Centro de Estudios®
Espinosa Yglesias
PROMOVEMOS LA IGUALDAD
DE OPORTUNIDADES

A Mirror of Status? Regional and Structural Correlates of Imputed Income in Mexico

Autores:

Pedro J. Torres L.
Department of Social Policy and International Inequalities Institute, London School of Economics and Political Science

Documento de trabajo núm.

09/2025

Centro auspiciado por: ESRU Fundación Espinosa Rugarcía

A Mirror of Status? Regional and Structural Correlates of Imputed Income in Mexico

Pedro J. Torres L.[1]

July 2025

Accurately measuring household income remains a critical yet challenging task in socioeconomic research, particularly in developing contexts like Mexico, where survey data often suffer from incompleteness or reporting bias. In this study, I apply the income imputation technique proposed by Torres *et al*. (2025) to reconstruct household income on the ESRU-EMOVI 2023 survey, using the ENIGH 2022 survey as a donor dataset. By combining education, occupation, and asset indicators, I construct multidimensional socioeconomic status (SES) indices and evaluates their capacity to predict imputed income. The analysis reveals substantial regional heterogeneity: education and wealth are stronger predictors in Mexico's southern regions, while occupational indicators have greater explanatory power in the northern and central areas. Gender differences are relatively modest but indicate slightly higher predictive performance of SES proxies among women. Overall, the findings underscore the necessity of context-specific proxy selection or imputation strategies to better inform policy targeting poverty and inequality.

Keywords: income imputation, socioeconomic status, regional inequality

JEL Classification: D1, D31, I32

- - - - - - - - - - **Documento de Trabajo CEEY núm. 09/2025** - - - - - - - - - -

[1] Department of Social Policy and International Inequalities Institute, London School of Economics and Political Science. p.torres-lopez@lse.ac.uk

1 Introduction

Accurate measurement of household income and consumption is fundamental to evaluating poverty, inequality, and broader socioeconomic dynamics. However, household surveys, particularly in low- and middle-income countries, often struggle to capture these key indicators reliably due to missing data, misreporting, or measurement errors (Deaton 2005; Beegle *et al.* 2012; Jerven 2013). These limitations pose significant challenges to policymakers and researchers aiming to design effective interventions to alleviate poverty and enhance socioeconomic equity (Banerjee and Duflo 2011; Ravallion 2016).

Traditional proxies such as educational attainment, occupational classification, and household asset ownership have long served as indirect measures of economic standing. Education is often treated as a marker of human capital and earning potential (Mincer 1974; Sturgis and Buscha 2015), while occupation reflects labour market segmentation and the institutional structure of economic opportunity (Goldthorpe 1981). Meanwhile, asset indices, pioneered by Filmer and Pritchett (2001), have become a staple in development economics for capturing wealth in the absence of consumption or income data. These indicators provide a multidimensional perspective on socioeconomic status (SES), capturing not only economic resources but also social and cultural capital, which play critical roles in shaping life outcomes (Bourdieu 1986; Coleman 1988).

Among these proxies, asset-based indices, which aggregate household assets, dwelling characteristics, and service access, have become particularly influential in developing-country contexts. Introduced by Filmer and Pritchett (2001), the asset index provides a practical measure of wealth without reliance on expenditure data, which are notoriously difficult and expensive to collect accurately. Similarly, occupation-based indicators, notably those advanced by (Song and Xie 2023), have contributed to refining our understanding of socioeconomic stratification by leveraging the educational and earnings profiles of occupations, thus shedding light on labour market segmentation and inequality dynamics.

Despite their practical advantages, the effectiveness of these proxy measures varies substantially across geographic regions and demographic groups, influenced by local economic structures, cultural contexts, and institutional frameworks. Regional disparities, urban-rural

divides, and gender dynamics significantly shape the validity and explanatory power of proxies, demanding careful consideration and context-specific calibration in analyses.

To enhance the robustness and accuracy of income proxies, recent methodological advancements have emphasized the importance of income imputation techniques, which statistically estimate household income based on observable characteristics. Prominent approaches include small-area estimation methods (Elbers *et al.* 2003; Singleton *et al.* 2020) and synthetic panel techniques (Dang, Lanjouw, *et al.* 2014; Dang and Lanjouw 2023; Roy *et al.* 2023), which draw from multiple datasets to overcome data constraints. Torres *et al.* (2025) further extend these methods by systematically correcting for prediction biases at both the cluster and within the cluster levels, thus preserving the relative income positions of individuals more accurately and improving the reliability of socioeconomic analyses.

In this study I apply the income imputation method proposed by Torres *et al.* (2025) to examine the relationships between imputed household income and key SES indicators: education, occupation, and wealth, in the context of Mexico. Using data from the 2023 ESRU-EMOVI survey, in this research I evaluate regional and gender-based variations in these relationships, constructing multidimensional SES indices through principal component analysis (PCA) to comprehensively understand socioeconomic stratification. Ultimately, in this paper I seek to demonstrate the critical necessity of context-specific selection of proxies and imputation methods, contributing to more precise policy interventions and a deeper understanding of the structural dimensions of inequality.

To do this I leverage the imputation technique proposed by Torres *et al.* (2025). This method involves two distinct stages. Initially, household income is imputed using a log-linear predictive model, constructed from common covariates shared between a source dataset (ENIGH 2022, containing observed income data) and a target dataset (ESRU-EMOVI 2023, lacking direct income measures). To enhance predictive accuracy and reduce systematic bias, the procedure includes cluster-level and within-cluster adjustments through computed ratios that align predicted and observed income distributions. The optimal weighting of these adjustments is determined through cross-validation, specifically prioritizing the minimization of differences in income distribution metrics, such as the Gini coefficient.

In addition, in this study I construct a multidimensional socioeconomic status (SES)

index using principal component analysis (PCA). The SES index integrates three complementary dimensions: educational attainment, occupational status, and asset ownership. Occupational status is uniquely operationalized as a continuous measure derived from the educational composition of occupational groups, allowing the capture of labour market stratification. The final analytical approach involves regression analyses to assess individual and combined associations between these SES components and the imputed household income, enabling a robust and context-sensitive evaluation of the predictors' efficacy.

I identify substantial variability in the efficacy of education, occupation, and wealth as proxies for household income, contingent on regional context and the analytical dimensions considered. Education and Asset-based indices significantly outperforms other measures as a predictor of household income in Mexico's southern regions, likely reflecting a more direct reliance on formal educational credentials in labour markets characterized by higher informality, as well as the acces to certain goods and services. Conversely, occupational indicators demonstrate stronger associations in the northern and central regions, suggesting distinct economic and labour market structures that privilege occupational differentiation.

From a gender perspective, the analysis reveals subtler distinctions, with education, occupation, and wealth proxies exhibiting relatively similar explanatory power across both men and women. However, the composite SES indices constructed through PCA and additive methods indicate slightly stronger predictive performance among women. These results highlight the necessity for interpretation and possibly supplementary qualitative data to thoroughly understand gender-specific socioeconomic dynamics. Overall, the findings underscore the critical importance of selecting income proxies and imputation methodologies in alignment with specific regional characteristics and targeted analytical objectives.

The remainder of the article is structured as follows. Section 2 explains the methodological aspects of the paper. Section 3 describes the main data used for this analysis. Section 4 discusses the findings and Section 5 concludes.

2 Methodology

This section outlines the empirical strategy I use to estimate household income, construct key socioeconomic measures, and analyse their interrelationships. I begin by describing the income imputation procedure, which allows me to recover a distribution of household income in the absence of direct income data. I then detail how I construct a multidimensional index of socioeconomic status (SES) using principal component analysis and measure occupational status. Finally, I explain how I examine the relationship between imputed income and different SES components, assessing their individual and joint contributions through a sequence of regression models and decomposition techniques.

Income Imputation

I adopt the method proposed by Torres *et al.* (2025), which provides a structured approach to income imputation when the goal is to analyse its association with variables not used during the imputation process. Rather than relying solely on regression predictions, this method adjusts for systematic distortions that arise when predicting our variable of interest.

This approach assumes the availability of two complementary datasets. The first, referred to as the source or donor survey (denoted by subscript 1), contains information on income. The second, the target survey (subscript 2), includes variables of interest that are absent from the source survey. Both datasets share a common set of covariates that provide predictive power for income.

In the first step, I reweight the source survey to match the distributional characteristics of the target sample. To do so, I estimate the probability that an observation from the source sample could also appear in the target sample, using a non-parametric reweighting procedure following DiNardo *et al.* (1996). These estimated probabilities are then used to adjust the original sampling weights, which improves the representativeness of the prediction model and the bias correction process.

I model household per capita income using a log-linear specification that includes only the covariates common to both surveys:

$$\log(y_{i1}) = \alpha_1 + X_{i1}\beta_1 + \varepsilon_{i1} \tag{1}$$

To identify prediction errors, I first compute predicted income in the source survey as $\hat{y}_{i1} = \exp(\hat{\alpha}_1 + X_{i1}\hat{\beta}_1)$, and then estimate two adjustment ratios that capture systematic imputation biases:

1. **Cluster-Specific Ratio (CS)**:

$$CS_c = \frac{\mu_{c1}}{\hat{\mu}_{c1}} \tag{2}$$

2. **Cluster-Rank Ratio (CR)**:

$$CR_{cr} = \frac{\mu_{cr1}}{\hat{\mu}_{cr1}} \tag{3}$$

where μ_c and $\hat{\mu_c}$ represent the weighted means of the observed and predicted values, respectively, within clusters c. The first ratio corrects for the mean value of each cluster c, while the second one corrects the mean value at each percentile rank r within each cluster.

Next, I predict income in the target survey as $\hat{y}_{i2} = \exp(\hat{\alpha}_1 + X_{i2}\hat{\beta}_1)$, and apply a convex combination of the two adjustment ratios to obtain the final imputed income values:

$$\tilde{y}_{icr2} = \gamma(CS_c \cdot \hat{y}_{icr2}) + (1-\gamma)(CR_{cr} \cdot \hat{y}_{icr2}) \tag{4}$$

I select the optimal value of $\gamma \in (0, 1)$ through cross-validation, using an 80/20 split of the source sample. The criterion for selection minimizes the difference between the predicted and observed Gini coefficients, thereby prioritizing alignment with the income distribution over traditional prediction accuracy.

To evaluate the robustness of my estimates and construct confidence intervals for the imputed statistics, I implement a non-parametric bootstrap with 200 replications.

For this study, I use the 2022 wave of the ENIGH survey as the source sample and the 2023 wave of the ESRU-EMOVI as the target. Appendix A lists the covariates included in the prediction model. Additional information on the ESRU-EMOVI survey is provided in the following section.

Socioeconomic Status Index

To construct a multidimensional measure of socioeconomic status (SES), I combine three core dimensions: household assets as a proxy for material well-being, educational attainment, and occupational status. This composite index is designed to capture distinct yet complementary aspects of an individual's position within the social and economic hierarchy.

I begin by constructing a standard asset index using Principal Component Analysis (PCA), following the approach of Filmer and Pritchett (2001). This method summarises a set of binary indicators, such as durable goods ownership, housing quality, and access to financial services, into a single ordinal score. I normalize the resulting index between 0 and 1, where higher values indicate greater household wealth. Formally, the index for household h is computed as:

$$w_h = \sum_{f=1}^{F} a_f \left(\frac{x_{fh} - \bar{x}_f}{s_f} \right) \tag{5}$$

where x_{fh} is the observed value of asset f for household h, $\bar{x}_f$ and s_f are the mean and standard deviation of asset f, and a_f is the weight derived from the first principal component. The variables used for its construction are listed in Appendix B.

To extend this to a broader SES measure, I estimate a second PCA model using three variables: years of education, the asset index described in Equation 5, and a continuous measure of occupational status. PCA requires that input variables reflect a common underlying dimension, typically assuming that higher values are positively correlated with higher status. This poses a challenge for occupation, which is often recorded as a categorical variable.

To generate a continuous indicator of occupational status, I follow the methodology proposed by Song and Xie (2023), which constructs a percentile-based occupational ranking from the educational distribution of workers. I adapt this method in two key ways. First, instead of focusing on temporal variation across birth cohorts, I capture *regional* variations using a single cross-sectional dataset. Second, I base the ranking on years of education rather than educational levels. I then calculate the average educational percentile of workers within each occupation-region cell as:

$$S_{ir} = \frac{1}{N_{ir}} \sum_{j \in (i,r)} Q_r(\text{educ}_j) \tag{6}$$

where S_{ir} denotes the occupational status score for occupation i in region r, N_{ir} is the number of individuals in that group, and $Q_r(\text{educ}_j)$ is the regional percentile rank of individual j's years of education.

To assign values of S_{ir}, I estimate the underlying ranks using the 2020 Mexican census and merge them into the analytic dataset using occupation codes from the CMO classification system.

Analysis

To understand how imputed income relates to socioeconomic status (SES), I implement a two-step analytical strategy that focuses on predictive strength, relative contribution, and structural overlap.

In the first step, I estimate a series of simple linear regressions in which each SES component: education, occupation, and assets; is entered individually as a predictor of imputed income. The specification takes the following form:

$$\log(\tilde{y}_{ir}) = \alpha_r + \beta_r X_{ir} + \varepsilon_{ir} \tag{7}$$

where $\tilde{y}_{ir}$ denotes per capita imputed household income for individual i in region r, and X represents the respective SES variable. I omit additional controls in this step to isolate the raw association between each SES dimension and income. I compute the coefficient of determination R^2 for each regression as:

$$R_r^2 = 1 - \frac{\sum_{i=1}^{n}(y_{ir} - \hat{y}_{ir})^2}{\sum_{i=1}^{n}(y_{ir} - \bar{y}_r)^2} \tag{8}$$

where $\hat{y}_{ir}$ denotes the predicted value for individual i, and $\bar{y}_r$ is the mean income in region r. These values provide a comparable metric to assess the explanatory power of each SES dimension across regions.

In the second step, I construct a final SES index using Principal Component Analysis,

following the structure introduced in Equation 5, but now including imputed income as an additional input alongside education, occupation, and assets. This allows me to evaluate how much of the overall SES variance is captured by income once I account for the joint variation shared across all components. By inspecting the loading associated with imputed income in the first principal component, I can assess the extent to which income is structurally embedded in the SES construct.

3 Data

I base this analysis on the 2023 wave of the ESRU-EMOVI Survey, conducted by the Centro de Estudios Espinosa Yglesias (CEEY). The survey is nationally representative of men and women aged 25 to 65 and uses a stratified design that covers five macro-regions as well as urban and rural areas.

The full sample comprises 17,833 individuals, of whom 17,541 report valid information on years of education. I restrict the analytical sample to respondents with complete education data. Individuals who do not report an occupation are assigned the category "Not employed."

Table 1 presents descriptive statistics for the national sample. The average respondent reports 13.1 years of education, suggesting that most individuals have completed upper secondary education. The sample is demographically balanced, with an average age of 42 years and a near-even gender distribution (53% women). Approximately 20% of respondents reside in rural areas, a figure that broadly aligns with national urbanization rates. The asset index has a mean of 0.58 (SD = 0.17), capturing variation in household wealth based on durable goods, housing quality, and services. Finally, the average imputed per capita household income is 7,252 pesos per quarter, or roughly 2,417 pesos per month.

Table 2 displays the regional distribution of observations and their weighted representation in the national population. While the number of survey respondents is relatively even across regions (approximately 3,500 observations each), the population weights differ substantially. The Central region accounts for the largest share, 38.4% of the weighted sample, reflecting the demographic and economic significance of Mexico City and its surrounding metropolitan area.

Table 1: Descriptive Statistics

	Mean	**SD**
HHI (Quarterly, MXN)	7,252	11,478
Log HHI	8.52	0.80
Years of Education	13.10	4.70
Age	42.00	12.00
Women	0.53	0.50
Rural	0.20	0.40
Assets Index	0.58	0.17

Note: This table reports descriptive statistics from the ESRU-EMOVI 2023 survey. The values represent weighted means and corresponding standard errors, calculated using survey sampling weights.

Table 2: Observations per Region

	Observations	**Individuals**	**Percentage**
National	17,551	58,993,122	100.0
Central	3,502	22,629,158	38.4
North	3,617	11,253,000	19.1
North-Central	3,480	8,137,407	13.8
Northwest	3,565	3,935,749	6.7
South	3,387	13,037,808	22.1

Note: This table reports the number of observations by region from the ESRU-EMOVI 2023 survey. It includes the unweighted counts, weighted counts, and corresponding percentages, with the latter two calculated using survey sampling weights.

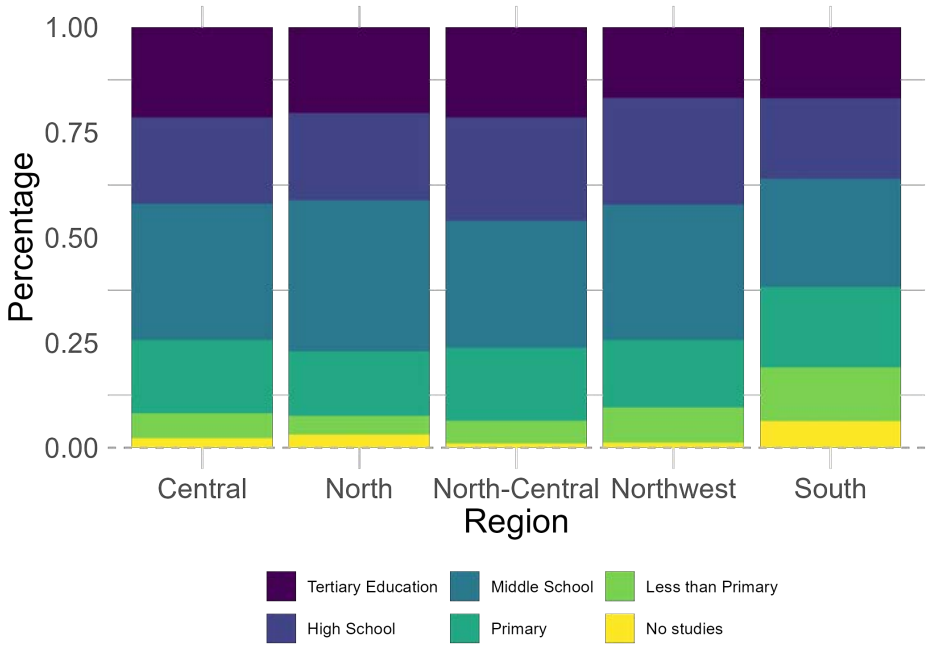
To explore regional disparities in socioeconomic conditions, I begin by examining educational attainment. Figure 1 shows clear spatial variation. The South registers the highest concentration of individuals with little or no formal education, while the Central and North-Central regions have higher shares of respondents with secondary and tertiary education.

Substantial differences also emerge in occupational structure. Figure 2 displays the distribution of occupational categories across regions. Agricultural and non-employed individuals are more prevalent in the South, whereas professional, technical, and industrial workers are more concentrated in the Central and Northern regions.

Figure 3 illustrates the distribution of the asset index across regions. The Central and North-Central regions exhibit higher and more concentrated levels of asset accumulation. In contrast, the South displays a flatter and left-skewed distribution, consistent with lower average household wealth and greater within-region inequality.

Finally, I examine the distribution of imputed household income. Figure 4 shows kernel

Figure 1: Education Level by Region



Note: This figure presents the educational composition across different regions of Mexico, based on data from the ESRU-EMOVI 2023 survey. Percentages are calculated using survey sampling weights to ensure representativeness.

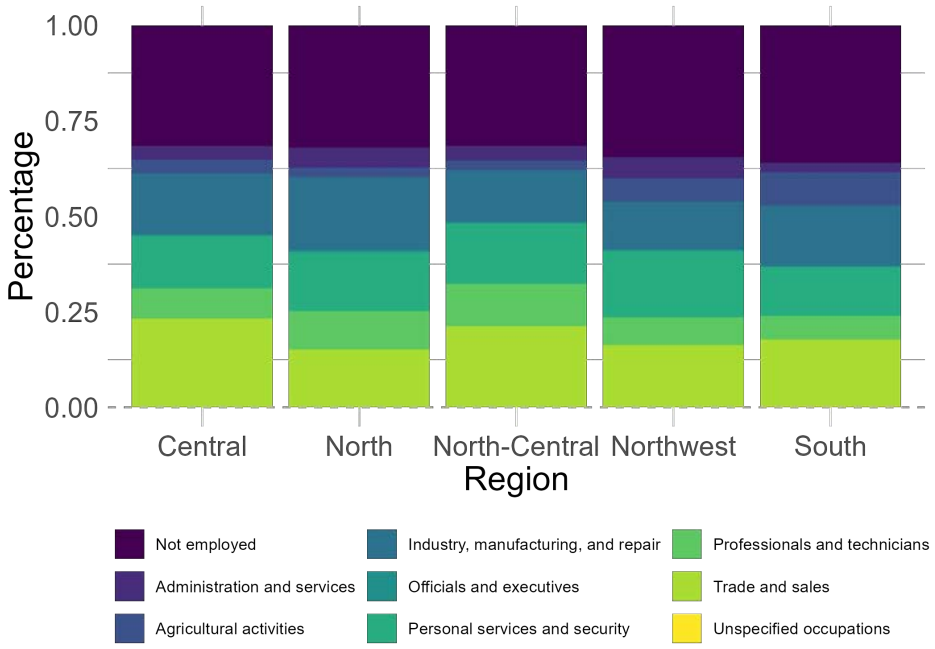
density estimates of log-transformed per capita income. The Central and North-Central regions register higher average incomes and narrower dispersions, indicative of more affluent and internally homogeneous populations. The South, by contrast, shows both lower average income and a wider spread, mirroring patterns observed in asset accumulation.

Together, these findings point to persistent and substantial regional inequalities in Mexico. Differences in educational attainment, occupational structure, and material wealth all contribute to variation in income outcomes. In the next section, I examine the degree to which each of these socioeconomic dimensions correlates with imputed household income across the country.

4 Results

Table 3 presents findings from a series of bivariate regressions where imputed household income is individually regressed on distinct dimensions of socioeconomic status (SES). The

Figure 2: Occupation by Region



Note: This figure presents the occupational composition across different regions of Mexico, based on data from the ESRU-EMOVI 2023 survey. Percentages are calculated using survey sampling weights to ensure representativeness.

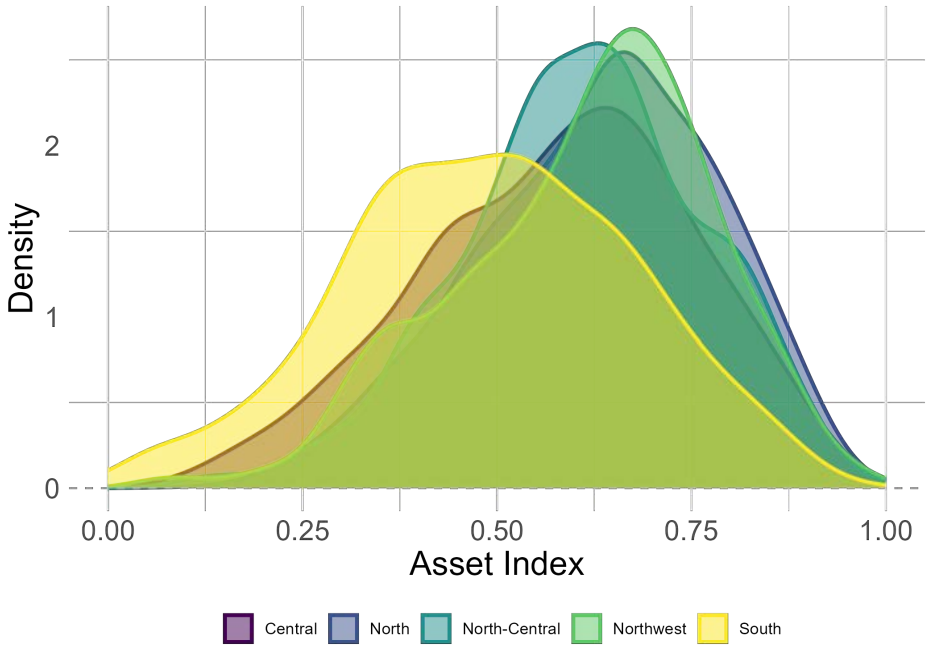
explanatory power, measured through the coefficient of determination (R^2), allows for a direct comparison of each variable's contribution to household income variance.

Among SES dimensions, education emerges as the strongest predictor of household income. When operationalized categorically, education explains 36.5% of the income variance, whereas the continuous measure based on years of schooling demonstrates lower explanatory power (28.4%). This divergence underscores the importance of discrete educational milestones, such as the completion of secondary or tertiary education, over the incremental accumulation of schooling years.

Occupational indicators exhibit comparatively modest associations. Broad occupational categories account for 22.1% of the variance, outperforming both class-based classifications (16.6%) and continuous occupation scores derived from educational composition. The notably low predictive power (12.9%) of occupation scores based on class suggests that finer classifications may be more informative for household income analyses.

Asset-based indicators nearly match education in predictive capability. The PCA-derived

Figure 3: Asset Index by Region



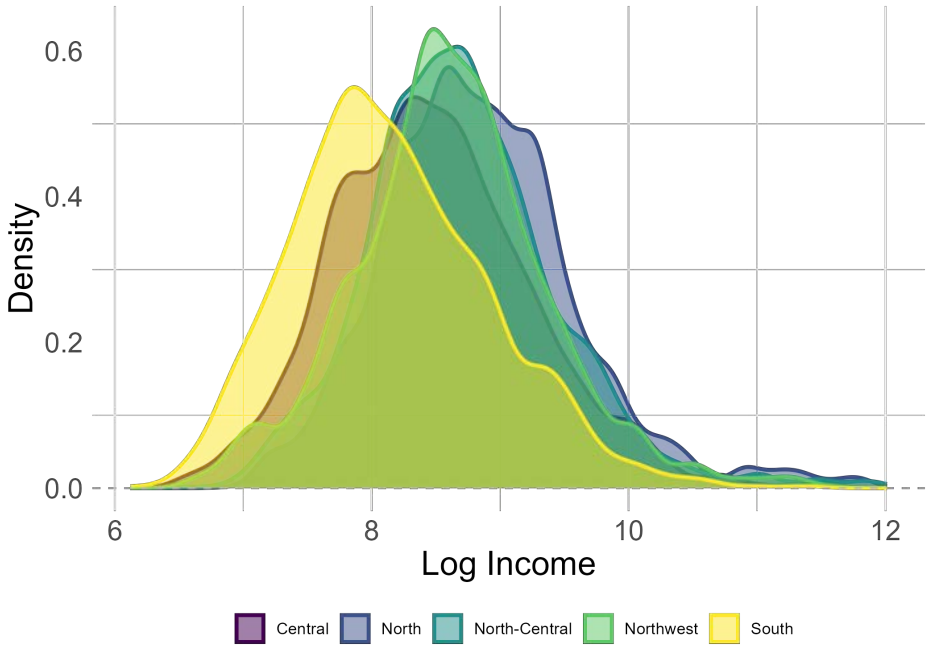
Note: This figure presents the asset index distribution across different regions of Mexico, based on data from the ESRU-EMOVI 2023 survey. Percentages are calculated using survey sampling weights to ensure representativeness.

asset index attains an R^2 of 0.340, closely paralleled by the percentile-based wealth ranking (0.331). These findings affirm the utility of tangible wealth measures, housing quality and durable goods ownership, as robust proxies for household economic standing, especially when direct income measures are unavailable.

When all SES dimensions are jointly included, the model's explanatory power significantly increases to $R^2 = 0.515$. This finding indicates that education, occupation, and assets capture distinct, complementary facets of socioeconomic advantage. Synthetic SES indices created via PCA exhibit notable explanatory power, particularly when based on occupational categories (0.391) rather than broader class-based measures (0.380). Despite some loss in explanatory power due to dimensionality reduction, these composite indices provide concise and interpretable proxies for SES.

Subsequent sections explore the heterogeneity of these associations across macro-regions, residential contexts, and gender, providing nuanced insights into the structural dimensions of socioeconomic inequalities.

Figure 4: Imputed Household Income by Region



Note: This figure presents the imputed per capita household income distribution across different regions of Mexico, based on data from the ESRU-EMOVI 2023 survey. Percentages are calculated using survey sampling weights to ensure representativeness.

Education

Figure 5 illustrates regional variation in the explanatory power of education. Panel 5a considers formal educational levels, while Panel 5b employs years of schooling.

The strength of the education-household income relationship varies markedly by region. Educational levels explain over 40% of income variation in the South, highlighting the pronounced signalling effect of formal credentials within this labour market context. Conversely, associations weaken in the Central and Northern regions (around 30%), suggesting greater diversification in income pathways or differential returns to education.

While the explanatory power of years of schooling is consistently lower across regions, the ordering remains similar, reinforcing that discrete educational milestones hold greater importance for household income determination than incremental educational progression, which may be stronger associated to individual income and earnings.

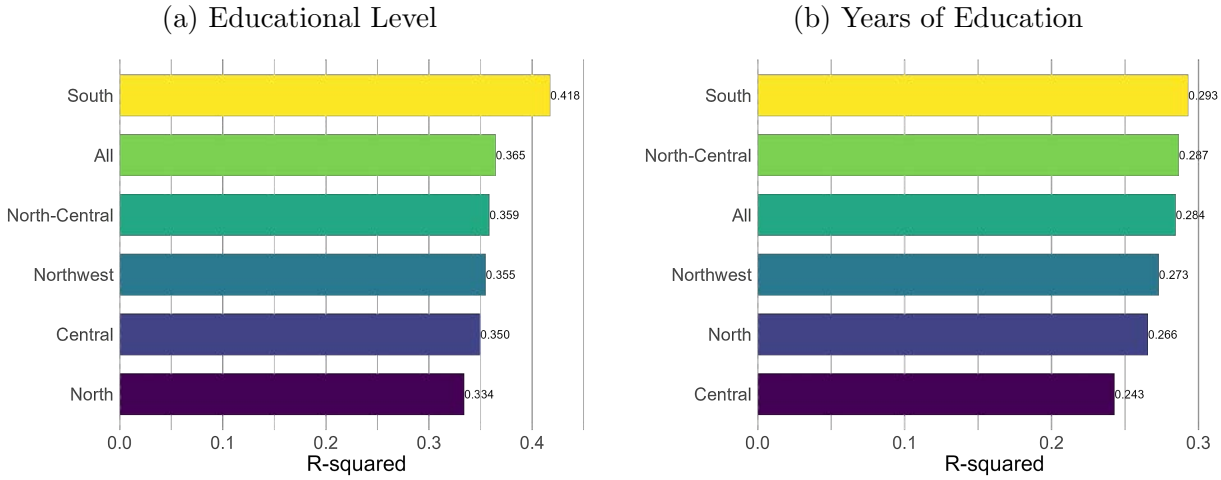
Figure 6 reveals urban–rural disparities. Educational levels demonstrate higher predictive power in urban areas. This gap widens further when examining years of education. This

Table 3: Associations between Key SES Variables and Imputed Income

	R^2
Education	
Level	0.365
Years	0.284
Occupation	
Categories	0.221
Class	0.166
Score (Categories)	0.216
Score (Class)	0.129
Assets	
Index	0.340
Rank	0.331
SES	
Independent	0.515
PCA (Categories)	0.391
PCA (Class)	0.380

Note: This table reports the coefficient of determination (R^2) from a series of simple regressions of imputed household income on individual socioeconomic status (SES) components at the national level, using data from the ESRU-EMOVI 2023 survey.

Figure 5: Education and Imputed Income across Regions

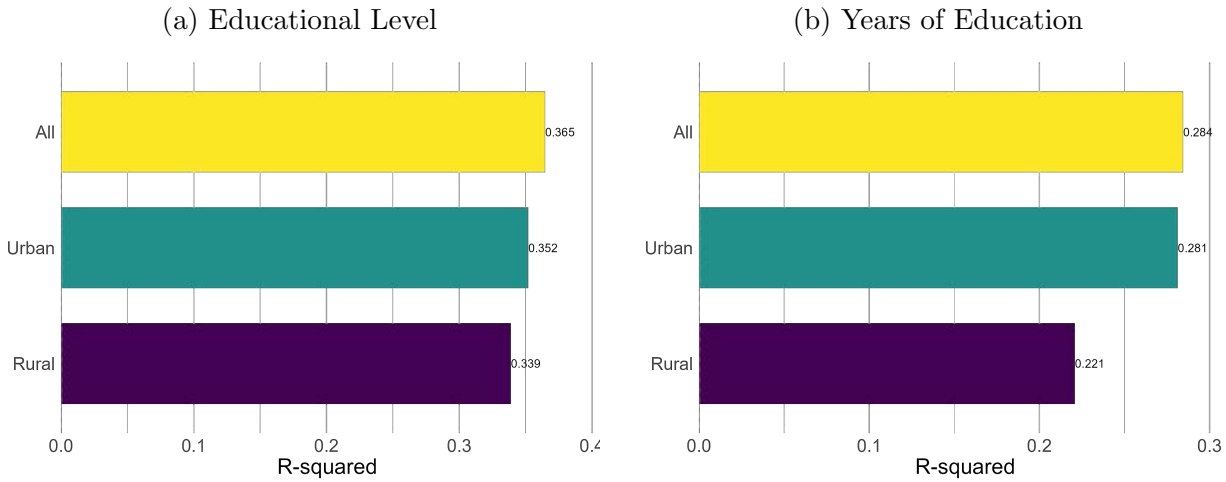


Note: This figure presents the coefficient of determination (R^2) for education across regions in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using categorical education levels, while Panel (b) uses years of education as a continuous measure.

likely reflects that for urban areas, the household heads characteristics may be a stronger signalling in the labour market than it is in rural areas.

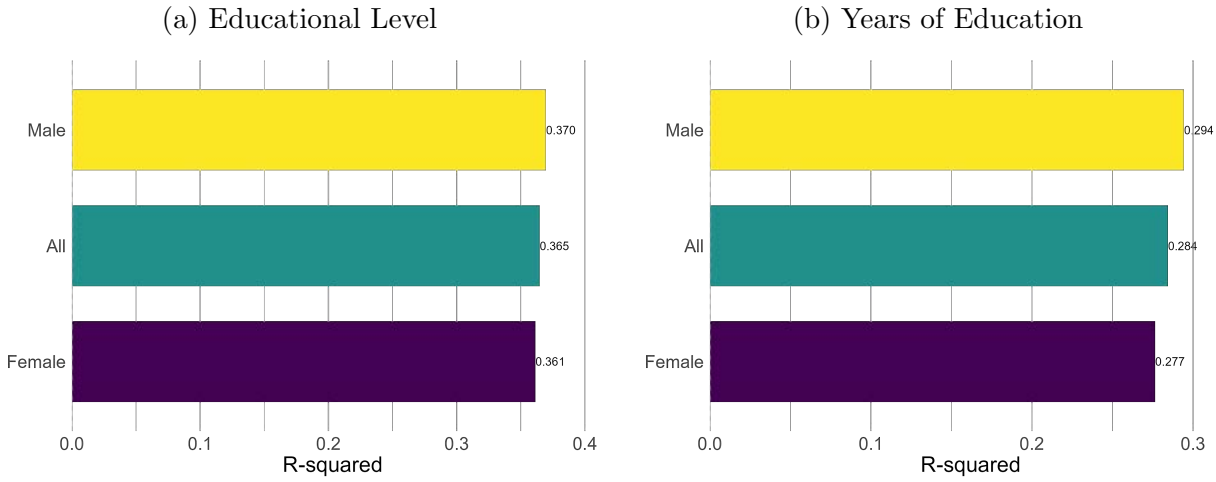
Gender-based analyses (Figure 7) uncover subtle yet significant differences. Formal edu-

Figure 6: Education and Imputed Income across Areas



Note: This figure presents the coefficient of determination (R^2) for education across areas in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using categorical education levels, while Panel (b) uses years of education as a continuous measure.

Figure 7: Education and Imputed Income across Gender



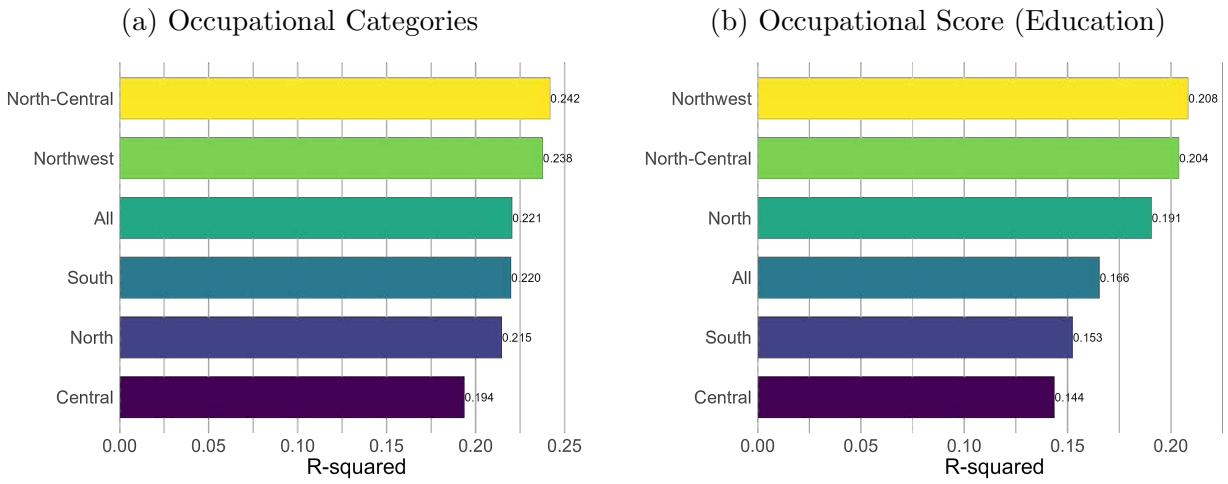
Note: This figure presents the coefficient of determination (R^2) for education across gender in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using categorical education levels, while Panel (b) uses years of education as a continuous measure.

cational levels equally predict household income for men and women; however, the continuous measure (years of education) reveals a steeper decline in predictive power for women. This pattern indicates potential gendered disparities in translating incremental education into household income. However, the association still remains very close.

Occupation

Figure 8 explores regional variations in occupational predictors. Broad occupational categories display the highest explanatory power in the North-Central and Northwest regions (around 24% and 23%, respectively), suggesting pronounced occupational segmentation. Conversely, continuous occupation scores exhibit weaker associations in Central and Southern regions, reflecting greater occupational heterogeneity.

Figure 8: Occupation and Imputed Income across Regions

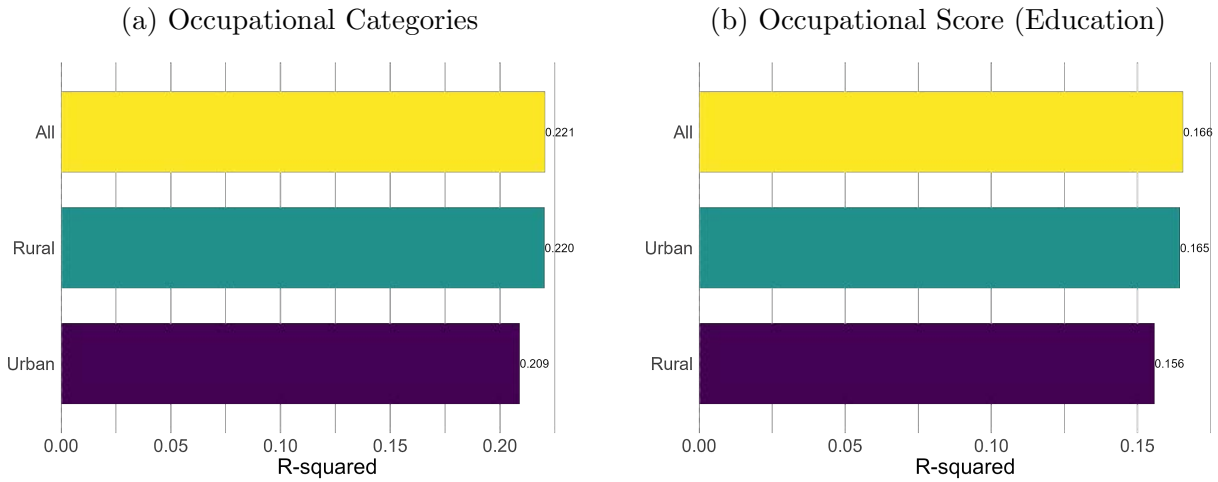


Note: This figure presents the coefficient of determination (R^2) for occupation across regions in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using categorical occupations, while Panel (b) uses a continuos score.

Urban-rural comparisons (Figure 9) reveal stronger predictive power of occupational variables in rural areas when taking the borad categories, indicating narrower occupational structures and greater dependence on occupational identities for rural incomes. When considering the score based on education, urban areas show stronger predictive power, probably reflecting the association of education and household income.

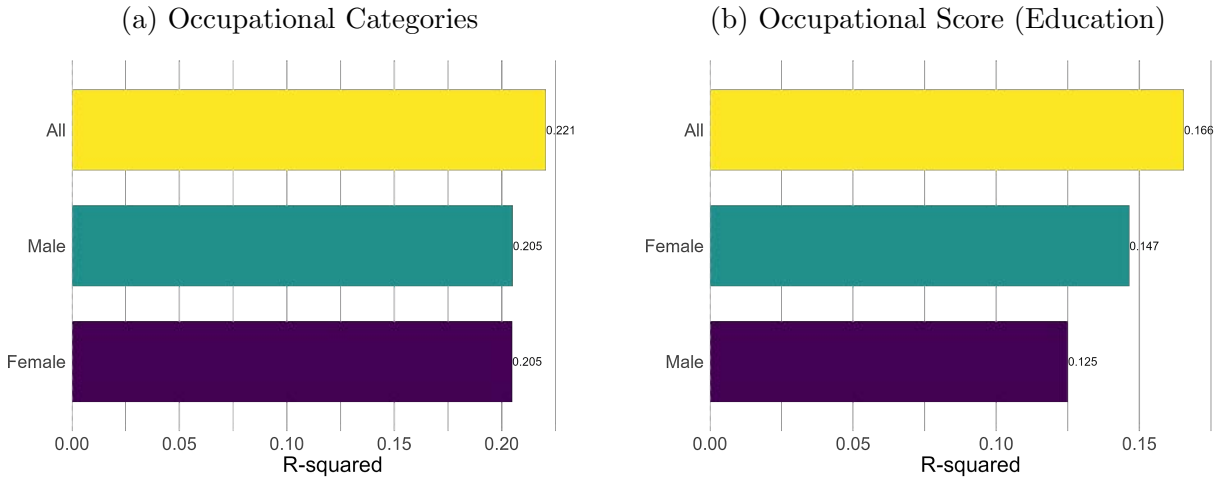
Gender comparisons in occupational predictors (Figure 10) show parity for categorical measures, yet a gender gap emerges with the continuous occupation score, performing better for women. This suggests greater alignment between educationally ranked occupations and household income for women, possibly due to restricted occupational opportunities or credential-based hiring practices.

Figure 9: Occupation and Imputed Income across Areas



Note: This figure presents the coefficient of determination (R^2) for occupation across areas in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using categorical occupations, while Panel (b) uses a continuos score.

Figure 10: Occupation and Imputed Income across Gender



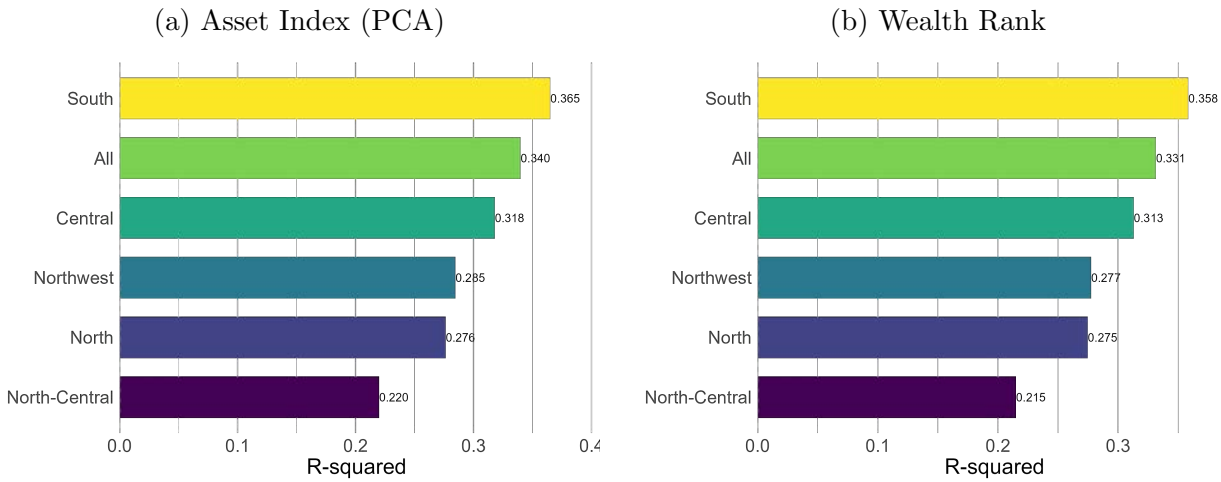
Note: This figure presents the coefficient of determination (R^2) for occupation across gender in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using categorical occupations, while Panel (b) uses a continuos score.

Wealth

Wealth indicators, analysed regionally in Figure 11, reveal the strongest predictive associations in the South (R^2 around 0.36). This implies heightened reliance on tangible assets as proxies for economic status in regions characterized by informal labour markets or economic

precarity. In contrast, the North-Central and Northern regions exhibit lower associations (below 0.30), signalling that household income in these areas depends more on formal employment and occupational returns.

Figure 11: Wealth and Imputed Income across Regions



Note: This figure presents the coefficient of determination (R^2) for the asset index across regions in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using the normalized value, while Panel (b) uses a percentile rank of the index.

Urban–rural contrasts (Figure 12) consistently highlight stronger asset–income associations in rural contexts. This pattern suggests that in less monetized rural economies, physical assets more directly reflect economic status compared to diversified urban income sources.

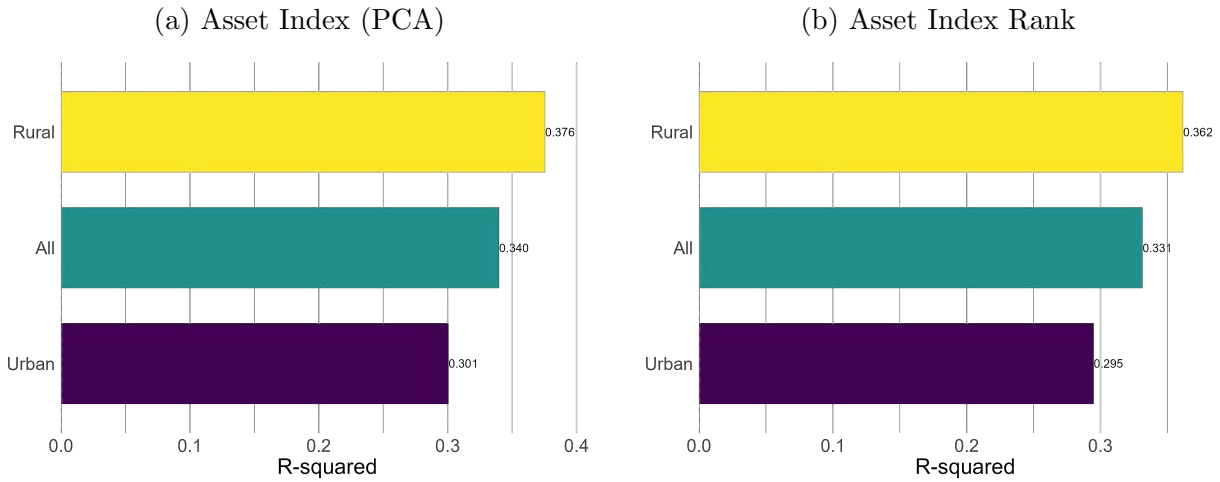
Gender analyses (Figure 13) display slightly higher predictive power for women across asset-based indicators, though differences remain modest. This subtle gender difference likely reflects household-level measurement, where asset ownership and household income are inherently shared across genders, obscuring individual-specific asset–income relationships.

Composite SES Indices

Figure 14 compares composite SES indices' explanatory power across macro-regions. Panel 14a shows results from a PCA-based SES index, whereas Panel 14b presents findings from an additive specification including years of education, occupational score, and asset index.

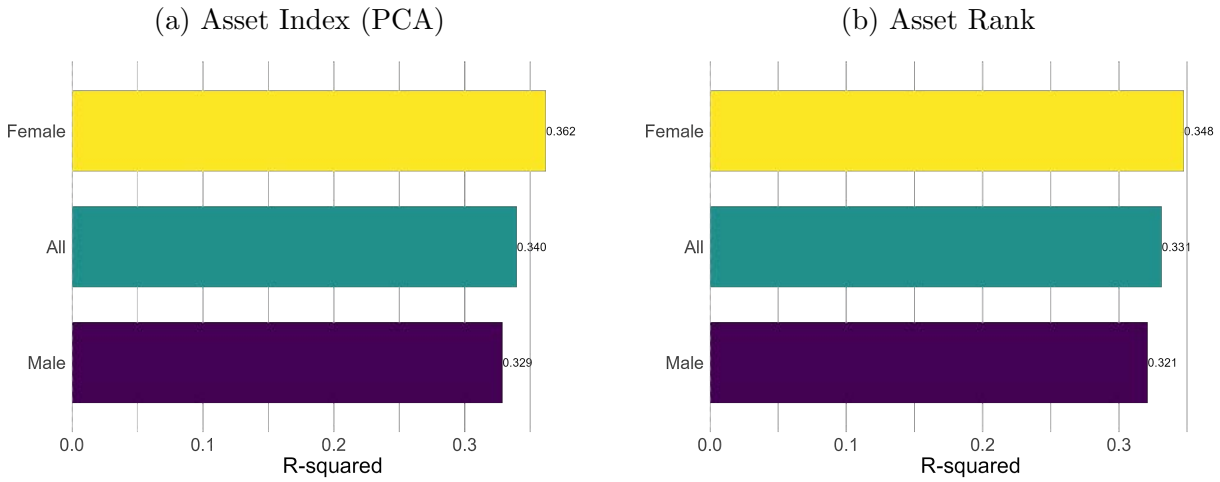
Both panels illustrate a clear regional hierarchy, with the South showing the highest explanatory power: approximately 0.41 in the PCA model and nearly 0.54 in the additive

Figure 12: Wealth and Imputed Income across Areas



Note:. This figure presents the coefficient of determination (R^2) for the asset index across areas in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using the normalized value, while Panel (b) uses a percentile rank of the index.par

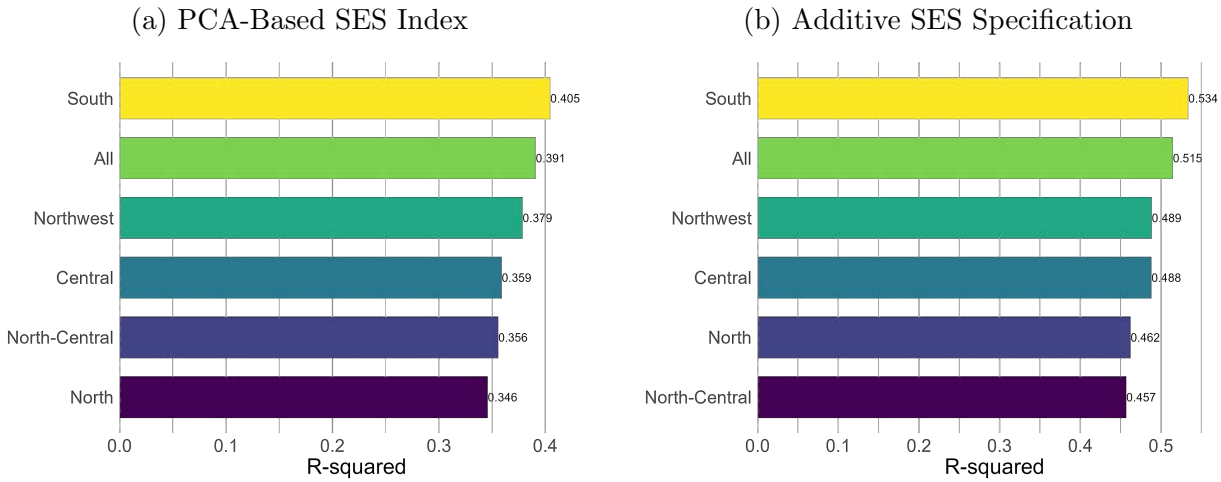
Figure 13: Assets and Imputed Income by Gender



Note: This figure presents the coefficient of determination (R^2) for the asset index across gender in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using the normalized value, while Panel (b) uses a percentile rank of the index.

model. The strong relationship in the South is probably due to educations influence over SES, as can be seen in Tables 4 and 5. The additive specification consistently outperforms the PCA index across all regions, particularly in the North-Central region, highlighting that disaggregating SES components yields greater predictive power.

Figure 14: SES Indices and Imputed Income across Regions



Note: This figure presents the coefficient of determination (R^2) for socioeconomic status across regions in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using the a PCA based index, while Panel (b) uses a each component separately as a regressor.

The PCA-based SES index in Panel 15a shows similar associations with household income across areas, likely reflecting the diminished variability introduced by PCA. Conversely, Panel 15b, employing an additive approach, exhibits notably stronger explanatory power in rural areas (above 0.5) compared to urban areas (below 0.5). This urban-rural gap underscores distinct component contributions within different residential contexts.

Both PCA-based and additive SES indices demonstrate higher explanatory power among women than men. Specifically, the PCA-based index explains approximately 0.39 of the income variance for women versus 0.34 for men. The additive model further amplifies this distinction, explaining nearly 0.52 of income variance among women compared to just below 0.47 for men. These patterns indicate that SES dimensions collectively offer greater predictive value for women's household income, highlighting potential gender-specific socioeconomic dynamics.

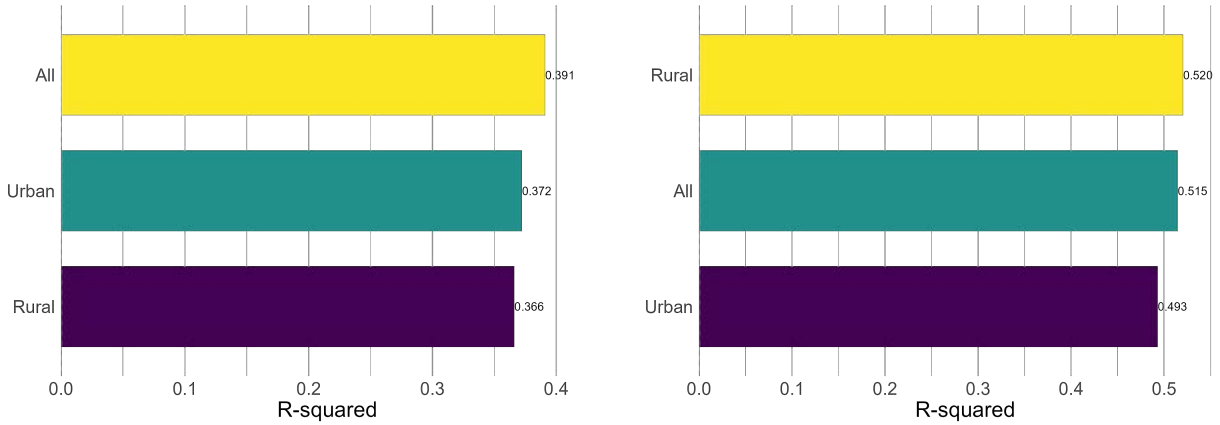
To assess the relative weights of SES dimensions within composite indices, Tables 4 and 5 present loading vectors from the first principal component (PCA) extracted under different specifications.

Table 4 shows loadings for education, occupation, and wealth without explicit income inclusion. All variables are standardized, ensuring that weights reflect each component's

Figure 15: SES and Imputed Income across Areas

(a) PCA-Based SES Index

(b) Additive SES Index (Education, Occupation, Wealth)

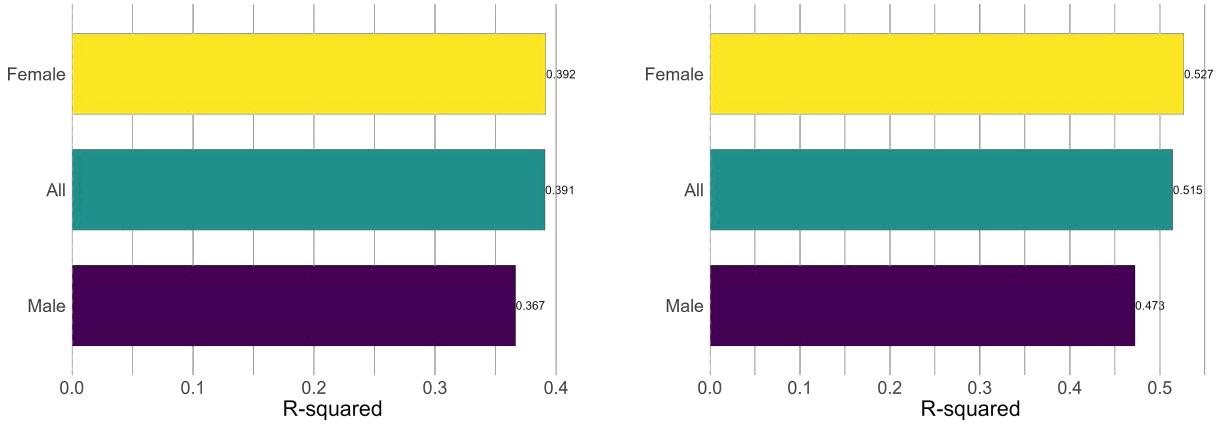


Note: This figure presents the coefficient of determination (R^2) for socioeconomic status across areas in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using the a PCA based index, while Panel (b) uses a each component separately as a regressor.

Figure 16: SES and Imputed Income by Gender

(a) PCA-based SES Index

(b) Education, Occupation, and Wealth



Note: This figure presents the coefficient of determination (R^2) for socioeconomic status across gender in Mexico, based on data from the ESRU-EMOVI 2023 survey. Panel (a) reports R^2 values using the a PCA based index, while Panel (b) uses a each component separately as a regressor.

independent association with SES.

Education carries the highest loading (0.64), confirming its primary role in shaping socioeconomic stratification, even after controlling for correlations with other dimensions. Occupation (0.45) and wealth (0.47) have closely aligned loadings, indicating their complementary

Table 4: Loading Vectors of SES Components

	Weight	Mean	Std. Dev.
Education	0.64	13.06	4.07
Occupation	0.45	32.75	26.24
Wealth	0.47	0.57	0.17

Note: This table presents the factor loadings (weights) of the variables used to construct the SES index via principal component analysis (PCA), based on data from the ESRU-EMOVI 2023 survey, calculated using survey sampling weights.

and nearly equivalent informational value regarding socioeconomic positioning.

Table 5 incorporates log household per capita income into the PCA, examining how explicit inclusion of income adjusts component weights.

Table 5: PCA Loadings with Log Income

	Weight	Mean	Std. Dev.
Education	0.52	13.06	4.07
Occupation	0.38	32.75	26.24
Wealth	0.48	0.57	0.17
Log Income	0.57	8.52	0.80

Note: This table presents the factor loadings (weights) as well as the mean and standard deviation of the variables used to construct the SES index via principal component analysis (PCA) when including imputed income, based on data from the ESRU-EMOVI 2023 survey, calculated using survey sampling weights.

Including log income significantly influences component loadings, highlighting its prominent role (0.57). Education remains important (0.52), though its weight decreases probably due to overlap with income. Occupation's loading is reduced further to 0.38, reflecting a higher correlation with income. Wealth maintains stability (0.48), emphasizing its unique contribution to capturing structural dimensions of SES not fully represented by income, education or occupation.

5 Conclusions

In this analysis I underscore the critical importance of selecting income proxies thoughtfully, contingent on the specific analytical objectives and dimensions being examined. The findings clearly illustrate that no single socioeconomic variable, education, occupation, or wealth, consistently serves as the optimal proxy for household income across all contexts. Rather,

the suitability of these proxies varies substantially based on whether the analysis is regionally or gender-focused.

At the regional level, education emerges as a notably powerful predictor of imputed income, particularly in the Southern regions of Mexico, where formal educational attainment significantly stratifies income levels. Conversely, in the Northern and Central regions, occupational categories and wealth-based measures demonstrate relatively higher explanatory power. These regional disparities likely reflect differing local economic structures and labor market dynamics. In areas characterized by informal or fragmented labor markets, such as the South, formal education may act as a stronger income signal. In contrast, regions with more formalized economies exhibit stronger associations between household income and occupational.

Gender-focused analyses, however, reveal subtler differences. Education, occupation, and wealth measures exhibit relatively consistent explanatory power between men and women, suggesting that gender-based socioeconomic disparities may not be optimally captured by simply switching proxies. Instead, nuanced interpretations and complementary data, such as detailed employment status or intra-household resource distribution, might enhance the understanding of gender dynamics within households.

Therefore, analysts and policymakers must carefully select proxies or adopt imputation techniques based on the critical understanding of their analytical focus. When regional disparities are central, proxy selection must reflect the distinctive economic and educational contexts of each region. For gender-focused research, supplementing proxy data with richer individual-level or qualitative indicators might offer more insightful conclusions. Ultimately, embracing such context-specific and flexible approaches enhances both the accuracy and the interpretability of socioeconomic assessments.

References

Banerjee, A. V. and E. Duflo (2011). *Poor Economics - A radical rethinking of the way to fight global poverty*. Vol. 22, pp. 356–358. ISBN: 1610390938. DOI: 10.3362/1755-1986.2011.037.

Beegle, K., J. De Weerdt, J. Friedman, and J. Gibson (2012). "Methods of household consumption measurement through surveys: Experimental results from Tanzania". In: *Journal of Development Economics* 98.1. Symposium on Measurement and Survey Design, pp. 3–18. ISSN: 0304-3878. DOI: https://doi.org/10.1016/j.jdeveco.2011.11.001. URL: https://www.sciencedirect.com/science/article/pii/S0304387811001039.

Bourdieu, P. (1986). "The Forms of Capital". In: *Handbook of Theory and Research for the Sociology of Education*, pp. 241–258.

Coleman, J. S. (1988). "Social Capital in the Creation of Human Capital". In: *The American Journal of Sociology* 94.

Dang, H. A. H. and P. Lanjouw (2023). "Regression-based imputation for poverty measurement in data-scarce settings". In: *Research Handbook on Measuring Poverty and Deprivation*. Edward Elgar Publishing Ltd., pp. 141–150. ISBN: 9781800883451. DOI: 10.4337/9781800883451.00023.

Dang, H. A. H., P. Lanjouw, J. Luoto, and D. McKenzie (2014). "Using repeated cross-sections to explore movements into and out of poverty". In: *Journal of Development Economics* 107, pp. 112–128. ISSN: 0304-3878. DOI: 10.1016/J.JDEVECO.2013.10.008.

Deaton, A. (2005). "Measuring Poverty In a Growing World (Or Measuring Growth In a Poor World)". In: *The Review of Economics and Statistics* LXXXVII.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach". In: *Econometrica* 64.5, pp. 1001–1044. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2171954 (visited on 03/13/2024).

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). "Micro–Level Estimation of Poverty and Inequality". In: *Econometrica* 71.1, pp. 355–364. DOI: https://doi.org/10.1111/1468-0262.00399. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-

0262.00399. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00399.

Filmer, D. and L. H. Pritchett (2001). "Estimating Wealth Effects without Expenditure Data-or Tears: An Application to Educational Enrollments in States of India". In: *Demography* 38.1, pp. 115–132. ISSN: 00703370, 15337790.

Goldthorpe, J. H. (1981). "The Class Schema of 'Social Mobility and Class Structure in Modern Britain': A Reply to Penn". In: *Sociology* 15.2, pp. 272–280. DOI: 10.1177/003803858101500209.

Jerven, M. (2013). *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*. Cornell University Press. ISBN: 9780801451638.

Mincer, J. A. (1974). "The Human Capital Earnings Function". In: *Schooling, Experience, and Earnings* I, pp. 83–96. URL: http://www.nber.org/books/minc74-1.

Ravallion, M. (2016). "Toward better global poverty measures". In: *Journal of Economic Inequality* 14 (2), pp. 227–248. ISSN: 15738701. DOI: 10.1007/S10888-016-9323-9/METRICS. URL: https://link.springer.com/article/10.1007/s10888-016-9323-9.

Roy, S. S., R. van der Weide, P. Choudhri, A. Dar, P. Ghosh, P. Gopalakrishnan, K. Kishore, discussants Santanu Pramanik, A. Malani, P. Sen, M. Vyas, S. Desai, P. Gupta, J. V. Meenakshi, P. Mukhopadhyay, R. Somanathan, J. K. Ahmad, Z. Allaoua, S. Bhalla, V. A. Nageswaran, A. Dabalen, I. Gill, K. Himelein, J. Hoogeveen, D. M. Jolliffe, A. Kraay, N. Krishnan, C. Lakner, A. Narayan, O. S. S. J. Ng, P. Olinto, N. Sarma, B. Ozler, C. Reinhart, B. Rijkers, P. A. C. Rodas, C. Sanchez-Paramo, H. Timmer, T. Vishwanath, and N. Yoshida (2023). "Poverty in India Has Declined over the Last Decade But Not As Much As Previously Thought *". In: URL: https://pib.gov.in/Pressreleaseshare.aspx?PRID=1591792..

Singleton, A., A. Alexiou, and R. Savani (2020). "Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation". In: *Computers, Environment and Urban Systems* 82, p. 101486. ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2020.101486.

Song, X. and Y. Xie (2023). "Occupational Percentile Rank: A New Method for Constructing a Socioeconomic Index of Occupational Status". In: *Sociological Methods and Research*. ISSN: 15528294. DOI: 10.1177/00491241231207914.

Sturgis, P. and F. Buscha (2015). "Increasing inter-generational social mobility: is educational expansion the answer?" In: *The British Journal of Sociology* 66 (3), pp. 512–533. ISSN: 1468-4446. DOI: 10.1111/1468-4446.12138.

Torres, P. J., L. A. Monroy-Gómez-Franco, and R. Vélez-Grajales (2025). "Survey to Survey Imputation when External Covariates Matter: Estimating Inequality of Opportunity in Mexico". In: *Centro de Estudios Espinosa Yglesias Working Papers*.

A Shared Covariates

Table A.1: Shared covariates between ENIGH 2022 and ESRU-EMOVI 2023

	ENIGH 2022	**EMOVI 2023**
Landline	0.35 (0.49)	0.38 (0.47)
Cellphone	0.95 (0.23)	0.92 (0.27)
Paid TV	0.43 (0.49)	0.52 (0.50)
Internet	0.63 (0.49)	0.70 (0.45)
Water Pipes	0.77 (0.43)	0.92 (0.25)
Electricity	0.99 (0.12)	0.99 (0.10)
Car	0.48 (0.50)	0.51 (0.50)
Property	0.69 (0.45)	0.75 (0.43)
Mood Floor	0.03 (0.16)	0.02 (0.13)
Cement Floor	0.46 (0.50)	0.55 (0.50)
Other Floor	0.52 (0.50)	0.43 (0.50)
Sex (HH)	0.69 (0.46)	0.52 (0.50)
Age (HH)	49.7 (13.99)	41.91 (12.02)
# Men	0.49 (0.24)	0.49 (0.26)
# Occupied	0.54 (0.29)	0.48 (0.3)
Speaks Indigenous	0.07 (0.28)	0.07 (0.22)
Major Adults	0 (0)	0.15 (0.31)
Rural	0.23 (0.48)	0.20 (0.40)
No education (HH)	0.05 (0.22)	0.03 (0.14)
Less than Primary Education (HH)	0.12 (0.34)	0 (0.02)
Primary Education (HH)	0.18 (0.4)	0.2 (0.37)
Secondary Education (HH)	0.31 (0.47)	0.30 (0.45)
High-School (HH)	0.18 (0.37)	0.22 (0.46)
University (HH)	0.14 (0.32)	0.19 (0.39)
Postgraduate (HH)	0.03 (0.14)	0.01 (0.11)

Note: This table presents the covariates shared between the ENIGH 2022 and ESRU-EMOVI 2023 surveys. Values represent weighted means, with standard deviations in parentheses. All statistics are computed using survey sampling weights.

B Shared Covariates

Table B.1: Assets and Services used in PCA

	Weight	**Mean**	**Std. Dev.**
Plumbing	0.24	0.92	0.25
Stove	0.26	0.94	0.21
Electricity	0.14	0.99	0.10
Tv	0.25	0.86	0.34
Fridge	0.27	0.93	0.23
Washing Machine	0.26	0.81	0.37
Microwave	0.26	0.54	0.49
Cable Tv	0.26	0.52	0.50
Internet	0.28	0.70	0.45
Cellphone	0.24	0.92	0.27
Computer	0.21	0.34	0.47
Videogames	0.19	0.18	0.38
Other Land	-0.05	0.07	0.24
Automobile	0.09	0.10	0.29
Bank Account	0.17	0.27	0.46
Credit Card	0.19	0.20	0.41
Insurance	0.14	0.12	0.33
Premises	-0.20	0.11	0.31
Farming Machinery	-0.21	0.04	0.18
Working Animals	-0.29	0.07	0.22

Note: This table presents the assets and services used to construct the Asset Index usinf principal component analysis (PCA) based on the ESRU-EMOVI 2023 surveys. Values represent the weights from the PCA, weighted means and standard deviations. All statistics are computed using survey sampling weights.