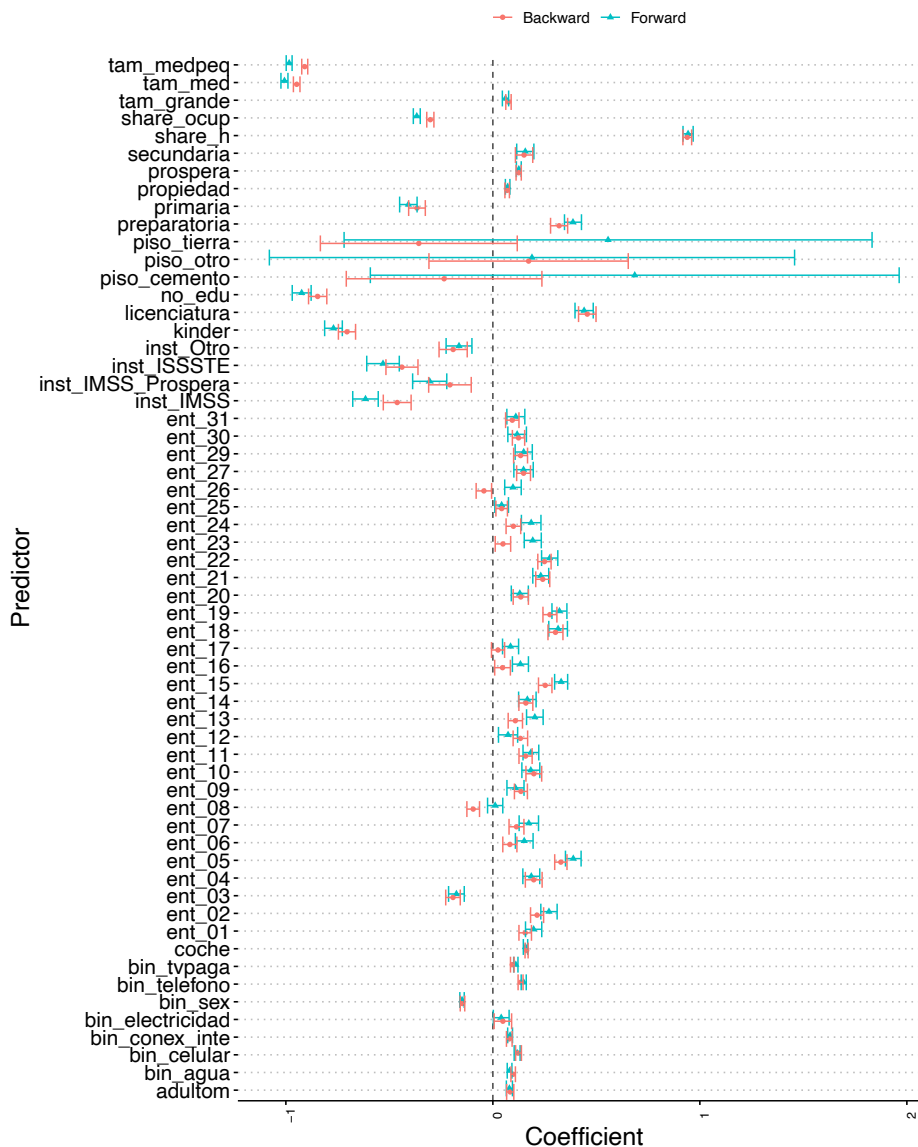
Finally, we apply our methodology to the ESRU-EMOVI 2017 and estimate inequality of opportunity (IOp). We compare our results with those using the asset index approach (Filmer and Pritchett 2001; Filmer and Scott 2012) and the multiple imputation approach (Ferreira, Gignoux, and Aran 2011; Dang and Lanjouw 2023b).

Data

The ESRU-EMOVI collects information on the current household and the household of the respondent at the age of 14. Recalling information from childhood allows for intergenerational analysis. Respondents are asked about their parents' education and occupation and about the availability of assets in both households.

In the absence of direct income data, researchers often resort to approximating economic well-being using an asset index, as shown in previous studies (e.g. Ferreira, Gignoux, and Aran 2011; Torche 2015; Vélez-Grajales *et al.* 2019; Delajara *et al.* 2022; Plassot *et al.* 2022). Similarly, through factor analysis, we use this approach to construct an asset index for the respondent's current household. Specifically, we use Principal Component Analysis

Figure 9: Stability



Notes: Points show each predictor's average point estimate of the OLS. Confidence intervals are estimated using bootstrap from step 2 of our methodology.

(PCA) to derive socioeconomic well-being as

$$y_h = \sum_{f=1}^F a_f \left(\frac{x_{fh} - \bar{x}_f}{s_f} \right) \quad (9)$$

where the F -dimensional vector of weights a is chosen to maximise the sample variance of y , subject to the constraint $\sum_f a_f^2 = 1$. \bar{x}_f is the mean of the f th asset and s_f is its standard deviation. The details of the PCA can be found in Table 8, Figure 10 shows its distribution.

Table 8: Household Asset Index

	Mean	Standard Deviation	Weight (a_f)
Plumbing	0.921	0.269	0.159
Stove	0.940	0.237	0.175
Electricity	0.987	0.111	0.055
TV	0.849	0.358	0.188
Fridge	0.925	0.263	0.196
Washing Machine	0.793	0.405	0.250
Landline	0.383	0.486	0.296
DVD Bluera	0.418	0.493	0.274
Microwave	0.521	0.500	0.307
Cable TV	0.523	0.500	0.264
Internet	0.428	0.495	0.340
Cellphone	0.178	0.383	0.261
Computer	0.324	0.468	0.318
Other Housing	0.047	0.212	0.124
Other Land	0.028	0.164	0.066
Automobile	0.097	0.296	0.202
Bank account	0.223	0.416	0.231
Credit Card	0.152	0.359	0.241
Premises	0.038	0.192	0.094
Working Parcels	0.074	0.262	0.099
Working Machinery	0.014	0.119	0.020
Working Animals	0.035	0.183	-0.010
Livestock	0.039	0.193	-0.031

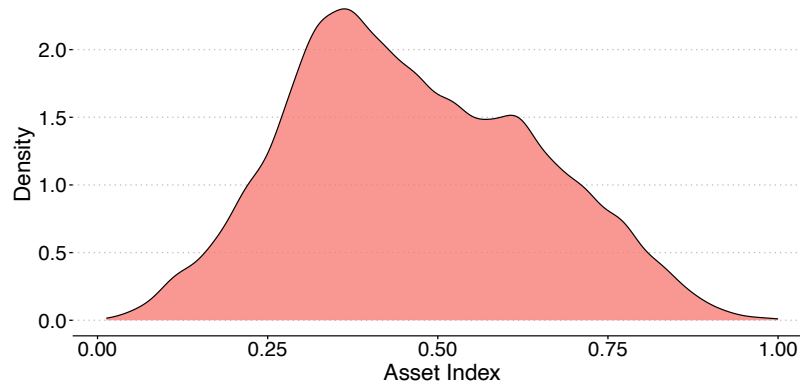
Notes: The table shows the characteristics of the assets used to construct the asset index as in Equation 9. The weights are derived using Principal Component Analysis.

Another common approach is the multiple imputation procedure (MIP) (McKenzie 2005; Ferreira, Gignoux, and Aran 2011). In this method, we estimate Equation 3 from a bootstrap sample drawn from our source survey. To account for variance, we predict household income (HHI) for each household in the target survey as:

$$\hat{y}_h = \exp(\hat{\alpha}_B + X_h * \hat{\beta}_B + \tilde{\varepsilon}_{Bh}) \quad (10)$$

where $\hat{\alpha}_B$ and $\hat{\beta}_B$ are specific to each bootstrap sample B , with $\tilde{\varepsilon}_{Bh}$ representing an additional error term randomly drawn from the empirical error distribution of the model. This procedure is repeated R times, and the measure of IOp is taken as the average of the estimated

Figure 10: Distribution of Asset Index



Notes: Density distribution of the normalised asset index using survey weights. The index is derived using PCA.

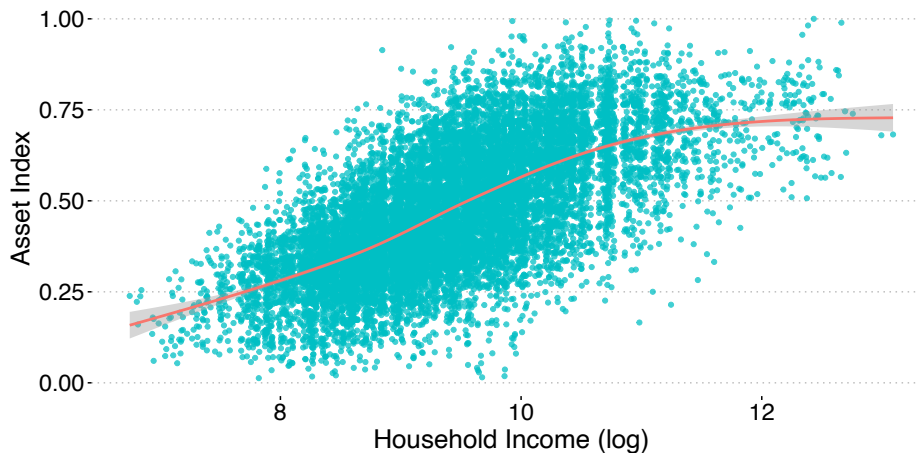
IOP values.

To assess the effectiveness of our method, we compare it with the above approaches, using 100 bootstrap replicates to estimate confidence intervals. In contrast to multiple imputations, a notable advantage of our procedure is the ability to obtain an average imputation. This is illustrated in Panel (b) of Figures 3 and 5, facilitating a direct comparison between the imputed measure and our asset index. Figure 11 provides insights into the relationship between our asset index and the imputed measure of income, along with the distribution of the average imputation derived from our methodology.

We observe a correlation of 0.45 between our imputed measure and the asset index, indicating a positive association between the two. The association is weaker at the top and bottom of the distribution, consistent with the notion that such asset indices cannot distinguish the very poor from the very rich (Ferreira, Gignoux, and Aran 2011).

The positive correlation between the two measures suggests that although they are related, each measure captures a different aspect of economic well-being. These findings are consistent with those of Filmer and Scott (2012) that the asset index is more closely related to long-term consumption than household income.

Figure 11: Asset Index and Imputation
(a) Correlation



(b) Density of HHI



Notes: Distribution of our average imputed measure. Panel (a) shows its relationship to the asset-based index. The red line is the association derived from a LOWESS regression. Panel (b) shows the distribution of imputed HHI using survey weights.

Conceptual Framework and Estimation Strategy

The literature on IOp follows Roemer (1998) and defines outcome y of individual i to be expressed by an additively separable function of effort e and circumstances C specific to each individual

$$y_i = f(C_i, e_i) \quad (11)$$

the population can then be divided into k non-overlapping groups based on their circumstances and m tranches of effort. We assume equality of opportunity (EOp) if there is no

difference in the expected outcome between groups (ex-ante) or within tranches (ex-post). When considering a weak ex-ante criterion, we look at the mean of each specific type:

$$\hat{y}_i = \mu_j \quad \forall j \in [1, \dots, k] \quad (12)$$

Using the parametric approach, we follow Ferreira and Gignoux (2011) and estimate Equation 11 through a linear regression. We use gender, ethnicity, region of birth, father’s and mother’s education and parental occupation as circumstances:

$$y_i = \alpha + C_i\beta \quad (13)$$

To get a measure of IOp, we first predict our outcome to approximate Equation 12 and then estimate the proportion of total inequality explained by circumstances as:

$$IOp = \frac{Gini(\hat{y})}{Gini(y)} \quad (14)$$

The circumstances chosen for this purpose are not exhaustive and will yield a lower bound estimation of ex-ante IOp.

Results

We analyze the differences in Inequality of Opportunity (IOp) across the three measures: Asset Index, Predicted, Multiple Imputations, and Adjusted. Table 9 presents a summary of our findings.

Table 9: Inequality Of Opportunity

	Asset Index	Prediction	Multiple Imputations	Adjustment
Total Inequality	0.220 (0.216, 0.224)	0.368 (0.363, 0.372)	0.477 (0.461, 0.493)	0.511 (0.475, 0.547)
Absolute IOp	0.107 (0.104, 0.110)	0.209 (0.204, 0.214)	0.209 (0.198, 0.221)	0.296 (0.274, 0.318)
Relative IOp	0.486 (0.473, 0.499)	0.569 (0.559, 0.579)	0.439 (0.414, 0.464)	0.580 (0.566, 0.593)

Notes: Point estimates are computed using survey weights. Confidence intervals are estimated through bootstrap estimations. We bootstrap our source survey from step 2 in our methodology for the prediction, multiple imputation, and adjusted measure. For the asset index, we only bootstrap the sample from which we estimate IOp.

Total inequality varies significantly across the different measures. The Asset Index pro-

duces a Gini coefficient of 0.22, which is notably low for income measures.⁵ The Predicted measure, which does not include additional variance, shows a Gini coefficient of 0.368, an increase of more than 14 points compared to the Asset Index but still lower than the observed values in both ENIGH 2016 and 2018. The Multiple Imputation method yields a Gini coefficient of 0.477, further increasing the inequality measure by more than 10 points compared to the Predicted method. Lastly, the Adjusted imputation method results in the highest Gini coefficient of 0.511, a substantial rise of 15 points.

It is crucial to recognize that both the Multiple Imputation and Adjusted methods originate from the same initial forecast (column two of Table 9). Therefore, the observed differences in overall inequality and IOp are attributable to the distinct underlying assumptions in each imputation methodology.

A closer examination of Relative IOp reveals significant differences between the measures. The Asset Index and Predicted methods yield Relative IOp values of 0.486 and 0.569, respectively, indicating a moderate level of inequality of opportunity relative to total inequality. However, the Multiple Imputation method produces a lower Relative IOp of 0.439, suggesting that introducing random error during the imputation process reduces the extent to which inequality can be attributed to opportunity-related factors.

In contrast, the Adjusted method results in a Relative IOp of 0.580. This difference underscores the impact of imputation techniques on measuring inequality of opportunity. The Adjusted method's ability to produce a higher Relative IOp suggests that it may more accurately capture the relationship between socioeconomic circumstances and income, minimising the attenuation effect caused by random error in the Multiple Imputation approach.

Furthermore, the narrower confidence intervals observed in the Adjusted method for both Absolute and Relative IOp indicate more stability and precision in these estimates compared to the Multiple Imputation method. This further supports the Adjusted method as a more reliable approach for measuring IOp, particularly when the goal is to preserve individuals' rankings within the income distribution.

⁵The [World Bank](#) estimates a value of 0.232 for the Slovak Republic as the lowest Gini index.

6 Conclusions

The lack of data on income and consumption in surveys poses significant challenges for distributional analysis. Researchers typically employ two solutions to address this issue: substituting these variables or reconstructing them within surveys.

The first approach often involves the use of asset-based indices. Our findings indicate that these indices are inadequate for distinguishing the very rich from the poor, consistent with the observations of Filmer and Scott (2012) and McKenzie (2005). This inadequacy results in lower variance and, consequently, lower levels of reported inequality compared to income-based measures.

Alternatively, regression-based imputations are commonly used, with the multiple imputation procedure being particularly prevalent (McKenzie 2005; Ferreira, Gignoux, and Aran 2011; Dang and Lanjouw 2023b). These regression models typically exhibit reduced distribution variance due to the error term's exclusion. A random error term from the empirical distribution of errors of the model is usually included in the prediction to account for this variance.

Our findings indicate that while this method attempts to account for variance, it introduces a downward bias when the analysis involves relationships with variables not included in the prediction model. This bias stems from the correlation between the added error term and these external variables, which can distort individuals' positions within the income distribution. We illustrate this effect through our Inequality of Opportunity (IOP) analysis, a relevant case study since circumstances are typically not included in surveys containing income information. However, this bias could also affect analyses of poverty and its dynamics.

To overcome this issue, we propose an alternative methodology that avoids incorporating the error term. Instead, we adjust the predicted values by aligning them with the observed values in the survey used for model estimation. This approach involves calculating the deviations between our predictions and the original data and adjusting the predictions to reflect the empirical distribution more accurately.

References

- Björklund, A. and M. Jäntti (1997). “Intergenerational Income Mobility in Sweden Compared to the United States”. In: *The American Economic Review* 87.5, pp. 1009–1018. ISSN: 00028282. URL: <http://www.jstor.org/stable/2951338>.
- Bloise, F., P. Brunori, and P. Piraino (2021). “Estimating intergenerational income mobility on sub-optimal data: a machine learning approach”. In: *Journal of Economic Inequality* 19.4, pp. 643–665. ISSN: 15738701. DOI: [10.1007/s10888-021-09495-6](https://doi.org/10.1007/s10888-021-09495-6).
- Campion, W. M. and D. B. Rubin (1989). “Multiple Imputation for Nonresponse in Surveys”. In: *Journal of Marketing Research*. Wiley Series in Probability and Statistics 26.4, p. 485. ISSN: 00222437. DOI: [10.2307/3172772](https://doi.org/10.2307/3172772). URL: <https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316696>.
- Chen, Y. and Y. Yang (2021). “The One Standard Error Rule for Model Selection: Does It Work?” In: *Stats* 4.4, pp. 868–892. ISSN: 2571-905X. DOI: [10.3390/stats4040051](https://doi.org/10.3390/stats4040051). URL: <https://www.mdpi.com/2571-905X/4/4/51>.
- Corral, P., I. Molina, A. Cojocar, and S. Segovia (2022). “Guidelines to Small Area Estimation for Poverty Mapping”. In: *Guidelines to Small Area Estimation for Poverty Mapping*. DOI: [10.1596/37728](https://doi.org/10.1596/37728).
- Cowell, F., E. Karagiannaki, and A. Mcknight (2018). “Accounting for Cross-Country Differences in Wealth Inequality”. In: *Review of Income and Wealth* 64.2, pp. 332–356. DOI: <https://doi.org/10.1111/roiw.12278>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/roiw.12278>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12278>.
- Dang, H.-A. H. (2021). “To impute or not to impute, and how? A review of poverty-estimation methods in the absence of consumption data”. In: *Development Policy Review* 39, pp. 1008–1030. DOI: <https://doi.org/10.1111/dpr.12495>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/dpr.12495>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/dpr.12495>.
- Dang, H.-A. H. and P. F. Lanjouw (2023a). “Measuring Poverty Dynamics with Synthetic Panels Based on Repeated Cross Sections”. In: *Oxford Bulletin of Economics and Statis-*

- tics* 85.3, pp. 599–622. DOI: <https://doi.org/10.1111/obes.12539>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/obes.12539>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12539>.
- Dang, H.-A. H. and P. F. Lanjouw (2023b). “Regression-based imputation for poverty measurement in data-scarce settings”. In: *Research Handbook on Measuring Poverty and Deprivation*. Edward Elgar Publishing Ltd., pp. 141–150. ISBN: 9781800883451. DOI: [10.4337/9781800883451.00023](https://doi.org/10.4337/9781800883451.00023).
- Delajara, M., R. M. Campos-Vazquez, and R. Velez-Grajales (2022). “The regional geography of social mobility in Mexico”. In: *Regional Studies* 56.5, pp. 839–852. ISSN: 13600591. DOI: [10.1080/00343404.2021.1967310](https://doi.org/10.1080/00343404.2021.1967310). URL: <https://www.tandfonline.com/doi/abs/10.1080/00343404.2021.1967310>.
- DiNardo, J., N. M. Fortin, and T. Lemieux (1996). “Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach”. In: *Econometrica* 64.5, pp. 1001–1044. ISSN: 00129682, 14680262. URL: <http://www.jstor.org/stable/2171954> (visited on 03/13/2024).
- Downes, M., L. C. Gurrin, D. R. English, J. Pirkis, D. Currier, M. J. Spittal, and J. B. Carlin (2018). “Multilevel Regression and Poststratification: A Modeling Approach to Estimating Population Quantities From Highly Selected Survey Samples”. In: *American Journal of Epidemiology* 187.8, pp. 1780–1790. ISSN: 0002-9262. DOI: [10.1093/aje/kwy070](https://doi.org/10.1093/aje/kwy070). eprint: <https://academic.oup.com/aje/article-pdf/187/8/1780/25369165/kwy070.pdf>. URL: <https://doi.org/10.1093/aje/kwy070>.
- Duan, N. (1983). “Smearing Estimate: A Nonparametric Retransformation Method”. In: *Journal of the American Statistical Association* 78.383, pp. 605–610. ISSN: 01621459. URL: <http://www.jstor.org/stable/2288126> (visited on 03/12/2024).
- Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). “Micro-Level Estimation of Poverty and Inequality”. In: *Econometrica* 71.1, pp. 355–364. DOI: <https://doi.org/10.1111/1468-0262.00399>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00399>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00399>.

- Ferreira, F. H. and J. Gignoux (2011). “The measurement of inequality of opportunity: Theory and an application to Latin America”. In: *Review of Income and Wealth* 57.4, pp. 622–657. ISSN: 00346586. DOI: [10.1111/j.1475-4991.2011.00467.x](https://doi.org/10.1111/j.1475-4991.2011.00467.x).
- Ferreira, F. H., J. Gignoux, and M. Aran (2011). “Measuring inequality of opportunity with imperfect data: The case of Turkey”. In: *Journal of Economic Inequality* 9.4, pp. 651–680. ISSN: 15691721. DOI: [10.1007/S10888-011-9169-0/METRICS](https://doi.org/10.1007/S10888-011-9169-0/METRICS). URL: <https://link.springer.com/article/10.1007/s10888-011-9169-0>.
- Filmer, D. and L. H. Pritchett (2001). “Estimating Wealth Effects without Expenditure Data-or Tears: An Application to Educational Enrollments in States of India”. In: *Demography* 38.1, pp. 115–132. ISSN: 00703370, 15337790.
- Filmer, D. and K. Scott (2012). “Assessing Asset Indices”. In: *Demography* 49.1, pp. 359–392.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). “The Elements of Statistical Learning”. In: Springer Series in Statistics. DOI: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7). URL: <http://link.springer.com/10.1007/978-0-387-84858-7>.
- Horvitz, D. G. and D. J. Thompson (1952). “A Generalization of Sampling Without Replacement From a Finite Universe”. In: *Journal of the American Statistical Association* 47.260, pp. 663–685. ISSN: 01621459. URL: <http://www.jstor.org/stable/2280784> (visited on 03/26/2024).
- Kiewiet de Jonge, C. P., G. Langer, and S. Sinozich (2018). “Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression with Post-stratification (MRP)”. In: *Public Opinion Quarterly* 82.3, pp. 419–446. ISSN: 0033-362X. DOI: [10.1093/poq/nfy023](https://doi.org/10.1093/poq/nfy023). eprint: <https://academic.oup.com/poq/article-pdf/82/3/419/26102399/nfy023.pdf>. URL: <https://doi.org/10.1093/poq/nfy023>.
- Lehtonen, R. and A. Veijanen (2009). “Chapter 31 - Design-based Methods of Estimation for Domains and Small Areas”. In: *Handbook of Statistics*. Ed. by C. Rao. Vol. 29. Handbook of Statistics. Elsevier, pp. 219–249. DOI: [https://doi.org/10.1016/S0169-7161\(09\)00231-4](https://doi.org/10.1016/S0169-7161(09)00231-4). URL: <https://www.sciencedirect.com/science/article/pii/S0169716109002314>.

- McKenzie, D. J. (2005). “Measuring Inequality with Asset Indicators”. In: *Journal of Population Economics* 18.2, pp. 229–260. ISSN: 09331433, 14321475. URL: <http://www.jstor.org/stable/20007957> (visited on 03/26/2024).
- Newhouse, D., S. Shivakumaran, S. Takamatsu, and N. Yoshida (2014). *How survey-to-survey imputation can fail*. Policy Research Working Paper Series 6961. The World Bank.
- Park, D. K., A. Gelman, and J. Bafumi (2017). “Bayesian Multilevel Estimation with Post-stratification: State-Level Estimates from National Polls”. In: *Political Analysis* 12.4, pp. 375–385. DOI: [10.1093/pan/mp024](https://doi.org/10.1093/pan/mp024).
- Pfeffermann, D. (2013). “New Important Developments in Small Area Estimation”. In: *Statistical Science* 28.1. DOI: [10.1214/12-sts395](https://doi.org/10.1214/12-sts395). URL: <https://doi.org/10.1214/12-sts395>.
- Plassot, T., I. Soloaga, and P. Torres (2022). “Inequality of Opportunity in Mexico and its Regions: A Data-Driven Approach”. In: *The Journal of Development Studies*, pp. 1–17. ISSN: 0022-0388. DOI: [10.1080/00220388.2022.2055465](https://doi.org/10.1080/00220388.2022.2055465). URL: <https://www.tandfonline.com/doi/abs/10.1080/00220388.2022.2055465>.
- Poirier, M. J. P., K. A. Grépin, and M. Grignon (2020). “Approaches and Alternatives to the Wealth Index to Measure Socioeconomic Status Using Survey Data: A Critical Interpretive Synthesis”. In: *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement* 148.1, pp. 1–46. DOI: [10.1007/s11205-019-02187-1](https://doi.org/10.1007/s11205-019-02187-1).
- Rodas, P. C., I. Molina, and M. Nguyen (2021). “Pull your small area estimates up by the bootstraps”. In: <https://doi.org/10.1080/00949655.2021.1926460> 91.16, pp. 3304–3357. ISSN: 15635163. DOI: [10.1080/00949655.2021.1926460](https://doi.org/10.1080/00949655.2021.1926460).
- Roemer, J. E. (1998). “Equality of Opportunity”. In: *Cambridge, MA: Harvard*. Harvard University Press. URL: <https://www.hup.harvard.edu/catalog.php?isbn=9780674004221>.
- Schaible, W. (2014). “Composite Estimators”. In: *Wiley StatsRef: Statistics Reference Online*. John Wiley Sons, Ltd. ISBN: 9781118445112. DOI: <https://doi.org/10.1002/9781118445112.stat05696>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat05696>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05696>.

- Sinha Roy, S. and R. van der Weide (2023). “Poverty in India Has Declined over the Last Decade But Not As Much As Previously Thought”. In.
- Suárez–Arbesú, C., M. R. Vicente, and A. J. López-Menéndez (2024). “An approach to social mobility in African countries: Is there a transmission of education, occupation, or income from parents to children?” In: *Research in Social Stratification and Mobility* 90, p. 100893. ISSN: 0276-5624. DOI: <https://doi.org/10.1016/j.rssm.2024.100893>. URL: <https://www.sciencedirect.com/science/article/pii/S0276562424000064>.
- Torche, F. (2015). “Intergenerational mobility and gender in Mexico”. In: *Social Forces* 94.2, pp. 563–587. ISSN: 15347605. DOI: [10.1093/sf/sov082](https://doi.org/10.1093/sf/sov082).
- Vélez-Grajales, R., L. Monroy-Gómez-Franco, and G. Yalonetzky (2019). “Inequality of Opportunity in Mexico”. In: *Journal of Income Distribution* 27.3-4. URL: <https://eprints.whiterose.ac.uk/135665/>.
- Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). “Forecasting elections with non-representative polls”. In: *International Journal of Forecasting* 31.3, pp. 980–991. ISSN: 0169-2070. DOI: <https://doi.org/10.1016/j.ijforecast.2014.06.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0169207014000879>.