Apples to Oranges: Imputing Household Income Using Survey Data

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

Roberto Vélez-Grajales [‡]

February, 2024

[Most recent version]

Abstract

Accurate and consistent estimation of household income is essential for socioeconomic research, policy development, and decision-making. However, survey data often lack complete income information, which challenges the validity and reliability of conclusions drawn from such data. Researchers commonly employ Small Area Estimation (SAE) techniques to impute household income across surveys. SAE can be categorized into two main approaches: 1) Model or Regression-based methods and 2) Design-based methods. In our study, we propose a hybrid approach that combines both methods. We use Machine Learning algorithms for the regression-based component and refine our estimates using empirical survey data containing household income information. Specifically, we demonstrate our approach by performing two exercises: imputing household income from the ENIGH 2016 survey to the ENIGH 2018 survey and vice versa. Additionally, we apply this imputation technique to the EMOVI 2017 dataset and estimate inequality of opportunity measures. Our results indicate a promising improvement compared to the uncorrected imputation.

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. These 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 have been looking into incorporating bootstrap estimation (Rodas et al., 2021) and machine learning to enhance the estimation process in regression-based settings (Corral et al., 2022; Singleton et al., 2020; Viljanen et al., 2022).

For regression-based methods to yield reliable imputations, these have to meet three key assumptions (Newhouse et al., 2014). *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 auxiliary variables are sufficiently correlated with the outcome. According to Newhouse's findings, a consistent downward bias emerges in estimating poverty measures when employing the traditional SAE approach and 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, complemented by design-based estimations. Design-based estimations can enhance the robustness of model-based imputations, ensuring reliability in estimating income 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 SAE 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 base points below the estimated Gini for Mexico in 2010 when using the ENIGH, aligning with the downward estimation bias observed by Newhouse et al. (2014). We argue that this reduction results from a mild correlation between the auxiliary and outcome variables.

We impute HHI from the ENIGH 2016 into the 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 more than 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). We propose correcting the model-based estimations with a design-based approach to enhance the imputation process. 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).

Finally, we impute HHI on the EMOVI 2017 and compare inequality of opportunity (IOp) estimations using our imputed measure and an asset index. Both measures of economic well-being show a positive correlation of 0.4. Our imputation yields a Gini coefficient of 0.481, close to the 0.472 observed in the ENIGH 2016, showing an improvement to that of Vélez-Grajales et al. (2019). The asset index shows a coefficient of 0.175. Using the parametric approach (Ferreira & Gignoux, 2011), we find a more than 30 base points difference when estimating absolute IOp and a difference of more than 10 points for relative IOp.

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 4; Section 5 presents an illustration using the EMOVI 2017; Finally Section 6 concludes.

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. 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 related to HHI. 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

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

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.

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. When predicting income or consumption, it is advisable to include variables closely related to these, such as dummies that denote if a household bought a certain good or service (Sinha Roy & van der Weide, 2023), 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, which will, in turn, imply a downward bias in poverty and inequality measures (Newhouse et al., 2014).

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:

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

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

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

XGBoost regressor and the ones from a K-Nearest-Neighbour regression in Appendix A.

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

$$BC^{c}_{ratio} = \frac{\mu_c}{\hat{\mu}_c} \tag{4}$$

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

$$WC^{cr}_{ratio} = \frac{\mu^{r}_{c}}{\hat{\mu}^{r}_{c}} \tag{5}$$

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

In contrast to Elbers et al. (2003), Newhouse et al. (2014), and Sinha Roy and van der Weide (2023), we choose our clusters depending on the representativeness of our two surveys. We do this assuming 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 B.

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 rank individuals within each cluster based on their predicted outcomes.

Our setting lets us see how far off our estimations are on completely unseen data because the prediction is made over a labelled set. 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.

$$AR_{cr} = \alpha BC^{c}_{ratio} + (1 - \alpha) WC^{cr}_{ratio} \tag{6}$$

our final prediction is thus $exp(\hat{y}_{hcr}) * AR_{cr}$. 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).[3] 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 ratio. 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 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.

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

However, the within-cluster adjustment does not account for individuals who are miss-classified in a different rank than where they should originally be; thus, any systematic biases introduced by the model will be exacerbated, overestimating our inequality measure. 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.[4] 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 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.

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.

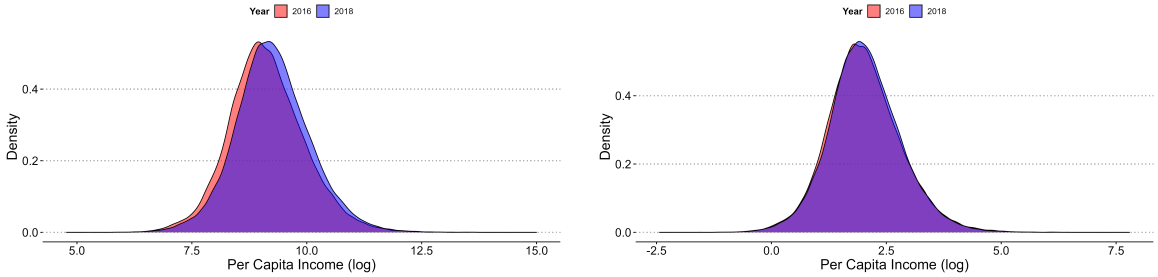
[4]When this does not hold, Sinha Roy and van der Weide (2023) propose a re-weighting procedure to ensure representability, this is out of the scope of this paper.

2.1 Data

The ENIGH is a survey representative of men and women at the national and state levels for urban and rural areas. It collects information about the 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 HHI. 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 the ENIGH 2016, ENIGH 2018 and EMOVI 2017 as auxiliary variables to ensure that assumption *i* holds. Specifically, we use household assets, composition, social security status, household head characteristics, and regional variables. Appendix C contains a comprehensive list of variables.

Figure 1: Distribution of HHI



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 HHI in the three surveys. Mean income increased by around 13% in current terms. In real terms, however, this amounts to an increase of less than 1%.

2.2 Setting

We test to see if assumption *ii* holds by accounting for different scenarios as done by Newhouse et al. (2014). These are summarised in Table 2. We do the exercise of imputing

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. $*$ Denotes HHI rescaled by the poverty line to represent real per-capita terms.

forwards (from 2016 to 2018) and backwards (from 2018 to 2016).

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. Once this is done, we test how well our methodology is doing by looking at the imputation in the target survey.

3 Results

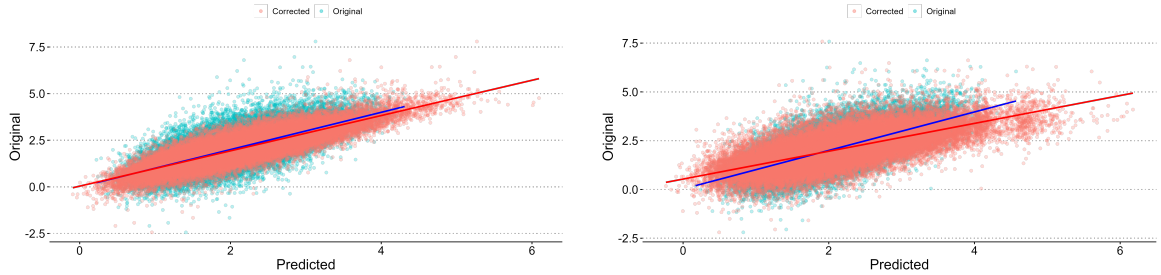
We test our methodology by applying it to the ENIGH 2016 and 2018. We first present results for the forward imputation, training a model over the ENIGH 2016 and imputing on the ENIGH 2018. We then present the results for the backward imputation. All results in this section are estimated using linear 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 D.1 shows the distribution of the error terms, reflecting that our original prediction and the correction respect the normality of the error distribution 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 base points below the real one, consistent with Newhouse et al. (2014) findings.

As expected, the BC^c_{ratio} improves the prediction by correcting for the mean of the clusters. However, the increase in the Gini coefficient is only marginal. The WC^{cr}_{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. Further-

Table 3: Forward Imputation Results

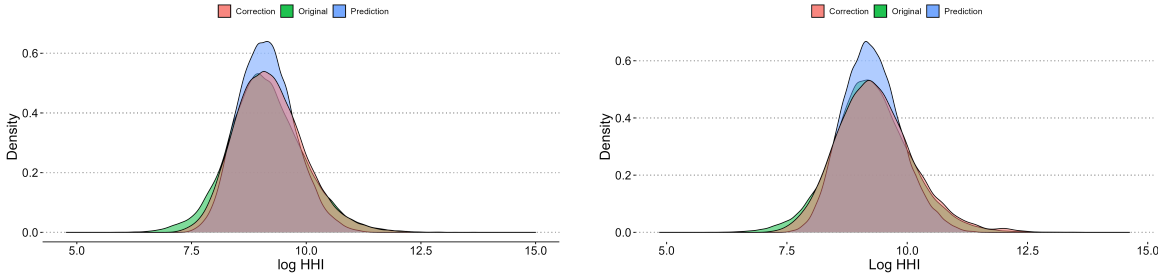
	Gini	**RMSE**
ENIGH 2018	0.464	
Original Prediction	0.342	19,945
Between Clustesr	0.344	19,624
Within Clusters	0.556	30,736
Weighted Average	0.472	23,923

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.

more, 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 empirical 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.

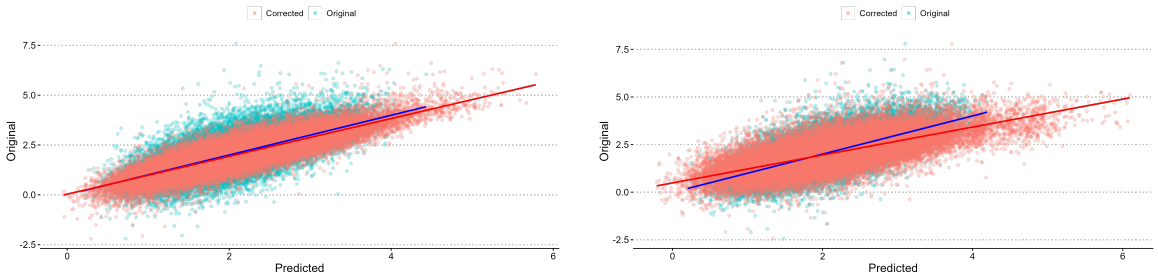
The similarity in the observed and corrected distributions suggests that the model is stable between 2016 and 2018; thus, assumption *ii* holds.

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 base 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

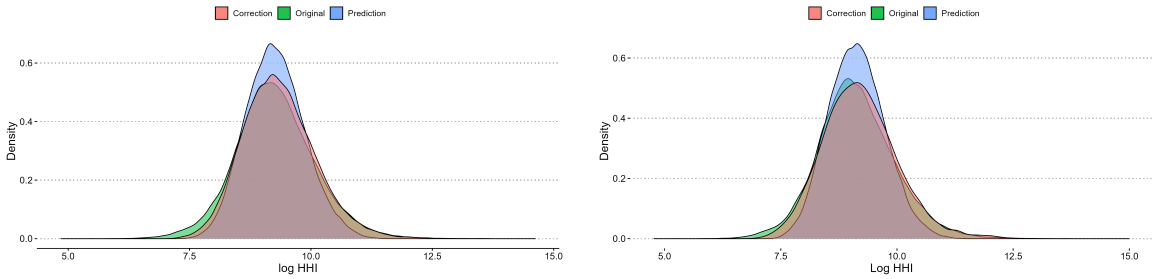
Notes: The Table shows the results of the forward imputation. It shows what each adjustment ratio implies for the Gini coefficient and the RMSE of the real values.

The weighted average yields a Gini of 0.467, closest to the true value. Furthermore, Figure 5 shows the distribution of the log of HHI. As in the forward process, the original prediction shows a higher concentration around the mean. The correction improves and is closer to the empirical distribution.

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

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.

process.

4 Robustness Checks

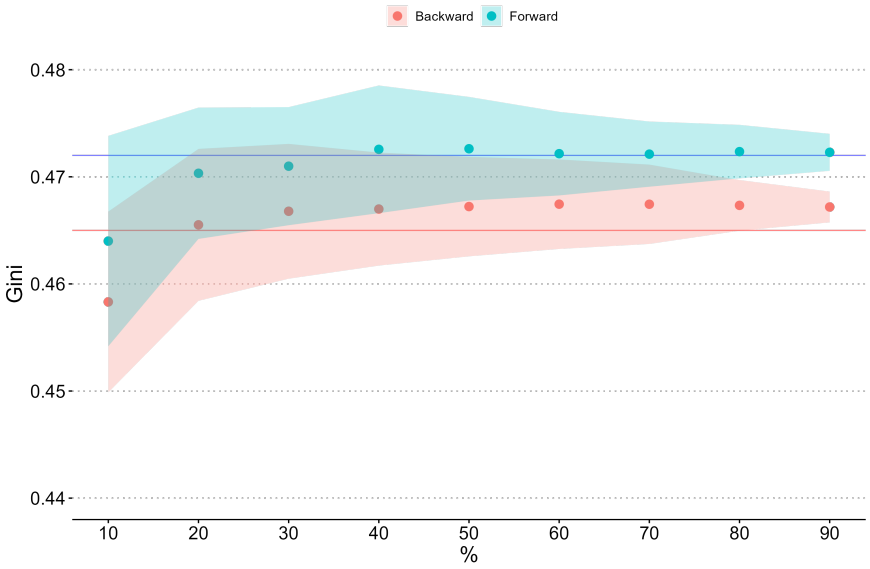
Since both ENIGHs are constructed under a similar sampling framework, we acknowledge that our results may be partially data-driven. We test our procedure by modifying our database to account for this fact.

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 to be drawn from a different sampling procedure. Figure 6 shows the results.

Overall, the procedure is robust to different sample sizes and specifications. 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 base point apart from the true value. From the previous section, we

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.

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.

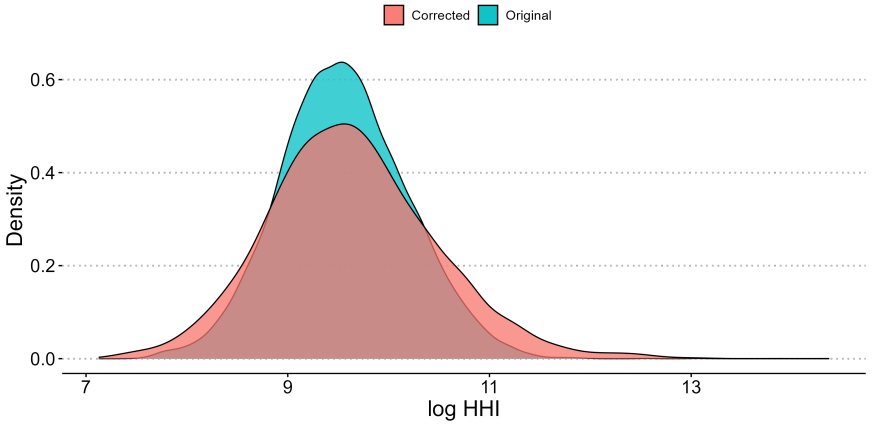
5 IOp in Mexico: Assets vs. Income

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 abovementioned 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.3. We use the forward approach and impute from the ENIGH 2016 into the EMOVI 2017. Figure 7 shows the distribution of our imputed income measure.

Figure 7: Density of Imputed Income



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

The original distribution is highly concentrated around the mean. This high concentration results in a Gini coefficient of 0.35, more than 10 base points below the expected value given the ENIGH 2016. After applying our correction, values have higher variance. This results in a Gini coefficient of 0.48, closer to the expected value.

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.

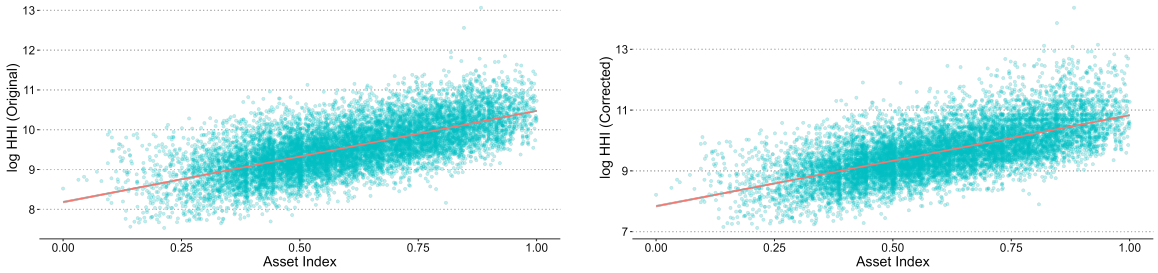
Since no income data exists, researchers often approximate economic well-being through an asset index (eg. Delajara et al., 2022; Torche, 2015; Vélez-Grajales et al., 2019). The index is constructed separately for both households using Principal Component Analysis.

We follow the same procedure. The specifics of the PCA can be found in Appendix E.1. Figure 8 shows the correlation between the asset index and our imputation of per-capita HHI.

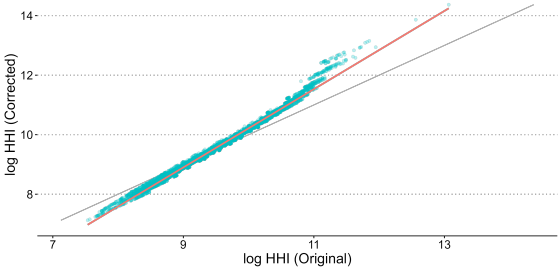
Figure 8: Correlation Between HHI and Asset Index

(a) HHI Original & Index

(b) HHI Corrected & Index



(c) HHI Original & Corrected



Notes: The Figure shows the correlation between our asset index and the imputed HHI value. HHI is presented in a logarithmic scale. The asset index is normalised to take values between 0 and 1. In Panel (c), the grey line represents the 45° line.

We find a positive correlation between the imputed household income value and the PCA-built asset index, with coefficients of 0.55 (uncorrected) and 0.39 (corrected). Additionally, the two income measures exhibit a strong correlation of 0.92. The correlation of our asset index with both income measures is stronger in the middle part of the distribution. The lower and upper tails seem to be less correlated. We find a similar pattern for our income measures at the lower tail when comparing one to the other. Our corrected measure is more dispersed at the upper tail than our original imputation. When looking at Panel (c), we see that our corrections are decreasing the predictions at the bottom and increasing them at the top. The middle part of the distribution stays relatively similar, consistent with what the densities in Figures 3 and 5 show.

While income captures short-term flows, the PCA index provides a multidimensional view of household assets related to long-run consumption (Filmer & Scott, 2012). Consumption

tends to be less volatile and to have lower variance, hence showing lower levels of inequality. Given that our original prediction has a stronger correlation to the index and is heavily correlated with our corrected measure, we can expect lower levels of absolute IOp in our original prediction. However, this would not necessarily result in lower levels of relative IOp due to a decrease in total inequality.

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 yield a lower bound estimation of ex-ante IOp.

5.3 Results and Comparisson

Table 5 shows the results of Equation 7 for both the asset index and our imputed HHI. We find a similar pattern of associations regardless of our outcome variable. Females and indigenous tend to have lower index levels and household income. Parental education (both father and mother) is positively associated with all outcomes.

Table 5: Correlation Between Circumstances and Outcomes

	Dependent variable:		
	Asset Index	HHI (Original)	HHI (Corrected)
	(1)	(2)	(3)
Female	−0.017***	−5,590.331***	−10,867.020***
	(0.003)	(219.602)	(605.060)
Indigenous	−0.042***	−3,324.623***	−5,174.758***
	(0.005)	(362.460)	(998.671)
Education Father	0.005***	392.633***	831.247***
	(0.0004)	(31.351)	(86.379)
Education Mother	0.008***	521.862***	941.169***
	(0.0005)	(32.649)	(89.956)
Constant	0.635***	21,116.550***	28,655.460***
	(0.016)	(1,159.065)	(3,193.528)
Other Circumstances:			
Occupation Father	X	X	X
Region of Birth	X	X	X
Observations	12,015	12,015	12,015
R^2	0.190	0.208	0.111
Adjusted R^2	0.189	0.206	0.110
Residual Std. Error (df = 11996)	0.165	11,839.700	32,621.480
F Statistic (df = 18; 11996)	156.489***	174.599***	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.

Interestingly, in accordance with the results of Ciaschi et al. (2023), mothers' education seems to have a stronger association with the economic well-being of their children than that of the father. All our point estimates take a higher value after the correction. However, this change is accompanied by higher variance, increasing the residual standard error.

Finally, we analyse differences in IOp between the two measures. Table 6 summarises

our findings. Total inequality varies substantially between the three measures. We estimate a Gini coefficient of 0.175 for our asset index.[5] The predicted income has a coefficient of 0.351, an increase of almost 20 base points compared to the index. However, it is still 11 base points lower than the one estimated using the ENIGH 2016. After applying our corrections, we estimate a coefficient of 0.481, an increase of more than 30 base points compared to the indexes and 13 base points compared to that of the original prediction and closest to that estimated in the ENIGH.

Table 6: Inequality Of Opportunity

	Asset Index	HHI (Original)	HHI (Corrected)
Total Inequality	0.175	0.351	0.481
Absolute IOp	0.074 [0.073; 0.075]	0.195 [0.194, 0.196]	0.284 [0.282; 0.285]
Relative IOp	0.426 [0.424; 0.428]	0.557 [0.556, 0.559]	0.590 [0.588; 0.592]

Notes: The Table shows the estimation of different measures of inequality of opportunity for our three variables of interest.

When looking at the absolute values of IOp, we find an important increase between the asset index and our income imputation. As expected, we find the estimated absolute value of IOp to be different between our two income measures due to the lower variance of the original distribution. However, the estimated value of relative IOp is very close between the two due to the strong association. Nevertheless, it results in an underestimation of almost 3 base points.

6 Conclusions

Understanding household income is essential as it provides insights into a household's resources and overall well-being while reflecting the economic conditions and opportunities in a specific region or country. However, its absence or unreliability in survey data can pose challenges for accurate analysis.

In recent years, there have been efforts to fill that gap either by the use of household asset indexes (Filmer & Pritchett, 2001; Filmer & Scott, 2012) or through an imputation

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

process using auxiliary information from a survey containing information on household income (Elbers et al., 2003; Pfeffermann, 2013; Rodas et al., 2021). The former has proven to be more closely related to long-run consumption (Filmer & Scott, 2012) and is normally expected to be less volatile and present lower levels of inequality than income or current consumption; the latter assumes a strong statistical relationship between the auxiliary variables and income. In many cases, such as the Mexican one, this assumption is only met partially.

When this assumption is not fully met, the imputation process will result in a downward bias of poverty estimates and inequality measures (Newhouse et al., 2014). We propose enhancing the imputation in such cases through design-based bias estimation that arises during the imputation process. Our procedure is based on estimating between and within-cluster biases in the source survey, then using this information in the target survey and correcting our imputation to obtain more accurate poverty and inequality results.

We test the reliability of our results by first imputing from the ENIGH 2016 into the ENIGH 2018 and vice versa. Both surveys have income present. Therefore, we can see how far off our estimations are. The original prediction results in a difference of more than 10 base points compared to the true Gini value. After applying the corrections, our estimate of the Gini is less than 1 base point different from the expected one. This result holds when accounting for sample size and sampling design.

We compare our estimation to an asset-based index as proposed by Filmer and Pritchett (2001). Additionally, we compare inequality of opportunity (IOp) measures using the index, the original imputation and our corrected imputation. The imputed measure of income shows a positive but mild correlation to the index, which holds after applying the corrections. However, the asset-based index presents much lower inequality than both income measures. Similarly, our original imputation presents a Gini coefficient of more than 10 base points below the one observed in the ENIGH 2016. Our correction estimates a Gini coefficient close to the expected one. This reduction in variance on the original prediction has a small effect on relative IOp, though it underestimates it by around 3 base points. However, the absolute values of IOp differ dramatically between income measures, showcasing the importance of our procedure.

This procedure proves useful when one wants to impute income or consumption from

one survey to another in contexts where the correlation between the auxiliary variables and the outcome is not too strong. However, it has its limitations. By correcting for biases in the source survey, the procedure seeks to replicate its distribution in the target survey. If one wants to estimate changes in poverty over time, as is necessary in the case of India, this procedure should not be used (in that case refer to Sinha Roy & van der Weide, 2023). Also, if one believes that the sample in the source and target surveys are substantially different, this procedure will present biased results.

Last, we showcase the general procedure and implications by using linear regression and following the model proposed by Elbers et al. (2003). Since the goal of the imputation is not to infer the implications of auxiliary variables over income, this procedure can enhance the prediction in a bootstrap estimation (Rodas et al., 2021) or when using machine learning models for the regression task (Suss et al., 2023; Viljanen et al., 2022) and is therefore not an alternative but rather a complement in settings where the correlation between auxiliary and outcome variables is not strong enough to yield reliable results.

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

Ciaschi, M., Marchionni, M., & Neidhöfer, G. (2023). Intergenerational Mobility in Latin America : The Multiple Facets of Social Status and the Role of Mothers.

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

Delajara, M., Campos-Vazquez, R. M., & Velez-Grajales, R. (2022). The regional geography of social mobility in Mexico. *Regional Studies*, *56*(5), 839–852. https://doi.org/10.1080/00343404.2021.1967310

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.

Torche, F. (2015). Intergenerational mobility and gender in Mexico. *Social Forces*, *94*(2), 563–587. https://doi.org/10.1093/sf/sov082

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

Table A.1: Model Specifications

	XGB	KNN
alpha	0.43	0.38
R^2 Original	0.55	0.44
R^2 Corrected	0.73	0.74

Notes: The Table shows the selected α and R^2 of an XGBoost regression and a KNN regression.

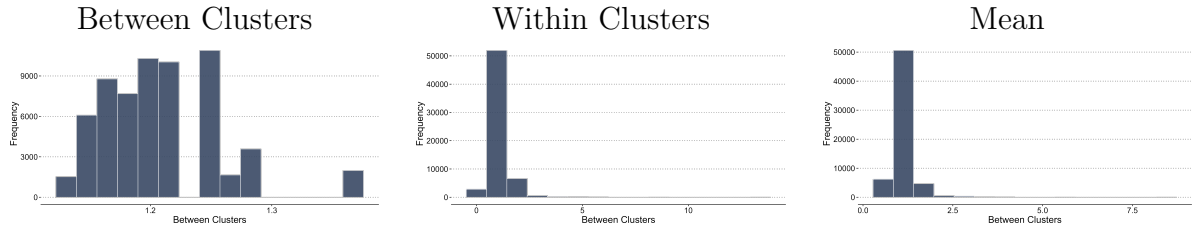
Table A.2: Results with ML Algorithms

	XGB		KNN	
	Gini	RMSE	Gini	RMSE
ENIGH 2018	0.464		0.464	
Original	0.352	19,479	0.297	20,784
Between Clusters	0.355	19,154	0.31	20,165
Within Clusters	0.545	29,469	0.555	29,907
Mean	0.467	22,975	0.467	24,103

Notes: The Table shows the procedure results when applied to an XGBoost regression and a KNN regression.

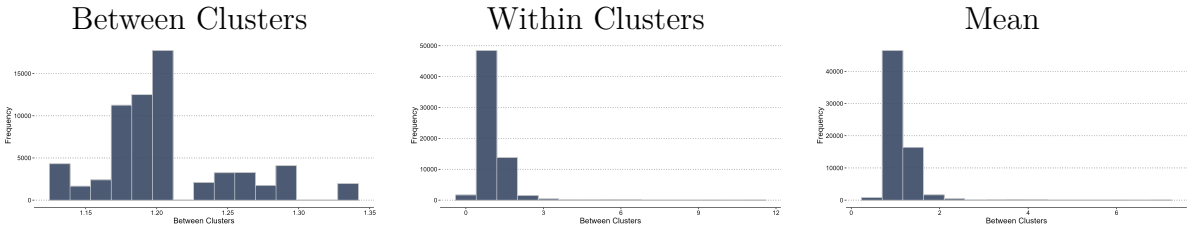
B Ratios

Figure B.1: Distribution of Adjustment Ratios - Forward



Notes: The Figure shows the distribution of the bias estimated as in Equations 4, 5 and 6 in the forward imputation process. Clusters are defined at the state level.

Figure B.2: Distribution of Adjustment Ratios - Backward



Notes: The Figure shows the distribution of the bias estimated as in Equations 4, 5 and 6 in the backward imputation process. Clusters are defined at the state level.

Figure B.3: Distribution of Adjustment Ratios - EMOVI 2017

Notes: The Figure shows the distribution of the bias estimated as in Equations 4, 5 and 6 in the imputation process. Clusters are defined at the state level.

C Variables

Table C.1: Variables

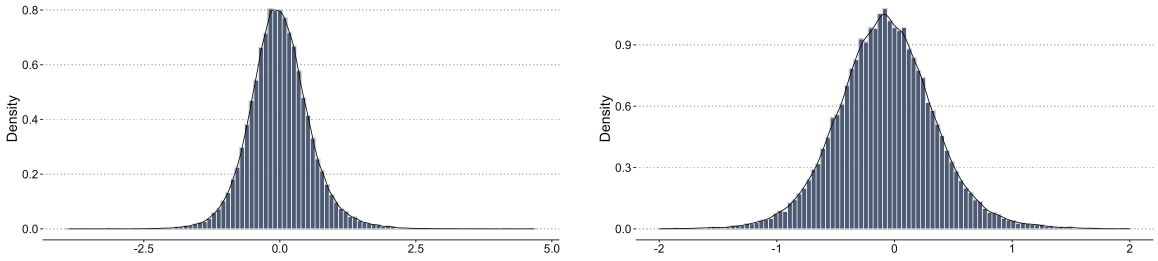
	ENIGH 2016	ENIGH 2018	EMOVI 2017
Telephone	0.29 (0.46)	0.28 (0.45)	0.38 (0.49)
Cellphone	0.86 (0.35)	0.88 (0.32)	0.85 (0.36)
TV	0.5 (0.5)	0.45 (0.5)	0.52 (0.5)
Internet Connection	0.31 (0.46)	0.36 (0.48)	0.43 (0.49)
Water	0.71 (0.45)	0.72 (0.45)	0.92 (0.27)
Electricity	0.98 (0.12)	0.98 (0.13)	0.79 (0.41)
Gender (HH)	0.75 (0.43)	0.73 (0.44)	0.61 (0.49)
Car	0.47 (0.5)	0.47 (0.5)	0.57 (0.77)
Owns Property	0.72 (0.45)	0.72 (0.45)	0.61 (0.49)
Dirt Floor	0.03 (0.17)	0.03 (0.17)	0.03 (0.16)
Cement Floor	0.55 (0.5)	0.55 (0.5)	0.56 (0.5)
Other Floors	0.42 (0.49)	0.42 (0.49)	0.41 (0.49)
Share Men	0.5 (0.23)	0.5 (0.23)	0.49 (0.23)
Share Occupied	0.51 (0.28)	0.52 (0.28)	0.45 (0.28)
IMSS	0.39 (0.49)	0.39 (0.49)	0.55 (0.5)
IMSS Prospera	0.01 (0.08)	0 (0.05)	0.02 (0.13)
ISSSTE	0.01 (0.11)	0.01 (0.11)	0.08 (0.27)
Other (Social Security)	0.59 (0.49)	0.59 (0.49)	0.35 (0.48)
PEMEX (Social Security)	0.01 (0.09)	0.01 (0.09)	0.01 (0.09)
Speaks Indigenous	0.08 (0.27)	0.08 (0.27)	0.13 (0.33)
Prospera	0.21 (0.41)	0.21 (0.41)	0.12 (0.33)
Elderly	0.08 (0.27)	0.07 (0.26)	0.06 (0.24)
Big Town	0.38 (0.49)	0.36 (0.48)	0.09 (0.28)
Median Town	0.13 (0.34)	0.13 (0.33)	0.18 (0.38)
Small/Median Town	0.13 (0.34)	0.13 (0.34)	0.2 (0.4)
Small Town	0.36 (0.48)	0.38 (0.48)	0.53 (0.5)
No Education (HH)	0.06 (0.24)	0.06 (0.24)	0.04 (0.19)
Kindergarden (HH)	0.17 (0.37)	0.16 (0.36)	0 (0.04)
Primary School (HH)	0.22 (0.41)	0.21 (0.41)	0.25 (0.43)
Secondary School (HH)	0.31 (0.46)	0.31 (0.46)	0.3 (0.46)
High School (HH)	0.14 (0.34)	0.14 (0.35)	0.2 (0.4)
Graduate Degree (HH)	0.09 (0.29)	0.1 (0.3)	0.15 (0.35)
Postgraduate Degree (HH)	0.02 (0.13)	0.02 (0.13)	0.01 (0.11)
State of Residency	X	X	X

Notes: The Table shows the variables used for the estimation. The values correspond to the mean in each survey; standard errors are presented in parentheses. HH refers to the household head.

D Errors

Figure D.1: Error Distribution - Forward

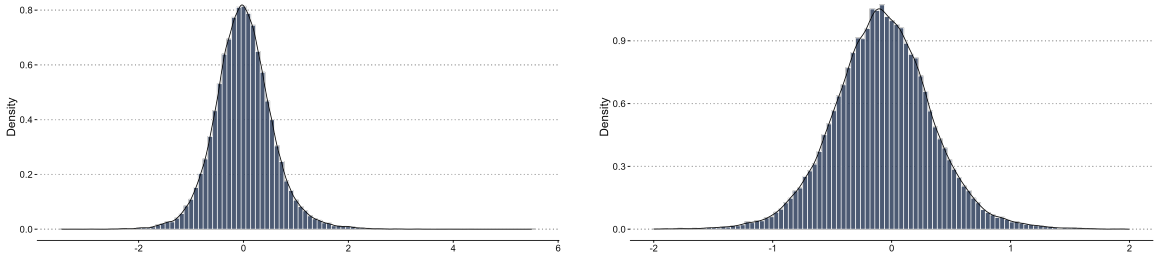
(a) Original Prediction (b) Correction



Notes: The Figure shows the distribution of the errors in our forward imputation. Panel (a) shows the original model and Panel (b) shows the errors after our adjustment.

Figure D.2: Error Distribution - Backward

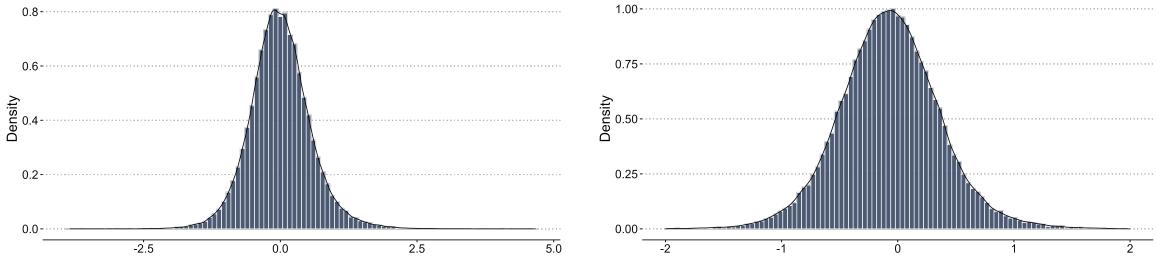
(a) Original Prediction (b) Correction



Notes: The Figure shows the distribution of the errors in our backward imputation. Panel (a) shows the original model and Panel (b) shows the errors after our adjustment.

Figure D.3: Error Distribution - EMOVI

(a) Original Prediction (b) Correction



Notes: The Figure shows the distribution of the errors in our imputation. Panel (a) shows the original model and Panel (b) shows the errors after our adjustment.

E Asset Index

Table E.1: Loadings Vector - PCA

	Weights
Plumbing	0.2351561
Stove	0.2477052
Electricity	0.1850971
TV	0.2245552
Fridge	0.2531215
Washing Machine	0.2564601
Landline	0.2405322
DVD	0.2161941
Microwave	0.2443960
Cable TV	0.2270869
Internet	0.2462148
Cellphone	0.1850880
Computer	0.2360006
Other Housing	0.0095109
Other Land	-0.1393694
Car	0.1560348
Bank Account	0.1917910
Credit Card	0.1879082
Premises	-0.0889106
Working Parcels	-0.0677728
Working Machinery	-0.2309658
Working Animlas	-0.2361236
Livestock	-0.2545471

Notes: The Table shows the loadings vector of the PCA used for constructing the asset index.