Apples to Oranges: Imputing Household Income Using Survey Data

Pedro J. Torres L. [*] Luis A. Monroy-Gómez-Franco [†]

Roberto Vélez-Grajales [‡]

September, 2023

Abstract

Accurate and consistent estimation of household income is a crucial aspect of socioeconomic research, policy formulation, and decision-making. However, survey data often suffer from missing or incomplete income information, posing challenges to the validity and reliability of derived conclusions. Typically, researchers resort to Small Area Estimations (SAE) to impute household income (HHI) from survey to survey. SAE can be broadly divided into two groups: *1)* Model-based approaches, and *2)* Design-based approaches. In this paper, we propose combining both methods, resorting to Machine Learning Algorithms for the model-based part and correcting our estimates through empirical estimates of the survey containing information on household income. We show our results by doing 2 exercises: imputing HHI from the ENIGH 2016 into the ENIGH 2018 and then the other way around. We test the robustness of our procedure by looking at performance in different scenarios. Our method can reproduce inequality measures very similar to the empirical ones and approximate empirical distributions even when these are highly uneven.

JEL classification: C40, C53, D63.

Keywords: Machine Learning, Small Area Estimation, Mexico.

[*]Department of Social Policy and III - LSE, London

[†]Department of Economics - University of Massachusetts, Amherst

[‡]Centro de Estudios Espinosa Yglesias, CDMX

Introduction

A key challenge to distributional analyses in developing countries is that information on income or consumption is either completely unavailable or not present in surveys that include other relevant variables for such analyses, such as the conditions of origin of the respondent. In the first case, that of the complete lack of information about income or consumption, alternatives such as household asset indexes have been developed to fill that gap (Filmer & Pritchett, 2001; Filmer & Scott, 2012; Poirier et al., 2020). For 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). Our contribution is to this second branch of the literature. We provide an alternative survey-to-survey method that, combining statistical modelling with survey design-based bias estimations, allows us to reconstruct the distribution of income present in a survey dedicated to recovering that type of information using information available in a survey with detailed socio-demographic information about the same population.

Among the imputation methods that have been developed, a substantial part of the literature has focused on developing regression-based imputation methods, known as small area estimations (SAE). These methods have their origin in the seminal contribution by (Elbers et al., 2003). SAE combines survey data with auxiliary information, such as census or administrative records, to generate more precise estimates for small geographic areas or subpopulations.

SAE methods can be broadly categorised into two groups (Corral et al., 2022):

1. **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 & Veijanen, 2009; Pfeffermann, 2013).

2. **Model-based methods** assume a statistical model that relates the variable of interest to auxiliary variables and accounts for the random variation between and within small areas. A commonly used approach is the Random Effects Linear Model (Elbers et al., 2003; Pfeffermann, 2013).

The applications of SAE methods, especially model-based ones, have been driven by the increasing demand for more detailed and timely information for countries where we lack precise income or consumption data. As new data sources and computational tools arise, SAE methods have been applied to a wider range of domains and contexts, including the evolution of poverty in India (Sinha Roy & van der Weide, 2023), health (Viljanen et al., 2022), wealth (Suss et al., 2023), digital inequalities (Singleton et al., 2020) and poverty mapping (Corral et al., 2022). From a methodological aspect, researchers are looking into incorporating machine learning and more flexible regression tools to enhance the estimation process in model-based settings (Corral et al., 2022; Singleton et al., 2020; Viljanen et al., 2022).

When looking at model-based estimations Newhouse et al. (2014) show that three key assumptions have to be met for the imputations to be reliable: *i)* The two surveys have the same set of questions that explain HHI. *ii)* The functional form, specifically the coefficients, must be stable and consistent across surveys (or over time). *iii)* The auxiliary variables are sufficiently correlated with HHI.

According to Newhouse's findings, a consistent downward bias emerges in estimating poverty measures when employing the traditional Small Area Estimation (SAE) approach when any of the three assumptions is not met.

Building upon this, we assert that *i* and *ii* require strong adherence for imputations to yield reliable estimates. However, *iii* can be met with a lower threshold, supplemented by design-based estimations. Even if assumption *iii* is weakly met, incorporating design-based estimations can enhance the robustness of model-based imputations, ensuring reliability in estimating HHI and related measures. This nuanced approach addresses the challenges posed by imperfect correlation between auxiliary variables and our variable of interest, strengthening the validity of survey-to-survey imputations.

In Mexico, two surveys are suitable for this study. The ESRU EMOVI Survey for Social Mobility, conducted every five years by the Centre for Studies Espinosa Yglesias (CEEY), captures information on parental background. However, it falls short in providing a measure of household income or consumption, making it challenging to analyse social mobility and inequality of opportunity in economic well-being. On the other hand, the National Survey

on Household Income and Spending (ENIGH), conducted every two years by the National Institute for Geography and Information (INEGI), offers valuable income and consumption data but lacks information on parental background.

To address this gap, Vélez-Grajales et al. (2019) employed the Small Area Estimation approach proposed by Elbers et al. (2003). They imputed household income from the ENIGH 2010 into the EMOVI 2011. Notably, their findings revealed a Gini coefficient 10 basis points below the estimated Gini for Mexico in 2010 when using the ENIGH 2010, aligning with the downward estimation bias observed by Newhouse et al. (2014).

We impute HHI from ENIGH 2016 into ENIGH 2018 and vice versa to test the validity of the three assumptions. we confirm that assumptions *i* and *ii* are met. However, a relatively modest correlation between auxiliary variables and HHI results in a 10-point reduction in the Gini coefficient, in line with the imputation in Vélez-Grajales et al. (2019) and the results of Newhouse et al. (2014).

To enhance the imputation process, we propose correcting the model-based estimations with a design-based approach. Our procedure integrates statistical modelling through machine learning (model-based) while adjusting estimations using parameters from the original survey (design-based). Our results demonstrate a notable improvement in the estimation of inequality measures compared to those presented by Vélez-Grajales et al. (2019).

The remaining of the project is structured as follows: Section 1 presents the framework form where we build our imputation approach; Section 2 discusses the data and transformations that are done; Section 3 present our main findings followed by robustness checks in Section **??**; Section 5 presents an illustration using the EMOVI 2017; Finally Section 6 cocnludes.

1 Methods

To impute household income from one survey to the other, our task is to approximate HHI through a model of the form:

$$y_h = f(X_h) + \varepsilon_h \tag{1}$$

where y represents per capita HHI for household h explained through a functional form f of some covariates X that have predictive power over HHI and are specific for each household, making it a unit-level model approach.

Traditionally, following the model proposed by Elbers et al. (2003) we estimate

$$log\ y_{hc} = \alpha + X_{hc}\beta + \eta_c + \varepsilon_{hc} \tag{2}$$

where c denotes the cluster or area to which household h belongs.[1] X denotes a matrix of observable auxiliary variables common to both surveys. β represents a vector of coefficients, indicating the direction and marginal contribution of each observable variable in the household income estimation. η signifies the random effect term, capturing unobserved or constant characteristics specific to each location or region that may be related to our error term. ε represents the error term, accounting for the random variability or unexplained factors affecting household income, which the model does not capture.

It is essential to acknowledge that we lose information during the process. 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:

$$Var(\hat{y}) < Var(y) \tag{3}$$

where the reduction in variance during the prediction process can be attributed to three primary factors (Corral et al., 2022):

1. Observable Factors $f(X_{hc})$: These are measurable and known factors contributing to the variation in household income, such as the assumed relation and interactions between variables.

2. Unobservable Factors ε_{hc}: These are unmeasured factors that influence household income but are not directly captured in the data. They represent the random variability or error in the prediction model and are assumed to be normally distributed around 0.

[1] Note that clusters or areas are not necessarily geographic areas. These can also represent population subgroups.

3. Location Specific Factors η_c: This factor arises due to the concentration of income, wealth or consumption in specific geographic areas. It could result from the location of industries, job opportunities, or other regional economic factors.

Note that β is not intended to capture the effect of x over y but rather to approximate (or predict) y as closely as possible through the observable factors.[2] When estimating income or consumption, it is advisable to include variables closely related to income or consumption (Elbers et al., 2003; Sinha Roy & van der Weide, 2023). Sinha Roy and van der Weide (2023) use a dummy that denotes the consumption of a certain good or service, for example, which will decrease the loss of variance due to the observed factors. However, these variables may not always be available, as for the two surveys in Mexico. This translates to a greater loss in variance due to unobserved factors.

Another possible source of bias in our imputation may arise from the fact that the model proposed in Elbers et al. (2003) is the empirical best linear unbiased predictor (EBLUP) for the logarithm of HHI but not for actual HHI. When assessing poverty or inequality measures, we look at the untransformed value of HHI, which may introduce some bias (note that although we may introduce some bias, this would still produce the EBLP (Pfeffermann, 2013)).

To overcome this, we propose a comprehensive method to impute data from one survey (source) to another (target) by leveraging a statistical model and employing adjustment ratios. Our imputation process involves the following steps:[3]

1. **Training the model**: We train a statistical model over the source survey to approximate Equation 1. For our source survey, we use a 5-fold cross-validation procedure to evaluate the out-of-sample prediction of our model. We assess the accuracy using the R^2 metric to ensure we lose the least variance possible.[4]

   We showcase our methodology using an OLS following the model in Elbers et al. (2003). However, the goal in this setting is not to interpret the direct effect of the covariates

[2] One could then opt for more flexible algorithms to approximate HHI such as Random Forest or a Lasso regression as shown in Appendix **??**.

[3] Here we describe the general process of the methodology, specifics of the model and data used are discussed in a latter section of the paper.

[4] This procedure can be extended to using any metric such as MSE or RMSE.

x over y but rather to approximate $f(X)$ as closely as possible. This allows for the use of any regression-based algorithm, such as linear regression, a penalised version of this, random forest, support vector machines, etc. We show the results of an XGBoost regressor and the ones from a K-Nearest-Neighbour regression in Appendix **??**.

2. **Evaluating the model**: We evaluate the performance of our model by looking at how closely our approximations are to the original data in the source survey. As we estimate y using its logarithm, we re-scale our variable and compute two types of bias.

   For each cluster or area c, we estimate how far our estimations are getting in what we call the between-cluster ratio

   $$BC_{ratio} = \frac{\mu_c}{\hat{\mu}_c} \tag{4}$$

   that quantifies how much our model over or underestimates the true value on average for each cluster.

   Within each cluster, we sort individuals into percentile ranks and estimate how far off our predictions are within each rank r of a specific cluster, estimating what we call the within-cluster ratio

   $$WC_{ratio} = \frac{\mu_c^r}{\hat{\mu}_c^r} \tag{5}$$

   that measures how much our model over or underestimates the target variable on average for each rank within each cluster.

   In contrast to Newhouse et al. (2014), we choose our clusters depending on the representativeness of our two surveys. We do this to assume that the cluster mean and variance are the true values, which is essential for design-based methods as discussed in Pfeffermann (2013). Since the ENIGH is representative at the state level, we define our clusters at the state level.

   The distribution of the ratios can be found in Appendix **??**.

3. **Predicting over the target**: We then apply our trained model to predict HHI on the target survey using the same set of covariates used for the training procedure.

   We predict over a labelled set, enabling us to see how far off our estimations are

on completely unseen data. Hence, we have a measure of how far off our original predictions are outside of the training sample.

4. **Adjusting predictions**: Finally, we adjust our prediction on the target survey using the bias ratios computed in step 2. We employ a weighted average of both ratios.

$$AR = \alpha BC_{ratio} + (1 - \alpha) WC_{ratio} \tag{6}$$

our final prediction is thus $exp(\hat{y_h}) * AR$. Since our test data set is labelled, we are able to compute if our corrections are properly adjusting the data outside of the training sample.

The weight α is chosen to minimise the MSE and deviation of our preferred inequality or poverty measure on the source survey (Pfeffermann, 2013).[5] We use a cross-validation procedure over the source survey and select α at one standard deviation from the one that minimises both MSE and our measure to avoid overfitting (Chen & Yang, 2021; Hastie et al., 2009).

The α that minimises MSE and deviations from the Gini coefficient is set at 0.41 for the forward imputation. In the backward imputation process, it is set to 0.42.

We opt for the weighted mean of both ratios as a better solution because it balances out some of the advantages and disadvantages of each individual ratio. Using each adjustment ratio corrects for different types of bias.

The between-cluster ratio corrects for the difference in the mean of the target variable between each cluster in the prediction. Suppose the prediction has a higher mean than the original survey for a certain cluster. In that case, the cluster adjustment ratio will be less than one, reducing the predictions for that cluster in the second survey.

However, the between-cluster adjustment does not correct for the difference in the shape or the variance of the distribution of the target variable within each cluster. The within-cluster ratio adjusts for the difference in the value of the target variable between each rank

[5] We focus on the Gini coefficient in this example, but the procedure can be extended to any other accuracy and inequality/poverty measure.

within each cluster in the prediction and the source survey. Suppose the prediction has a higher value than the second survey for a certain rank within a certain cluster. In that case, the rank adjustment ratio will be less than one, reducing the prediction for that rank in that cluster in the prediction.

However, the within-cluster adjustment does not account for individuals that are misclassified in a different rank than where they should originally be; thus, any systematic biases introduced by the model will be exacerbated, resulting in an increase in the error.

The mean of the two adjustment ratios is a compromise between these two methods. It corrects for the difference in the mean and the distribution of the target variable between each cluster. However, it does not match the mean of either cluster or rank.

Following Elbers et al. (2003), Newhouse et al. (2014), Pfeffermann (2013), and Sinha Roy and van der Weide (2023), we assume that the set of covariates shared across surveys is the same and is measuring the same concept. Second, we assume there is stability between surveys, meaning that $f(X)$ holds between the two surveys. In contrast, we only assume that X correlates with y, but it does not have to be strong. Additionally, we assume that both surveys represent the same population at least at a similar cluster level.[6]

In our setting, assumptions *i*, *ii* and *iv* are necessary to ensure that the adjustment ratios will yield reliable predictions. Assumption *iii* can be met loosely and enhanced using the estimated parameters of the source survey.

2 Data & Setting

To exemplify our methodology, we resort to two datasets in Mexico. We take advantage of the periodicity of the National Survey on Household Income and Spending (ENIGH), which is conducted every two years by the National Institute for Statistics Geography and Information (INEGI). We impute HHI 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 over which the model was not trained.

[6]When this does not hold, Sinha Roy and van der Weide (2023) propose a re-weighting procedure to ensure representability, this is out of the scope of this paper.

Furthermore, we acknowledge that both ENIGHs are constructed under a similar sampling design, and the results may be data-driven. As a robustness check, we carefully subset the ENIGH under different scenarios. By predicting this subset of the ENIGH, we ensure that our methodology extends beyond a similar sampling design.

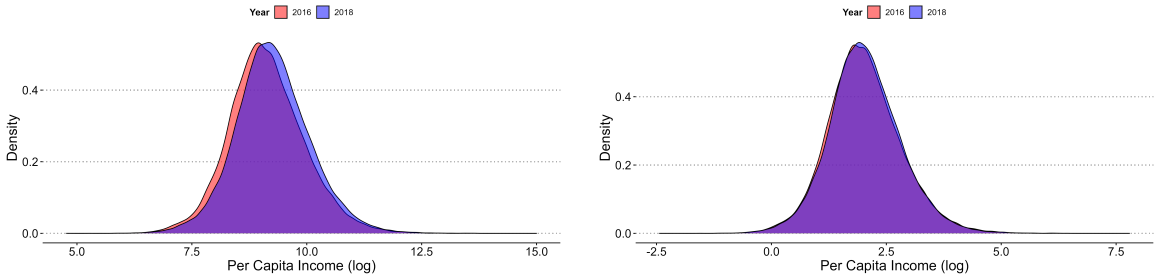
2.1 Data

The ENIGH is a national survey representative of men and women at the national and state levels for urban and rural areas. It collects information about the household composition, income and spending at the household level. It is conducted biennially, and comparability between waves from 2016 onwards is ensured.

We focus on disposable income at the household level and divide it by the number of individuals in the household to measure per-capita household income. We account for economic growth and inflation by re-scaling household income by the poverty line in July each year, as published by COVENAL. We use variables present in ENIGH 2016, ENIGH 2018 and EMOVI 2017 as covariates to ensure that assumption i holds. Specifically, we use household assets, composition, social security status, household head characteristics, and regional variables. Appendix **??** contains a comprehensive list of variables.

Figure 1: Distribution of HHI

(a) Current Income (b) Scaled Income



Notes: The Figure shows the densities of the log of HHI for both ENIGH 2016 and 2018. HHI is rescaled by the poverty line and divided by the number of individuals in a household to represent real per capita terms.

Table 1 shows some summary statistics of the three surveys.

Table 1: Summary Statistics

	ENIGH 2016	ENIGH 2018	EMOVI 2017
HHI	13,555.53	15,283.84	-
HHI*	10.85	10.90	-
log HHI*	2.02	2.04	-
PL Urban	1,348.81	1,521.44	1,471.60
PL Rural	1,015.44	1,145.50	1,120.08

Notes: The Table shows the HHI mean, its log, and the rural and urban poverty lines used in each survey. HHI represents per capita household income. $*$ Dentotes HHI rescaled by the poverty line to represent real per capita terms.

2.2 Setting

We create scenarios for potential problems, as stated by Newhouse et al. (2014) and extend these to the case of missing values. These are summarised in Table 2. We do the exercise of imputing forwards (from 2016 to 2018) and backwards (from 2018 to 2016) to test if the stability of the models holds (assumption *ii*).

Table 2: Imputation Scenarios

	Source	Target
Forward	ENIGH 2016	ENIGH 2018
Backward	ENIGH 2018	ENIGH 2016

Notes: The Table shows the scenarios we account for.

In each scenario, we train our model over 80% of the source survey and validate its result using the remaining 20%. We estimate our adjustment ratios over the whole sample of the source survey using the bootstrap procedure as proposed by Rodas et al. (2021). Once this is done, we test how well our methodology is doing by looking at the imputation in the target survey.

3 Results

We test our methodology by applying it to ENIGH 2016 and 2018. We first present results for the forward imputation, training a model over the ENIGH 2016 and imputing on the

ENIGH 2018. We then present the results for the backward imputation.

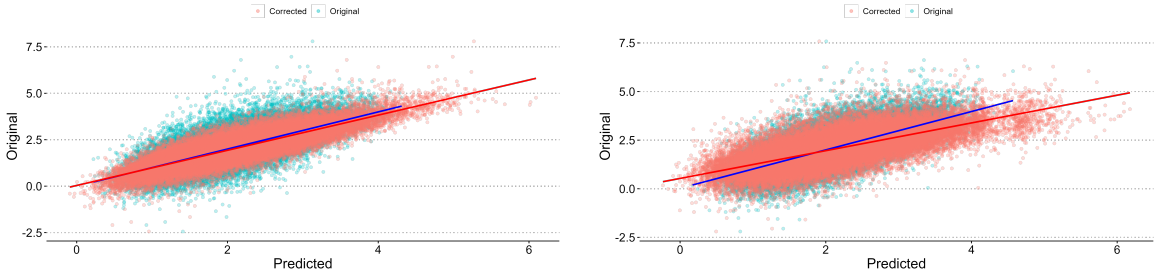
All results in this section are estimated using linear regression as proposed in Elbers et al. (2003). The results for more flexible algorithms can be found in the Appendix of this paper.

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. Over the source survey, our original model exhibits an R^2 of 0.50, which reflects a modest correlation between the predicted and observed values. Our correction procedure increases R^2 to 0.71, a substantial increase. The two can be compared in panel (a) of Figure 2. Appendix **??** shows the distribution of the error terms, reflecting that our original prediction and the correction respect the normality of the error distribution that is assumed in Equation 2.

Figure 2: Correlation Between Predicted and Original Values

(a) Source-2016 (b) Target-2018



Notes: The Figure shows the correlation of the predicted and observed values for the source survey (2016) and the target survey (2018) in panels (a) and (b), respectively. The values are shown in logarithms.

Regarding the target survey, the ENIGH 2018 exhibits a Gini coefficient of 0.464 when considering HHI. With respect to the correlation between the predicted and original values, we find a similar pattern as for the source survey; our correction process improves upon the correlation in the corrections. However, the improvement is milder, which is expected since the model was trained over the source survey, as seen in panel (b) of Figure 2.

The results of our predictions over the target survey and the corresponding corrections are summarised in Table 3. Our original prediction is 12.2 gini points below the real one, consistent with Newhouse et al. (2014) fndings.

Table 3: Forward Imputation Results

	Gini	**RMSE**
ENIGH 2018	0.464	
Original Prediction	0.342	19,945
Between Clustesr	0.344	19,624
Within Clusters	0.556	30,736
Weighted Average	0.472	23,923

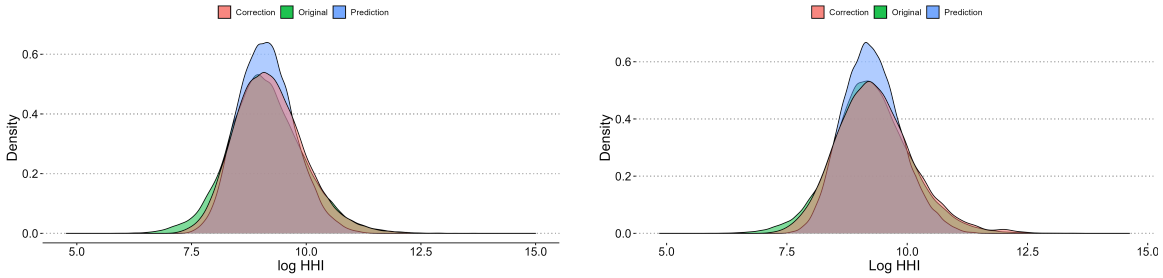
Notes: The Table shows the results of the forward imputation. It shows what each adjustment ratio implies for the Gini coefficient and the RMSE of the real values.

As expected, the BC_{ratio} improves the prediction by correcting for the mean of the clusters. However, the increase in the Gini coefficient is only marginal. The WC_{ratio} exacerbates the systematic errors of the model and increases the Gini to 0.556, overestimating the true coefficient. The weighted average yields a Gini of 0.472, closest to the true value.

Furthermore, Figure 3 shows the distribution of the logarithm of HHI. The original prediction is highly concentrated around the mean. The correction improves and is closer to the observed distribution.

Figure 3: Distribution of Predicted and Original Values

(a) Source-2016 (b) Target-2018



Notes: The Figure shows the distribution of HHI, the predicted and corrected values. Values are shown in the logarithm.

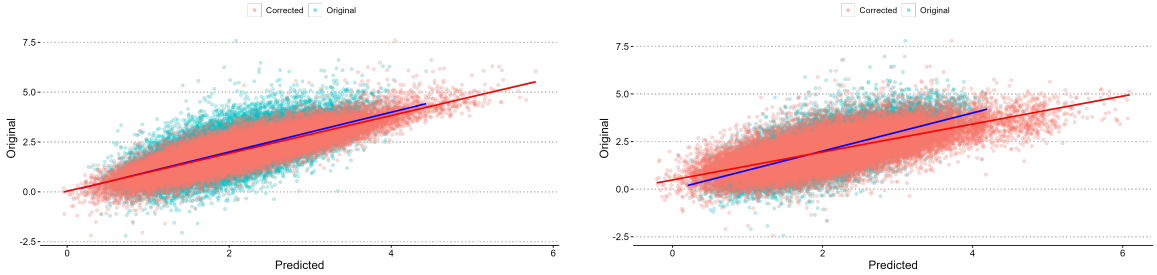
The similarity in the observed and corrected distributions suggests that the model is stable between 2016 and 2018; thus, assumption *ii* holds.

3.2 Backward Imputation

We impute HHI from the ENIGH 2018 to the ENIGH 2016 for the backwards imputation. The results are similar to the ones in the forward process. Over the source survey, our original model exhibits an R^2 of 0.49, which reflects a modest correlation between the predicted and observed values. Our correction procedure increases R^2 to 0.71, similar to what we observe in the forward process. The two can be compared in panel (a) of Figure 4.

Figure 4: Correlation Between Predicted and Original Values

(a) Source-2018 (b) Target-2016



Notes: The Figure shows the correlation of the predicted and observed values for the source survey (2018) and the target survey (2016) in panels (a) and (b), respectively. The values are shown in logarithms.

The results of our predictions over the target survey and the corresponding corrections are summarised in Table 4. The sample of the ENIGH 2016 has a Gini coefficient of 0.472. Our original prediction is 13.3 gini points below the real one.

Table 4: Backward Imputation Results

	Gini	**RMSE**
ENIGH 2016	0.472	
Original Prediction	0.339	21,221
Between Clustesr	0.342	20,929
Within Clusters	0.553	29,172
Weighted Average	0.469	23,786

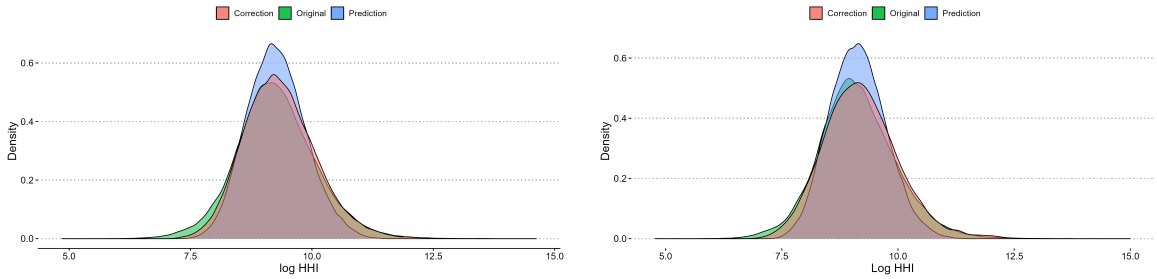
Notes: The Table shows the results of the forward imputation. It shows what each adjustment ratio implies for the Gini coefficient and the RMSE of the real values.

The weighted average yields a Gini of 0.467, closest to the true value. Furthermore, Figure 5 shows the distribution of the logarithm of HHI. As in the forward prediction, the

original prediction shows a higher concentration around the mean. The correction improves and is closer to the observed distribution.

Figure 5: Distribution of Predicted and Original Values

(a) Source-2018 (b) Target-2016



Notes: The Figure shows the distribution of HHI, the predicted and corrected values. Values are shown in the logarithm.

Looking closer at Tables 3 and 4, we see that our imputation matches the crossed Gini coefficient, meaning that the forward imputation correctly estimates the coefficient in the ENIGh 2016 and vice versa. We expect the distribution to match our source surveys in a context where we do not have data available. Additionally, Figure 5 shows that the stability of the model holds; however, stability seems to be stronger for the forward imputation process.

4 Robustness Checks

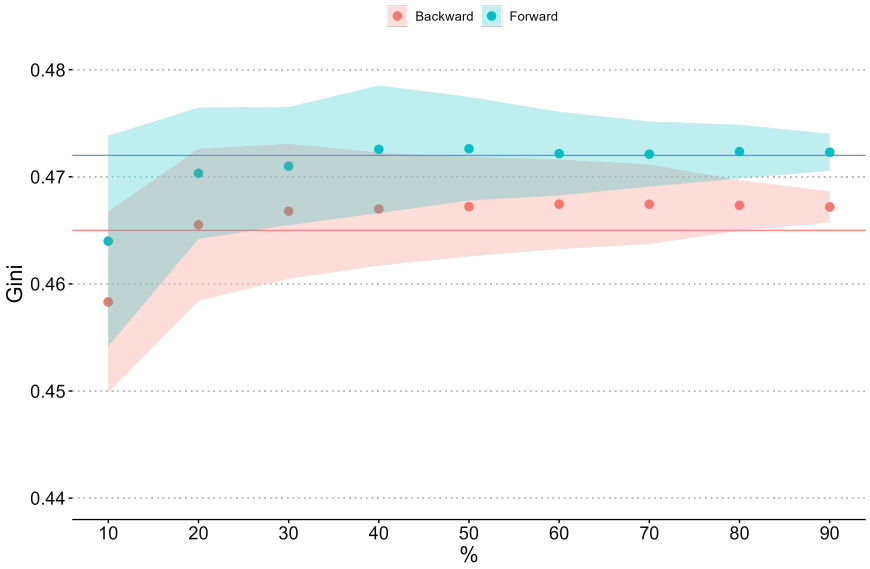
Since both ENIGHs are constructed under a similar sampling framework, we acknowledge that our results may be partially data-driven. We test our procedure by modifying our database to account for this fact.

We first account for sample size and impute over smaller sample sizes. We draw 100 bagged samples at each sample size. We randomly draw from the following specifications: *i)* Random sample, *ii)* higher probability of being drawn if an observation is at the bottom of the distribution, *iii)* higher probability of being drawn if an observation is at the top of the distribution, and *iv)* higher probability of being drawn if an observation is in the middle of the distribution.

By accounting for different specifications, we test if our framework can correctly impute

the Gini coefficient even when the sample is expected from a different sampling procedure. Figure 6 shows the results.

Figure 6: Sample Size and Gini



Notes: The Figure shows the point estimate for the Gini at different sample sizes. The shadowed area represents the 95% confidence interval estimated using 100 bagged samples. The solid lines represent the expected value of the Gini.

Overall, the procedure is robust to different sample sizes and specifications. We see that the confidence interval gets smaller as the sample size increases, signalling a finer estimation. The estimation of the Gini is more precise when we do the forward imputation (from 2016 to 2018). For the backward imputation, we overestimate the true Gini. However, our estimations are less than 1 gini point apart from the true value. From the previous section, we know that we are approximating the Gini from the source survey; we take this value as our benchmark.

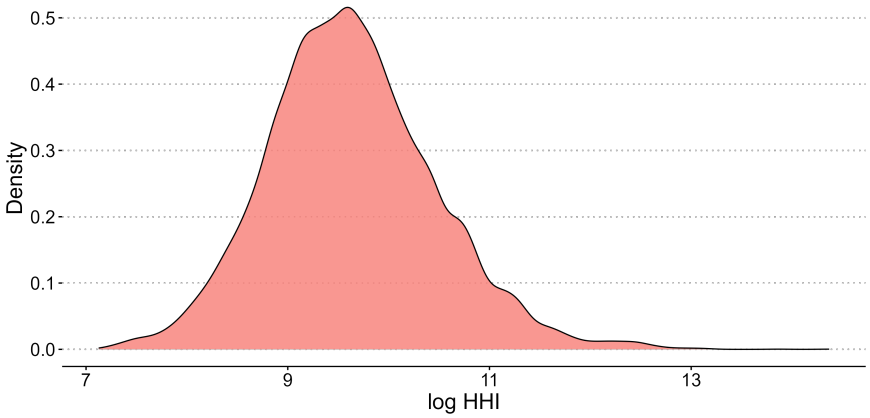
Following the discussion from the previous section, we find evidence for the stability of the model. This assumption holds even when changing the sample size and the distribution of the observations.

5 IOp in Mexico: Assets vs. Income

Finally, we apply our methodology to the EMOVI 2017 and compare our results to those using the asses index approach (Filmer & Pritchett, 2001; Filmer & Scott, 2012). The EMOVI 2017 is a national survey representative of men and women aged 25 to 64 at the national and regional level and for urban and rural areas. The regions are divided into 5 big regions: 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.

Since the EMOVI is only representative at the regional level, we modify the aforementioned procedure and set the clusters at the regional level. This modification results in an α of 0.43 and an R^2 of 0.68. The distribution of our ratios can be found in Appendix B. We use the forward approach and impute from the ENIGH 2016 into the EMOVI 2017. The distribution of our imputed income measure can be found in Figure 7.

Figure 7: Density of Imputed Income



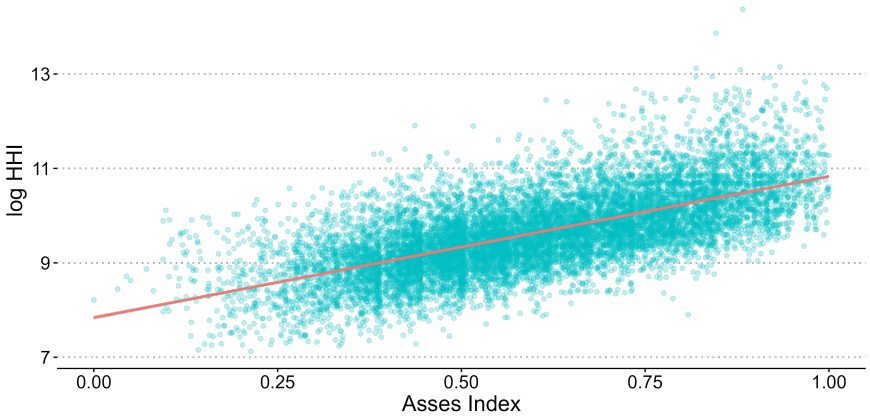
Notes: The Figure shows the distribution of the imputed HHI on the EMOVI 2017 survey.

5.1 Data

The survey collects information about the current household and that of the respondent when she was 14. The recollection of childhood information allows for intergenerational analysis. Respondents are asked about the education and occupation of their parents, as well as the availability of assets in both households.

Since no income data exists, researchers often approximate economic well-being through an asset index (Vélez-Grajales et al., 2019)[Agregar mas citas]. The index is constructed separately for both households using Principal Component Analysis. We follow the same procedure. The specifics of the PCA can be found in Appendix **??**. Figure 8 shows the correlation between the asset index and our imputation of per capita HHI.

Figure 8: Correlation Between HHI and Asset Index



Notes: The Figure shows the correlation between our asset index and the imputed HHI value. HHI is presented in logarithmic scale. The asset index is normalised to take values between 0 and 1.

We find a positive correlation ($\rho = 0.39$) between the imputed value for Household Income (HHI) and the PCA build asset index. This positive correlation signifies a mild linear relationship between the two variables. The HHI provides insights into income, while the PCA build asset index encapsulates a multidimensional representation of household assets more closely related to long-run consumption (Filmer & Scott, 2012). These findings suggest an important difference between what each variable measures and thus highlight the importance of complementing the analysis with HHI as a measure of economic well-being.

5.2 Conceptual Framework and Estimation Strategy

We follow Roemer (1998) and define our outcome to be expressed by an additively separable function of effort and circumstance

$$y_i = f(C_i, e_i) \tag{7}$$

the population can then be divided into k non-overlapping groups based on their circumstances and m tranches of effort that are assumed to be orthogonal to circumstances. We assume equality of opportunity (EOp) if there is no difference in the expected outcome between groups (ex-ante) or between tranches (ex-post). Note that the ex-ante approach requires no information about effort and can be estimated by looking only at the circumstances. When considering a weak ex-ante criterion, we look at the mean of each specific type

$$\hat{y}_i = \mu_j \ \forall j \in [1, ..., k] \tag{8}$$

where we define IOp as first-order stochastic dominance between types.

Using the parametric approach, we follow Ferreira and Gignoux (2011) and estimate Euqation 7 through a linear regression. We use gender, ethnicity, region of birth, father and mother's education, and parental occupation as circumstances. To get a measure of IOp, we first predict our outcome and then estimate the share of total inequality that is explained by the circumstances as:

$$IOp = \frac{Gini(\hat{y})}{Gini(y)} \tag{9}$$

The circumstances chosen for this purpose are not exhaustive and will, therefore, yield a lower bound estimation of IOp.

5.3 Results and Comparisson

Table 5 shows the results of Equation 7 for both the asset index and our imputed HHI. Our results suggest that circumstances have the same direction for both variables. We find a negative correlation to women and those who identify themselves as indigenous. Education

of both father and mother has a positive correlation with both variables in both cases; one year of education of the mother is associated with a bigger increase in outcomes than one year of education of the father.

Table 5: Correlation Between Circumstances and Outcomes

	Dependent variable:	
	Asset Index	Household Income
	(1)	(2)
Women	-0.017^{***}	$-10{,}867.020^{***}$
	(0.003)	(605.060)
Indigenous	-0.042^{***}	$-5{,}174.758^{***}$
	(0.005)	(998.671)
Education Father	0.004^{***}	831.247^{***}
	(0.000)	(86.379)
Education Mother	0.008^{***}	941.169^{***}
	(0.000)	(89.956)
Constant	1.067^{***}	$28{,}655.460^{***}$
	(0.635)	(3,193.528)
Controls:		
Birth Region	X	X
Occupation Father	X	X
Observations	12,015	12,015
R^2	0.190	0.111
Adjusted R^2	0.189	0.110
Residual Std. Error (df = 11996)	0.165	32,621.480
F Statistic (df = 18; 11996)	156.489^{***}	83.545^{***}

*p<0.1; **p<0.05; *p<0.01

Notes: The Table shows the results of an OLS of our normalised asset index and HHI over circumstances. Standard errors are presented in parentheses.

Furthermore, we analyse differences in IOp between the two measures. Table 6 summarises our findings. Compared to the asset index, HHI shows an increase in relative IOp of more than 10 percentage points. When analysing absolute values, our index shows a Gini coefficient of 0.175, which is more than 30 gini points below the one for HHI.

6 Conclusions

Table 6: Ineqaulity Of Opportunity

	Asset Index	Household Income
Total Inequality	0.175	0.481
Absolute IOp	0.074 [0.073; 0.075]	0.284 [0.282; 0.285]
Relative IOp	0.426 [0.424; 0.428]	0.590 [0.588; 0.592]

Notes: The Table shows the estimates of IOp for our normalised asset index and HHI. 95% Confidence intervals are estimated through the bootstrap procedure and presented in parentheses.

References

Chen, Y., & Yang, Y. (2021). The one standard error rule for model selection: Does it work? *Stats*, *4*(4), 868–892. https://doi.org/10.3390/stats4040051

Corral, P., Molina, I., Cojocaru, A., & Segovia, S. (2022). Guidelines to Small Area Estimation for Poverty Mapping. *Guidelines to Small Area Estimation for Poverty Mapping*. https://doi.org/10.1596/37728

Dang, H.-A. (2021). To impute or not to impute, and how? a review of poverty-estimation methods in the absence of consumption data. *Development Policy Review*, *39*, 1008–1030. https://doi.org/https://doi.org/10.1111/dpr.12495

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro–level estimation of poverty and inequality. *Econometrica*, *71*(1), 355–364. https://doi.org/https://doi.org/10.1111/1468-0262.00399

Ferreira, F. H., & Gignoux, J. (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth*, *57*(4), 622–657. https://doi.org/10.1111/j.1475-4991.2011.00467.x

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data-or tears: An application to educational enrollments in states of india. *Demography*, *38*(1), 115–132.

Filmer, D., & Scott, K. (2012). Assessing asset indices. *Demography*, *49*(1), 359–392.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. https://doi.org/10.1007/978-0-387-84858-7

Lehtonen, R., & Veijanen, A. (2009). Chapter 31 - design-based methods of estimation for domains and small areas. In C. Rao (Ed.), *Handbook of statistics* (pp. 219–249). Elsevier. https://doi.org/https://doi.org/10.1016/S0169-7161(09)00231-4

Newhouse, D., Shivakumaran, S., Takamatsu, S., & Yoshida, N. (2014). *How survey-to-survey imputation can fail* (Policy Research Working Paper Series No. 6961). The World Bank.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, *28*(1). https://doi.org/10.1214/12-sts395

Poirier, M. J. P., Grépin, K. A., & Grignon, M. (2020). Approaches and Alternatives to the Wealth Index to Measure Socioeconomic Status Using Survey Data: A Critical Interpretive Synthesis. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, *148*(1), 1–46. https://doi.org/10.1007/s11205-019-02187-

Rodas, P. C., Molina, I., & Nguyen, M. (2021). Pull your small area estimates up by the bootstraps. *https://doi.org/10.1080/00949655.2021.1926460*, *91*(16), 3304–3357. https://doi.org/10.1080/00949655.2021.1926460

Roemer, J. E. (1998). Equality of Opportunity. In *Cambridge, ma: Harvard*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674004221

Singleton, A., Alexiou, A., & Savani, R. (2020). Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation. *Computers, Environment and Urban Systems*, *82*, 101486. https://doi.org/10.1016/j.compenvurbsys.2020.101486

Sinha Roy, S., & van der Weide, R. (2023). Poverty in India Has Declined over the Last Decade But Not As Much As Previously Thought.

Suss, J., Kemeny, T., & Connor, D. (2023). GEOWEALTH: Spatial wealth inequality data for the United States, 1960-2020. (August), 1–24.

Vélez-Grajales, R., Monroy-Gómez-Franco, L., & Yalonetzky, G. (2019). Inequality of opportunity in mexico. *Journal of Income Distribution*, *27*(3-4). https://eprints.whiterose.ac.uk/135665/

Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kassteele, J. (2022). A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. *International Journal of Health Geographics*, *21*(1), 1–18. https://doi.org/10.1186/S12942-022-00304-5/FIGURES/5

Appendices

A Machine Learning Models

B Ratios

C Variables

D Errors