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

Understanding household income is crucial for assessing well-being and economic conditions in a region or country (Lenhart, 2019; Yu et al., 2020). This indicator offers insights into a household's resources and reflects the prevailing economic opportunities. Household income is a key metric for measuring and comparing living standards, poverty rates, inequality, and social mobility across diverse groups and geographical areas. However, the absence or unreliability of income data in surveys can pose challenges for accurate analysis.

In response to these challenges, regression-based imputation methods, known as small area estimations (SAE), have been developed (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, 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 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 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 **??** discusses the data and transformations that are done; Section **??** present our main findings followed by robustness checks in Section **??**; Finally Section **??** discusses our results.

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

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

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.

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)

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

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]

   As mentioned before, 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 using any regression-based algorithm, such as linear regression, a penalised version of this, random forest, support vector machines, or even neural networks.

   [COMING BACK AFTER ESTIMATING MODEL]

2. **Evaluating the model**: We evaluate the performance of our model in 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.

[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.

For each cluster or area c,[5] 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.

[COMING BACK AFTER ESTIMATING MODEL]

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.

   [COMING BACK AFTER ESTIMATING MODEL]

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

[5] 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).

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).[6] 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).

[COMING BACK AFTER ESTIMATING MODEL]

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 as shown in Figure **??**.

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 in the prediction as shown in Figure **??**.

However, the within-cluster adjustment does not account for individuals that are miss-classified 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

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

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.[7]

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.

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 both ENIGHs, balancing them with respect to the EMOVI 2017, conducted by the Center for Studies Espinosa Yglesias (CEEY). 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,

[7] 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.

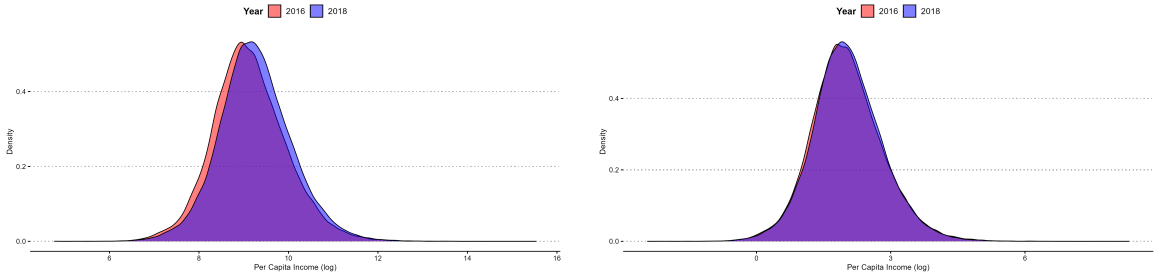
income and spending at the household level. It is conducted biennially, and comparability between waves from 2016 onwards is ensured.

The EMOVI is a survey conducted with the purpose of analysing social mobility. It is representative of men and women at the national level and for five big regions of Mexico and Mexico City, including urban and rural areas. It is conducted every five years by the CEEY and includes information about the respondent's current household and the one where she lived at 14.

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.

2.2 Setting

We create scenarios for potential problems, as stated by Newhouse et al. (2014). 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 1: Summary Statistics

	ENIGH 2016	ENIGH 2018	EMOVI 2017
HHI	10.97	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 is rescaled by the poverty line and divided by the number of individuals in a household to represent real per capita terms.

Table 2: Imputation Scenarios

	Source	Target
Full Data-Set		
Forward	ENIGH 2016	ENIGH 2018
Backward	ENIGH 2018	ENIGH 2016
Pseudo-EMOVI 2017		
Forward	ENIGH 2016	Subset of ENIGH 2018
Backward	ENIGH 2018	Subset of ENIGH 2016

Notes: The Table shows the scenarios we account for. The Full data set refers to the data used for our main results. The Pseudo-EMOVI 2017 refers to the data sets used as robustness checks.

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

4 Robustness Checks

5 Conclusions

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

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

Grajales, R. V., 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/

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

Lenhart, O. (2019). The effects of income on health: new evidence from the Earned Income Tax Credit. *Review of Economics of the Household*, *17*(2), 377–410. https://doi.org/10.1007/S11150-018-9429-X/TABLES/8

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

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

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.

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

Yu, G. B., Lee, D. J., Sirgy, M. J., & Bosnjak, M. (2020). Household Income, Satisfaction with Standard of Living, and Subjective Well-Being. The Moderating Role of Happiness Materialism. *Journal of Happiness Studies*, *21*(8), 2851–2872. https://doi.org/10.1007/S10902-019-00202-X/FIGURES/2