Identifying Hand-to-Mouth Households: Evidence from India

Vivek Gupta
Shiv Nadar Institution of Eminence

Fiorella Pizzolon
Hamilton College

Aditi Singh*
CAFRAL, Reserve Bank of India

Abstract

This paper investigates the prevalence and characteristics of poor hand-to-mouth (P-HtM) and wealthy hand-to-mouth (W-HtM) households in India—an emerging economy where high household savings coexist with limited liquidity and weak credit access. Using a harmonized dataset that combines two nationally representative surveys, we construct household-level balance sheets including income, consumption, and liquid and illiquid assets. We classify households into poor and wealthy HtM categories following the methodology of Kaplan et al. (2014), and use a range of machine learning models to impute missing income data. Our findings indicate that while P-HtM households account for 2–5% of the population, W-HtM households comprise a much larger share—15–27%, with total HtM prevalence ranging from 17–32%. Classifications based solely on net worth substantially understate this share. We also find that average propensities to consume are similar across HtM and non-HtM groups, highlighting the broader financial constraints facing Indian households. These results underscore the importance of distinguishing between liquidity and net worth when evaluating consumption behavior and the transmission of fiscal and monetary policy in developing economies.

*We gratefully acknowledge the excellent support provided by our research assistants, Ken Lam and Ashwath Damle, throughout this project.

1 Introduction

Understanding the prevalence and characteristics of hand-to-mouth (HtM) households is central to analyzing consumption behavior, credit constraints, and the transmission of fiscal and monetary policy. While traditional models often associate liquidity constraints with low wealth, recent research has drawn attention to a more nuanced group: wealthy hand-to-mouth (W-HtM) households—those who hold substantial illiquid assets but lack sufficient liquid resources to smooth consumption in the face of income shocks (Kaplan et al., 2014). Identifying such households is particularly important in developing economies, where informal labor markets, limited financial access, and asset-based savings practices (e.g., in land or gold) give rise to liquidity constraints that differ from those observed in high-income settings. Whereas W-HtM behavior in advanced economies often stems from institutional features such as retirement accounts or mortgage-backed wealth, in low- and middle-income contexts it typically reflects structural frictions, including weak credit markets and high transaction costs associated with asset liquidation.

This heterogeneity in household balance sheets has become central to modern macroeconomic models, as the responsiveness of aggregate demand to fiscal or monetary stimuli depends critically on how different households adjust consumption in response to income or interest rate shocks (Kaplan and Violante, 2014; Kaplan et al., 2018; Auclert, 2019; Cloyne et al., 2020). In this context, accurately estimating the share of poor and wealthy HtM households is essential for both theoretical modeling and effective policy design. While a growing body of work has explored this in high-income countries using detailed household balance sheet data (Kaplan et al., 2014; Hara et al., 2016b; Song, 2020; Cherchye et al., 2023), there remains a notable gap in the literature for emerging economies.

India offers a particularly compelling setting for this inquiry. Despite high aggregate household savings, over 65% of Indian household wealth is held in illiquid forms such as housing, land, or gold, while access to formal credit and liquid savings instruments remains limited (RBI, 2017). These patterns raise a central question: who are India's HtM households, and how do they influence the economy's response to policy shocks?

To address this, we construct harmonized household balance sheets by combining two nationally representative surveys: the All-India Debt and Investment Survey (AIDIS) and the Consumer Pyramids Household Survey (CPHS). AIDIS provides rich data on household assets and liabilities but lacks income information; CPHS, by contrast, includes income and consumption but not detailed asset holdings. We bridge these gaps by imputing income in AIDIS using CPHS data, employing a range of techniques including consumption-based binning, ordinary least squares (OLS), and several machine learning models—decision trees, random forests, neural networks, and XGBoost.

Following the classification method of Kaplan et al. (2014), we identify poor and wealthy HtM households and estimate their prevalence. We find that the share of poor HtM households in India ranges from 2% to 5%, while the share of W-HtM households lies between 15% and 27%. This implies that the overall share of HtM households—those with high consumption

but limited liquidity—varies between 17% and 32%. When classification is based solely on net worth, as in Zeldes (1989), the estimated share narrows to between 3% and 12%, underscoring the importance of distinguishing between liquid and illiquid asset positions in this context.

We also examine the average propensity to consume (APC) across these groups. Somewhat surprisingly, APCs are similar across HtM and non-HtM categories. This may reflect the broader prevalence of financial constraints in emerging economies like India, where even households with greater net worth often face limited access to liquid savings or formal credit, dampening their ability to smooth consumption over time.

Review of Literature: This paper contributes to two main literatures. The first is the growing body of work on HtM households and their empirical prevalence across different country contexts. Early work by Zeldes (1989) used net worth to detect liquidity constraints among U.S. households, showing violations of the permanent income hypothesis and underscoring the importance of borrowing constraints in consumption behavior. A major advance came with Kaplan et al. (2014), who introduced a more nuanced classification based on liquid and illiquid asset holdings and identified a substantial group of W-HtM households. These households, despite owning illiquid wealth, behave like poor HtM households (P-HtM) when faced with income shocks due to their limited liquidity. Subsequent studies have extended this framework to high-income countries in Asia and Europe (Hara et al., 2016a; Cui and Feng, 2017; Song, 2020; Cherchye et al., 2023; Arroyo and Tisnés, 2023), while findings from developing countries suggest even higher HtM prevalence, often exceeding 50% (**?**Bracco et al., 2021). However, such studies typically rely on coarse proxies for liquidity constraints due to limited data. By leveraging two rich, nationally representative datasets from India and adopting the Kaplan et al. (2014) framework, our study provides one of the first detailed estimates of PHtM and WHtM households in a large developing economy, offering a basis for meaningful cross-country comparison and policy analysis.

The second literature we contribute to is the application of machine learning methods for household income estimation in data-scarce, developing country settings. Accurate measurement of income is crucial for identifying HtM status, but household surveys in many developing countries often lack detailed and reliable income data. Recent studies such as Wan (2023) and Wang (2022) have shown that machine learning models can substantially improve income prediction using demographic and administrative variables. Our study builds on this approach by integrating the CPHS and AIDIS datasets and using supervised learning techniques to predict household income, enabling more precise classification of HtM status. We also find that machine learning approaches outperform traditional interpolation methods (Kose et al., 2024), particularly in capturing nonlinear patterns and heterogeneity in the income distribution. This contribution is especially relevant for contexts where data constraints are severe but demographic coverage is rich—an increasingly common feature of data infrastructure in low- and middle-income countries.

The remainder of the paper is structured as follows. Section 2 outlines the AIDIS and CPHS datasets and describes the key variables used in the analysis. Section 3 focuses on the

definition and identification of HtM households, including data harmonization and income imputation methods. Section 4 presents results on the shares of each HtM category and their average propensity to consume (APC). Finally, Section 5 concludes with directions for future research.

2 Data

Identifying hand-to-mouth (HtM) households requires detailed information on households' income and assets. In the Indian context, no single dataset contains all of these variables. We thus draw on two complementary surveys - CPHS for income data and AIDIS for asset data. In this section, we describe the two datasets in detail and outline the methodological approach used to integrate them.

2.1 AIDIS

The All-India Debt and Investment Survey (AIDIS), conducted periodically by the National Statistical Office (NSO), is a principal source of data on household assets, liabilities, indebtedness, and capital formation in both rural and urban India. Initiated in the 26$^{\text{th}}$ round of the National Sample Survey (NSS) in 1971–72, subsequent rounds were conducted in NSS rounds 37, 48, 59, 70, and most recently, round 77. The latest round, conducted from January to December 2019, provides detailed information on household balance sheets as of June 30, 2018, along with capital expenditures during the 2018–19 agricultural year, disaggregated into residential, farm, and non-farm investments.

Round 77 was implemented in two phases and covered a nationally representative sample of 116,461 households—69,455 rural and 47,006 urban—across 5,940 villages and 3,995 urban blocks. Its breadth enables robust analysis of asset distribution, debt burdens, and investment patterns across socio-economic strata.

We use the 77$^{\text{th}}$ round of AIDIS as the primary source for household balance sheet data. Following Kaplan et al. (2014), we classify assets as either liquid or illiquid. Liquid assets include cash, bank deposits, money-market funds, and tradable equities; illiquid assets comprise land, housing, gold, durable goods, long-term deposits, and provident fund accounts. Using this classification, we construct household-level measures of net liquid and illiquid wealth to identify hand-to-mouth (HtM) households.

Net liquid wealth is defined as liquid assets minus liquid liabilities. The baseline definition includes cash, current and savings account balances, mutual funds, equities, bonds, and cooperative shares, while liquid liabilities comprise loans for household consumption, education, medical treatment, and litigation.

To capture broader liquidity, we construct two expanded definitions. The first, *broad 1*, adds

fixed deposits, post office savings, other fixed investments, deposits in cooperative banks and non-bank financial institutions, and gold. The second, *broad 2*, further includes productive physical assets such as transport equipment, livestock, and agricultural and non-agricultural machinery, recognizing their collateral or resale value in times of financial need.

Net illiquid wealth is defined analogously, as illiquid assets minus housing-related liabilities. The baseline illiquid asset definition includes land and buildings, long-term financial instruments (e.g., provident and pension funds, life insurance), deposits in cooperative and non-bank financial institutions, and interest-free and personal/business loans. A narrower variant excludes fixed deposits, post office accounts, and cooperative or non-bank deposits.

Wealth estimates vary significantly by definition. Mean net liquid wealth is ₹12,976 under the baseline, increasing to ₹82,056 under *broad 1* and ₹157,727 under *broad 2*, underscoring the sensitivity of results to asset classification. A substantial share of households report negative net liquid wealth, indicating widespread indebtedness. In contrast, mean net illiquid wealth exceeds ₹1.7 million, although about 10% of households report zero or negative values. Gold holdings, highly relevant in the Indian context, are common, with a mean value of ₹61,534 and a median of ₹25,000, though 17% of households report no gold assets.

Since AIDIS does not report household income, a key variable in our classification, we impute income using data from the nationally representative CPHS survey. Details on the imputation methodology are provided in Section 3.

Table 1: Summary Statistics of CPHS and AIDIS Harmonized Variables

	CPHS					AIDIS				
	Min	Max	Mean	Median	% 0s	Min	Max	Mean	Median	% 0s
Consumption (₹)	1,001.67	86,602.78	7,127.08	6,607.67	0.00	0	700,000	8,681.29	7,000	0.02
Total Income (₹)	0	927,067.06	20,776.27	15,844.55	0.00					
Income 1 (₹)	0	315,166.09	14,409.03	11,775.20	1.09					
Income 2 (₹)	0	925,703.44	20,565.44	15,744.00	0.00					
Net liquid wealth baseline (₹)						-9,956,230	96,015,000	12,975.69	5,700	0.48
Net liquid wealth broad 1 (₹)						-9,875,430	96,425,600	82,056.38	34,000	0.44
Net liquid wealth broad 2 (₹)						-9,438,430	96,445,600	157,726.59	66,000	0.14
Net illiquid wealth baseline (₹)						-48,889,200	1,147,949,952	1,789,786.10	699,860	9.84
Net illiquid wealth narrow (₹)						-48,889,200	1,147,949,952	1,782,239.50	695,500	10.42
Gold holdings (₹)						0	9,000,000	61,534.15	25,000	17.09
Region (1=urban, 0=rural)	0	1	0.34	0	65.61	0	1	0.34	0	66.39
Gender (1=male, 0=female)	0	1	0.88	1	11.88	0	1	0.86	1	13.76
Age (years)	18	110	51.43	50	0.00	18	110	47.88	47	0.00
Education	1	7	3.86	4	0.00	1	7	3.24	3	0.00
Household size	1	29	5.03	5	0.00	1	30	4.32	4	0.00
Caste	1	9	4.40	3	0.00	1	9	4.22	3	0.00
Religion	1	7	1.21	1	0.00	1	7	1.29	1	0.00
Employment type	1	8	4.23	4	0.00	1	8	4.41	4	0.00
Has bank account	0	1	0.98	1	1.66	0	1	0.93	1	7.30
Has savings in life insurance	0	1	0.50	0	50.39	0	1	0.17	0	83.27
Has savings in fixed/recurring deposits	0	1	0.61	1	38.62	0	1	0.02	0	97.62
Has savings in post office account	0	1	0.19	0	80.60	0	1	0.07	0	92.57
Has savings in gold	0	1	0.99	1	1.11	0	1	0.83	1	17.09
Has savings in businesses	0	1	0.04	0	96.29	0	1	0.00	0	99.91
Has borrowings from banks	0	1	0.11	0	88.77	0	1	0.13	0	87.05
Has borrowings from employer	0	1	0.00	0	99.75	0	1	0.00	0	99.96
Has borrowings from relatives/friends	0	1	0.11	0	89.14	0	1	0.05	0	94.87
Has borrowings from NBFC/MFI	0	1	0.03	0	97.34	0	1	0.02	0	98.39
Has borrowings from Self Help Group	0	1	0.06	0	94.31	0	1	0.00	0	99.99
Has borrowings from chit fund	0	1	0.00	0	99.59	0	1	0.00	0	99.84
Has borrowings from shops	0	1	0.17	0	82.86	0	1	0.00	0	99.81
Has borrowings from money lender	0	1	0.05	0	95.41	0	1	0.00	0	99.95
Has borrowings for education	0	1	0.02	0	98.39	0	1	0.01	0	98.91
Has borrowings for medical treatment	0	1	0.02	0	98.22	0	1	0.03	0	97.01
Has borrowings for repayment	0	1	0.04	0	95.75	0	1	0.01	0	99.18
Has borrowings for housing	0	1	0.05	0	94.84	0	1	0.05	0	94.69
Has borrowings for household expenditures	0	1	0.25	0	74.76	0	1	0.11	0	89.14
Has borrowings for business/investment	0	1	0.06	0	93.97	0	1	0.12	0	88.00

Note: Asset and liability data refer to the second wave of 2018 (May–August), while consumption and income figures represent 2019 annual averages. Monetary variables are expressed in current rupees (₹). Categorical variables: Education (1=not literate, 2=below primary, 3=primary, 4=upper primary/middle, 5=secondary/higher secondary (including diploma/certificate), 6=graduate (including diploma/certificate), and 7=postgraduate and above); Caste (1=scheduled tribe, 2=scheduled caste, 3=other backward class, 9=other); Religion (1=Hinduism, 2=Islam, 3=Christianity, 4=Sikhism, 5=Jainism, 6=Buddhism, 7=Other); Employment type (1=urban, self-employed, 2=urban, regular wage earning, 3= urban, casual labor, 4=rural, self-employed in agriculture, 5=rural, self-employed in non-agriculture, 6=rural, regular wage earning, 7=rural, casual labor in agriculture, 8=rural, casual labor in non-agriculture) ."Has savings/borrowings" variables are binary indicators, where 1 denotes "yes" and 0 denotes "no."

2.2 CPHS

The second dataset used in this study is the *Consumer Pyramids Household Survey (CPHS)*, developed and maintained by the Centre for Monitoring Indian Economy (CMIE). CPHS is the world's largest high-frequency household panel dataset, collecting data since 2014. It surveys approximately 174,000 households three times annually, producing a panel that spans thirty three waves through December 2024. The survey employs a stratified, multi-stage sampling design representative of both urban and rural India.

CPHS comprises four modules that together capture rich information on household economic behavior. The *Consumption Pyramids* module records monthly recall-based expenditures across food, non-food, and durable goods. The *Income Pyramids* module collects monthly recall-based income at the household and individual levels, covering labor income, government transfers, remittances, and income from self-production. The *Aspirational India* module reports ownership and intended acquisition of key household assets, saving behavior, outstanding debt (by source and purpose), and perceptions of well-being. The *People of India* module provides demographic, socioeconomic, and employment characteristics for all household members. A notable limitation is that all asset-related variables are binary, restricting analysis on the intensive margin of asset holdings.

We construct three alternative definitions of monthly household income. The broadest includes all income earned by household members—wages, dividends, interest, rental income, transfers, business profits, and other sources. This measure yields a mean of ₹20,776 and a median of ₹15,845, indicating a right-skewed distribution. Following Kaplan et al. (2014), we also define two narrower measures. The narrowest includes only wages and government transfers. The intermediate definition, which is our preferred measure, includes wages, pensions, self-production, private and government transfers, and business profits. This intermediate income measure closely approximates total income, with a mean of ₹20,565 and a median of ₹15,744, suggesting it captures most of the relevant variation in household income.

2.3 Combining CMIE and AIDIS

Identifying hand-to-mouth (HtM) households following the methodology of Kaplan et al. (2014) requires data on household income, liquid and illiquid assets, and liabilities. While the CPHS dataset provides detailed information on household income, it lacks quantitative data on assets and liabilities. Conversely, AIDIS contains rich information on household assets and liabilities but does not report income. To address this limitation, we integrate information from both datasets to construct a measure of HtM status. This section outlines the construction of a harmonized dataset that includes variables common to both AIDIS and CPHS.

Both CPHS and AIDIS provide information on selected indicators of consumption, financial integration, and demographic characteristics. However, the two datasets differ substantially

in frequency and structure, necessitating specific assumptions to enable meaningful comparisons. CPHS is a panel dataset initiated in 2014, in which each household is surveyed three times per year. Although survey waves occur triannually, data on consumption and income are collected at a monthly frequency, relying on respondents' recall. In contrast, AIDIS is a cross-sectional survey conducted approximately every five years, with the most recent round completed in 2019. This round was administered in two phases: the first from January to August 2019, and the second from September to December 2019. The survey includes both contemporaneous data (i.e. pertaining to the time of the interview) for variables such as demographics and consumption, and retrospective data, based on respondents' recall of asset and liability holdings as of the end of June 2018.

To match the retrospective variables of AIDIS, which include assets and liabilities, we use information from the second wave of the 2018 CPHS (May–August). To match the contemporaneous variables of AIDIS, which include consumption and demographics, we use information from the three waves conducted in 2019 in CPHS. Accordingly, we restrict the CPHS sample to households that participated in the survey in both 2018 and 2019 and reported recalled monthly consumption in at least eight months of 2019.

To ensure comparability between AIDIS and CPHS data, we harmonize the consumption measures by restricting our analysis to reported expenditures on purchases. In AIDIS, household consumption is recorded as usual monthly consumer expenditure and includes both monetary and imputed components such as consumption from homegrown stock, wages in kind, freely collected goods, and gifts. Since these imputed items are not treated as expenditure categories in CPHS, we exclude them from our analysis.[1] We focus instead on CPHS consumption categories that align with purchased expenditures, including food, intoxicants (cigarettes, tobacco, and liquor), clothing and footwear, cooking fuel, electricity, utility and maintenance bills (such as water, society charges, and similar expenses), and rent.

Table 1 presents a comparison of key variables in the harmonized dataset, with values weighted using each survey's sampling weights. While consumption expenditure is reported in both CPHS and AIDIS, notable differences emerge in its distribution. Income data are available only in CPHS, whereas wealth-related variables are exclusive to AIDIS. Both datasets include demographic and financial variables, though the latter show more pronounced variation. Average monthly household consumption is ₹7,127 in CPHS and ₹8,681 in AIDIS. Despite similar medians, the maximum in AIDIS is substantially higher, suggesting outliers.

Demographic characteristics of household heads (restricted to those above 18) are broadly similar across surveys. Roughly one-third of households are urban in both datasets, indicating a primarily rural composition. Male headship dominates, at 88% in CPHS and 86% in AIDIS, consistent with prevailing gender norms. The average age of household heads is slightly higher in CPHS (51.4 years) than in AIDIS (47.9 years), though both medians are close to 50. Educational attainment is modest in both datasets, with most household heads reporting primary or middle-level schooling.

[1] In CPHS, self-consumption of agricultural produce is recorded as income rather than expenditure.

Average household size is larger in CPHS (just over five members) compared to AIDIS (4.3). Caste and religious compositions are comparable, with a concentration among Other Backward Classes and a Hindu majority. Employment patterns are also similar, with rural self-employment—particularly in agriculture—being the dominant occupation.

Financial inclusion differs markedly. Bank account ownership is nearly universal in both datasets (98% in CPHS; 93% in AIDIS), but formal savings are far more common in CPHS. Life insurance coverage is reported by 50% of CPHS households but only 17% in AIDIS. Fixed or recurring deposit use is 61% in CPHS versus 2% in AIDIS; similar gaps exist for post office and gold savings. Business-related savings are rare in both surveys, though virtually absent in AIDIS.

Borrowing patterns show divergence as well. Bank borrowings are relatively uncommon (11% in CPHS, 13% in AIDIS), but informal borrowing—from friends, shops, or moneylenders—is more frequently reported in CPHS. Purpose-specific borrowing (e.g., for education, medical needs, or business) is generally rare, yet more commonly observed in CPHS. Notably, borrowing for household consumption is reported by 25% of CPHS households, compared to only 11% in AIDIS, indicating possible differences in credit use or reporting conventions.

3 Methodology

The primary objective of this paper is to classify each household in the combined 2019 cross-section as poor hand-to-mouth (P-HtM), wealthy hand-to-mouth (W-HtM), or non-hand-to-mouth (N-HtM). We begin by outlining the theoretical framework for this classification, drawing on the methodologies proposed by Kaplan et al. (2014) and Zeldes (1989). Implementing this framework requires information on household liquid asset holdings and current disposable income—variables that are not jointly observed in either the AIDIS or CPHS datasets. To overcome this limitation, we develop a set of imputation strategies that combine econometric techniques and machine learning algorithms to match and harmonize the datasets. These approaches are designed to preserve the joint distribution of income, consumption, and wealth, ensuring consistency with the underlying economic behavior. This section provides a detailed discussion of the various matching methodologies employed in the classification exercise.

3.1 Defining Hand-to-Mouth Households

Hand-to-mouth (HtM) households are those that consume their entire income within the same pay period, without setting aside any resources for future consumption. Following the framework of Kaplan et al. (2014), we distinguish between two types of HtM households based on their asset portfolios: poor hand-to-mouth (P-HtM) and wealthy hand-to-mouth (W-HtM). P-HtM households lack both liquid assets and illiquid wealth, leaving them financially

constrained in both the short and long run. In contrast, W-HtM households hold substantial illiquid assets such as housing, retirement accounts, or time deposits, but possess little or no liquid wealth. As a result, despite being wealthier on paper, these households are unable to smooth consumption in response to small income shocks due to limited access to liquid resources.

Following Kaplan et al. (2014), a household i is deemed HtM if its real liquid asset position m_i satisfies

$$0 \leq m_i \leq \frac{y_i}{2} \tag{1}$$

or the household can be considered as the borrower if he has negative liquid assets:

$$m_i < 0 \quad \text{and} \quad m_i \leq \frac{y_i}{2} - \bar{m}_i. \tag{2}$$

where y_i denotes monthly disposable income and $\bar{m}_i$ is an exogenous formal credit limit that we set to one month of income. A HtM household with strictly positive illiquid wealth a_i is classified as W-HtM; one with $a_i = 0$ is P-HtM. All remaining households are Non-HtM.

To compare with traditional definition, we also calculate hand-to-mouth in terms of net worth (HtM-NW). A household i with net worth $n_i = a_i + m_i$ can be identified as HtM in terms of net worth if

$$0 \leq n_i \leq \frac{y_i}{2} \tag{3}$$

or

$$n_i \leq 0 \quad \text{and} \quad n_i \leq \frac{y_i}{2} - \bar{m}_i. \tag{4}$$

where a_i and m_i are the liquid and illquid asset holdings by household i.

3.2 Income Imputation

To generate income estimates in AIDIS for identifying HtM households, we employ five distinct approaches. The first is a straightforward back-of-the-envelope calculation based on consumption bins. The remaining four methods incorporate a broader set of household characteristics beyond consumption. Specifically, the second approach relies on ordinary least squares (OLS) regression, while the last three utilize machine learning techniques. In the sections that follow, we provide a concise description of each methodology and discuss their implementation in our context.

3.2.1 Consumption Bins

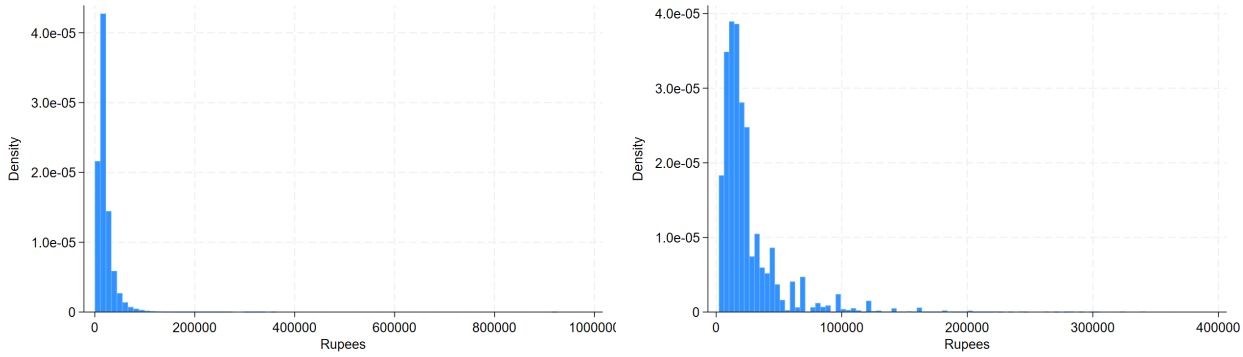
Our first approach to estimating household income for the AIDIS dataset employs a model-free method based on the concept of the average propensity to consume (APC). The APC is defined as the ratio of consumption (C) to income (Y), that is, APC = C/Y. This metric

reflects the average share of income that households allocate to consumption within a given time period, monthly in our case.

Recall that CPHS contains both consumption and income data, while AIDIS includes consumption but not income. To impute income in AIDIS, we begin by sorting households in both datasets into hundred consumption-based bins, defined identically across the two surveys. In CPHS, where both income and consumption are observed, we compute the average propensity to consume (APC) for each bin by averaging the APCs of all households within that bin. Using these bin-specific APCs, we impute income in AIDIS by applying the inverse relationship $Y = C/APC$, where C is household consumption.

Figure 1 presents a comparison between actual and predicted monthly household income distributions. The left panel displays the observed income distribution based on CPHS, while the right panel illustrates the distribution of predicted income using consumption bin classifications. The RMSE of 0.44 indicates that, on average, the predicted incomes deviate from actual incomes by approximately 44% of one unit of the income variable's scale. This reflects a moderate level of predictive accuracy.

Figure 1: Monthly Household Income Distributions using Consumption Bins



(a) CPHS

(b) Predicted in AIDIS using consumption bins

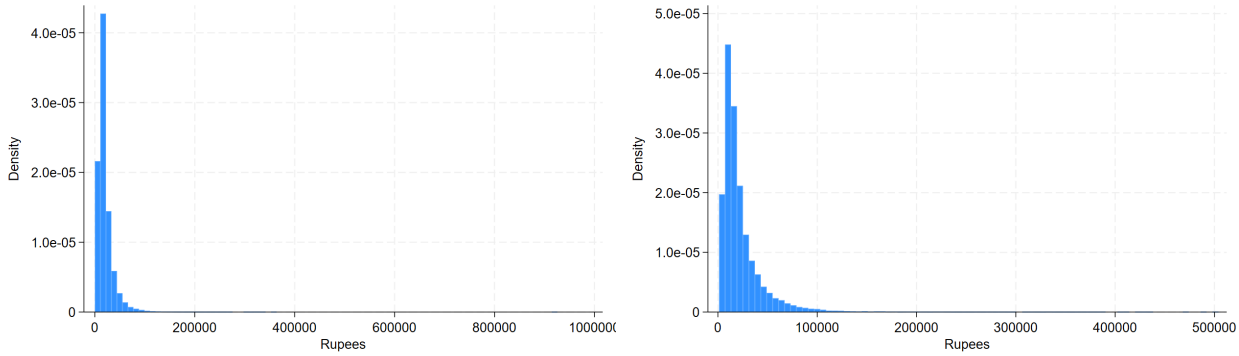
3.2.2 Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) regression (Greene, 2018; Wooldridge, 2019) is a widely used statistical method used to estimate the relationship between a dependent variable and one or more independent variables. In its standard form, OLS fits a linear model by minimizing the sum of squared residuals, which is the differences between observed and predicted values, yielding coefficient estimates that capture the direction and magnitude of association between independent and the dependent variables.

OLS is valued for its simplicity and interpretability: each coefficient reflects the marginal effect of an independent variable, holding others constant. It also facilitates hypothesis testing by identifying statistically significant predictors. However, the method rests on key assump-

tions, notably linearity in parameters. Violations of this assumption can result in biased or inefficient estimates. Moreover, multicollinearity among regressors can inflate standard errors, obscuring the effects of individual variables. It is therefore important to assess model diagnostics and consider alternative specifications or transformations when appropriate.

Figure 2: Monthly Household Income Distributions using OLS



(a) CPHS

(b) Predicted in AIDIS using OLS

Figure 2 presents a comparison between actual and predicted monthly household income distributions. The left panel displays the observed income distribution based on CPHS, while the right panel illustrates the distribution of predicted income using an OLS regression model. The root mean squared error (RMSE) of 0.37 indicates that, on average, the predicted incomes deviate from actual incomes by approximately 37% of one unit of the income variable's scale, suggesting a reasonably good level of predictive accuracy.

The model explains approximately 60% of the variation in (log) income. The estimated coefficients are broadly consistent with theoretical expectations. As anticipated, there is a strong positive association between (log) income and (log) usual consumption. Higher income levels are observed among households located in urban areas, those with a male household head, and those with older household heads. Income also increases with household size and educational attainment, particularly at the graduate and postgraduate levels. Individuals from upper caste groups and those engaged in regular wage employment, report significantly higher income levels.

The effects of financial variables on income are more mixed. Ownership of savings is generally associated with higher income, although savings in fixed deposits and gold are exceptions, showing negative or negligible effects. Borrowing from various sources is typically positively associated with income, with the exception of borrowing from shops, which is negatively associated. Similarly, the purpose of borrowing plays a role: loans taken for business investment or repayment are positively linked to income, whereas borrowing for education, medical treatment, or housing tends to be associated with lower income levels.

3.2.3 Decision Tree

A decision tree (Mienye and Jere, 2024) operates by recursively partitioning the data into subsets based on feature values, forming a tree-like structure where each internal node represents a decision rule on an attribute, and each leaf node corresponds to an outcome. For example, a split might occur at a node based on whether household expenditure falls below a threshold (e.g., 1500), a value determined during training. This rule-based structure allows the model to capture complex, non-linear relationships in the data without requiring prior feature scaling, such as normalization or standardization.

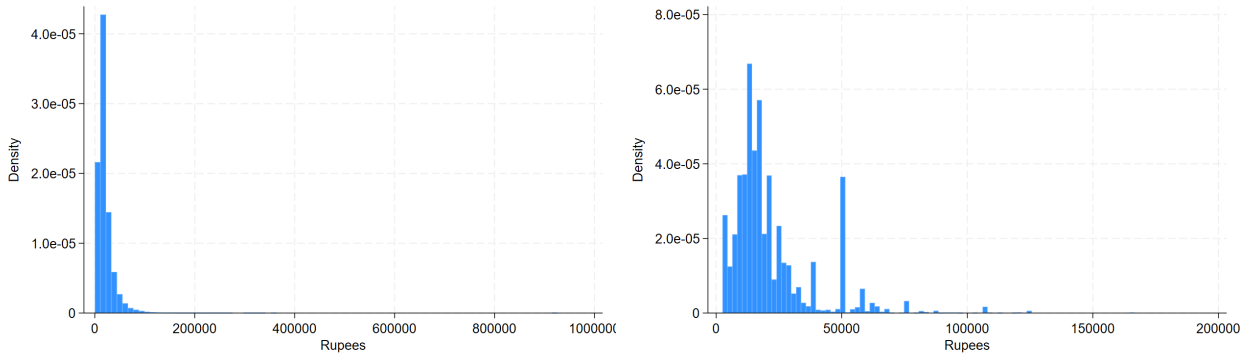
One of the primary advantages of using decision trees is the interpretability and transparency of the output, making it easier to visualize and understand the decision-making process. However, the method also has notable limitations. Decision trees are sensitive to outliers and may favor attributes with many levels, which can introduce bias. Additionally, while effective in modeling non-linear patterns, they often struggle with linear relationships and are prone to overfitting, reducing their generalization to unseen data. To mitigate computational inefficiencies, especially with large datasets, dimensionality reduction techniques like Principal Component Analysis (PCA) can be applied. PCA helps enhance computational performance by reducing the number of input features while preserving the most critical information from the original dataset.

Model robustness is ensured using five-fold cross-validation. Hyperparameters, which are variables that control different aspects of training, are tuned via a combination of Bayesian optimization and random search to enhance predictive accuracy while mitigating overfitting. Performance is evaluated using Root Mean Squared Error (RMSE) and the coefficient of determination (R^2), both calculated on validation folds. RMSE measures the average magnitude of prediction errors and is expressed in the same units as the outcome variable, enhancing interpretability. R^2 quantifies the proportion of variance in the target variable explained by the model, offering insight into its explanatory power.

The best-performing model used the following configuration: maximum tree depth of 9, unrestricted feature consideration at splits, a minimum of 7 observations per leaf node, and a minimum of 20 observations to initiate a split. This setup yielded a validation-weighted RMSE of 0.3580 and an R^2 of 0.6311. Feature importance scores were computed to quantify the contribution of each predictor, aiding interpretability in a non-linear, non-parametric framework.

Figure 3 presents a comparison between actual and predicted monthly household income distributions. The left panel displays the observed income distribution based on CPHS, while the right panel illustrates the distribution of predicted income using a decision tree model. The RMSE of 0.36 indicates that, on average, the predicted incomes deviate from actual incomes by approximately 36% of one unit of the income variable's scale, reflecting a strong predictive performance.

Figure 3: Monthly Household Income Distributions using Decision Tree



(a) CPHS

(b) Predicted in AIDIS using decision tree

3.2.4 Random forest

Random forests (Biau, 2012) are an ensemble learning method that enhances predictive accuracy by aggregating the outputs of multiple decision trees. A decision tree is a non-parametric model that recursively partitions the data based on feature thresholds, creating a tree-like structure where each leaf representing a prediction. Although decision trees are intuitive and can handle both categorical and numerical variables, they are prone to overfitting when used individually.

Random forests mitigate overfitting by training multiple decision trees on bootstrapped samples of the data and random subsets of features, then aggregating their predictions to improve accuracy and robustness. This method captures complex, non-linear relationships while reducing sensitivity to multicollinearity and outliers. Random forests also accommodate missing data via surrogate splits, though this form of implicit imputation may not be appropriate in all contexts. The method requires no data standardization and offers measures of feature importance. However, their complexity makes them computationally intensive and less interpretable than single decision trees.

Model evaluation was conducted using five-fold cross-validation. The dataset was partitioned into five equally sized folds; in each iteration, four folds were used for training and the fifth for validation. Repeating this process five times ensures that each subset serves as the validation set once. Such a strategy provides a reliable estimate of the model's out-of-sample performance and helps to mitigate overfitting.

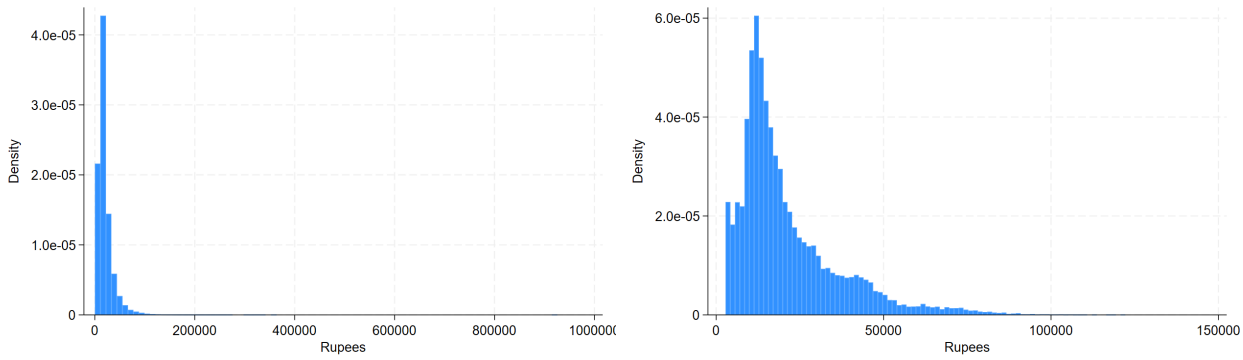
To optimize hyperparameters, we employed Bayesian optimization, which is a probabilistic search method that constructs a surrogate model to predict performance across the hyperparameter space and selects configurations based on expected improvement. This approach is more efficient than traditional grid or random search and enhances the likelihood of identifying high-performing settings. The optimization was conducted over a predefined space, including the number of estimators (50–300), tree depth (3–30), minimum samples to split

a node (2–20), and minimum samples per leaf (1–20). Categorical hyperparameters such as the maximum number of features, bootstrap usage, and the splitting criterion (squared or absolute error) were also tuned.

Model performance was assessed using two metrics: root mean squared error (RMSE) and the coefficient of determination (R^2). To interpret the fitted model, we examined feature importance scores derived from the ensemble of decision trees. These scores reflect each variable's contribution to reducing prediction error across splits, with higher values indicating greater influence on model accuracy.

Figure 4 presents a comparison between actual and predicted monthly household income distributions. The left panel displays the observed income distribution based on CPHS, while the right panel illustrates the distribution of predicted income using the random forest model described above. The RMSE of 0.32 indicates that, on average, the predicted incomes deviate from actual incomes by approximately 32% of one unit of the income variable's scale, demonstrating a high level of predictive accuracy.

Figure 4: Monthly Household Income Distributions using Random Forest



(a) CPHS

(b) Predicted in AIDIS using random forest

3.2.5 Neural Network

Neural networks (Warner and Misra, 1996) are computational models inspired by the human brain, consisting of interconnected layers of nodes (neurons) that process input data through a series of mathematical operations. In our application, the number of neurons in the input layer corresponds to the number of input variables. Data is passed forward through the network, producing an output that is compared against actual outcomes using a loss function. During training, the model iteratively adjusts the weights of the connections between nodes through backpropagation to minimize this loss, improving prediction accuracy.

Neural networks are powerful tools for modeling complex, non-linear relationships in data and benefit from a variety of optimization algorithms designed to find the global minimum of the loss function. However, they are often described as "black box" models due to limited

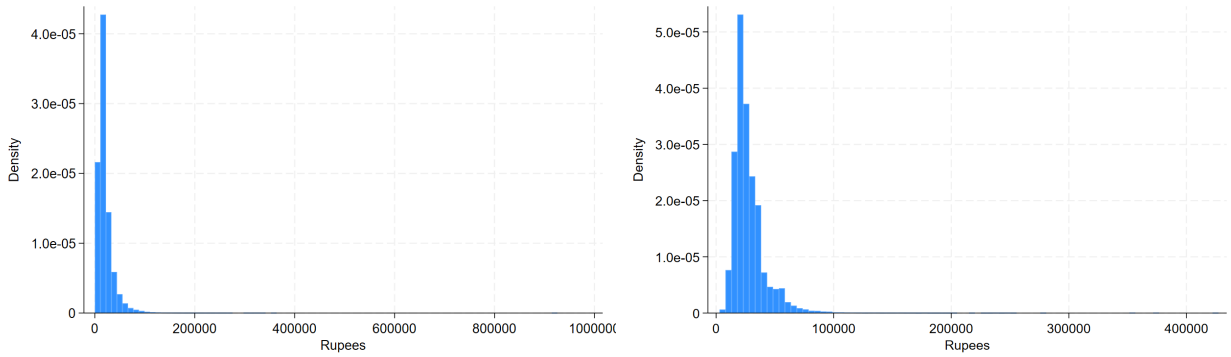
interpretability of their internal computations. Additionally, training neural networks can be computationally intensive, especially with large datasets, and requires careful tuning of hyperparameters such as the choice of optimizer, loss function, network architecture, and number of training epochs to achieve optimal performance.

We employed a fully connected feedforward neural network to predict income. The architecture comprised six layers in total: an input layer with 66 neurons (one for each input feature), followed by four hidden layers with 256, 128, 64, and 32 neurons, and a final output layer with a single neuron. To improve generalization and prevent overfitting, we applied a dropout rate of 0.4 to the hidden layers, randomly deactivating 40% of neurons during each training iteration. Batch normalization was used after each hidden layer to normalize intermediate activations, accelerating convergence and improving training stability. Additionally, L2 regularization with a coefficient of 0.001 was applied to discourage large weight values. The model was trained using early stopping with a patience of 20 epochs—training was halted if validation loss failed to improve for 20 consecutive epochs.

The loss function used for training was quantile loss with a quantile parameter $p = 0.9$. Unlike traditional loss functions such as mean squared error or Huber loss, which focus on minimizing average error, quantile loss allows the model to focus on a specific conditional quantile of the target distribution. In this case, the model was trained to make predictions such that 90% of actual incomes fall below the predicted value. This formulation penalizes underpredictions more heavily than overpredictions, making it particularly suitable for skewed distributions or scenarios where underestimation carries greater risk. The model's performance was evaluated using the Root Mean Squared Error (RMSE), providing an interpretable measure of prediction accuracy in the same units as the target variable.

Figure 5 presents a comparison between actual and predicted monthly household income distributions. The left panel displays the observed income distribution based on CPHS, while the right panel illustrates the distribution of predicted income using the neural network model described above. The RMSE of 0.60 indicates that, on average, the predicted incomes deviate from actual incomes by approximately 60% of one unit of the income variable's scale, suggesting a moderate level of predictive accuracy.

Figure 5: Monthly Household Income Distributions using Neural Network



(a) CPHS

(b) Predicted in AIDIS using neural network

3.2.6 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) is a powerful ensemble learning algorithm that constructs a series of decision trees in a sequential manner, where each new tree is trained to correct the residual errors of the previous ones. This "boosting" technique improves predictive accuracy by minimizing a specified loss function through gradient descent. XGBoost further enhances efficiency by optimizing tree splits and pruning branches that do not contribute meaningfully to error reduction, resulting in a model that is both accurate and computationally efficient.

One of XGBoost's key advantages lies in its flexibility and robustness. It can handle missing values internally by learning optimal assignments during training and performs well with both numerical and categorical features. Additionally, its ability to implicitly manage multicollinearity by selecting only one feature among a group of correlated predictors during splits makes it particularly suitable for high-dimensional data. While the model's complexity and sequential training process make it computationally demanding, these same characteristics also allow it to model intricate, non-linear relationships in the data.

In our implementation, XGBoost was trained using five-fold cross-validation to ensure generalizability and robustness. Hyperparameter tuning combined Bayesian optimization and random search. Bayesian optimization efficiently explored the hyperparameter space by leveraging a surrogate probabilistic model, typically a Gaussian process or a tree-structured Parzen estimator, to guide the search toward regions with higher expected improvement. In contrast, random search sampled hyperparameters uniformly from predefined distributions, offering broader and more diverse coverage of the search space.

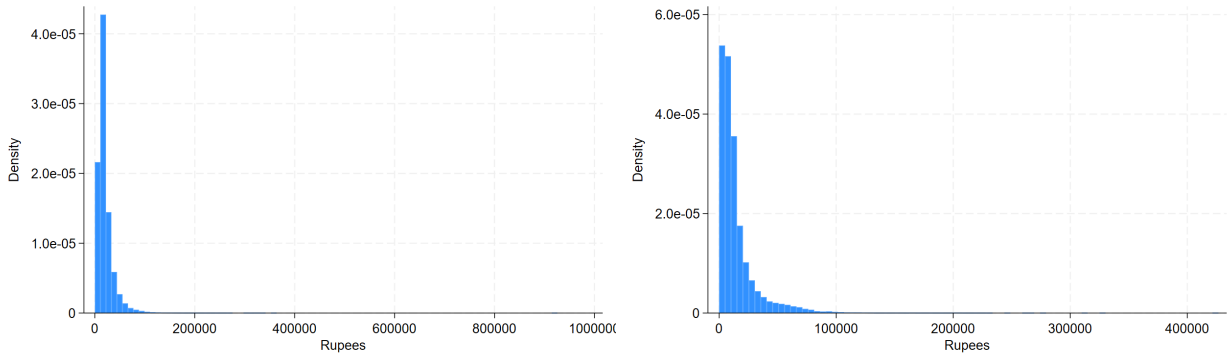
Model performance was assessed using root mean squared error (RMSE) and the coefficient of determination (R^2), providing a comprehensive evaluation of predictive accuracy and explanatory power. To enhance interpretability, we examined feature importance scores, which quantify the frequency and effectiveness with which each variable contributes to reducing

prediction error across the ensemble of trees.

The best-performing XGBoost model configuration included a learning rate of approximately 0.038, 303 estimators (i.e., gradient-boosted trees), and a maximum tree depth of 8. Additional tuned parameters included a minimum sum of instance weights (hessian) of 4, and a minimum loss reduction of 0.459 required to make a split. The model also employed a column subsampling rate of 0.737 and a row subsample rate of 0.923, controlling the fraction of features and observations used in each tree, respectively. Regularization was incorporated through an L1 penalty of 0.017 and an L2 penalty of 0.763. This optimized configuration effectively balanced model complexity, generalization, and predictive accuracy.

Figures 6 and 7 present a comparison between actual and predicted monthly household income distributions. The left panel displays the observed income distribution based on CPHS, while the right panel illustrates the distribution of predicted income using the XGBoost model described above. Figure 6 corresponds to the default model, while Figure 7 corresponds to the fine tuned model. The RMSE of 0.32, in both cases, indicates that, on average, the predicted incomes deviate from actual incomes by approximately 32% of one unit of the income variable's scale, suggesting a moderate level of predictive accuracy.

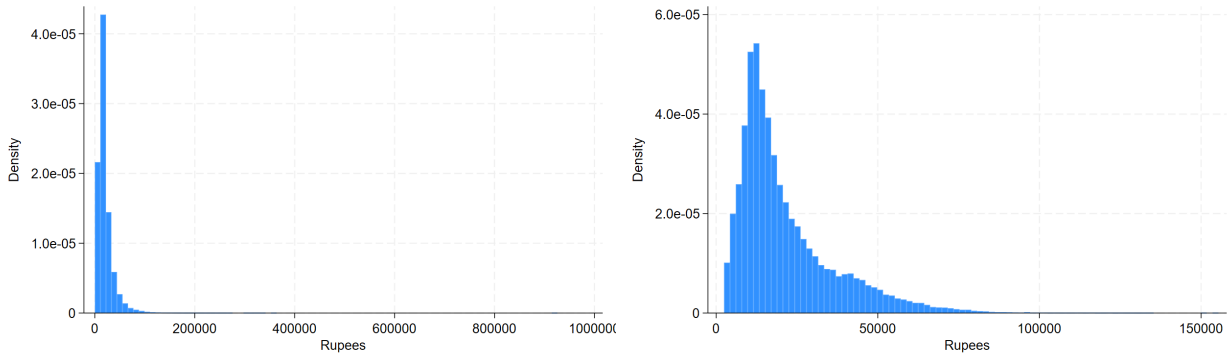
Figure 6: Monthly Household Income Distributions using XGBoost, Default



(a) CPHS

(b) Predicted in AIDIS using XGBoost, default

Figure 7: Monthly Household Income Distributions using XGBoost, Fine Tuned



(a) CPHS

(b) Predicted in AIDIS using XGBoost, fine tuned

4 Results

Table 2 reports the estimated shares of poor hand-to-mouth (P-HtM) and wealthy hand-to-mouth (W-HtM) households in the overall population, as well as the share of HtM households identified based on net worth (HtM-NW). The share of P-HtM households ranges from 2% to 5%, while the proportion of W-HtM households lies between 15% and 27%. Consequently, the total share of HtM households varies from 17% to 32%. When classification is based solely on net worth, the share of HtM households falls within a narrower range, between 3% and 12%, highlighting the importance of distinguishing between liquid and illiquid assets and liabilities when implementing this classification.

Table 2: Shares of HtM Households

Income prediction method	P-HtM	W-HtM	HtM-NW
Consumption Bins	.0336	.1886	.0416
OLS	.0509	.2343	.1208
Decision Tree	.0320	.1829	.0401
Random Forest	.0309	.1810	.0388
Neural network	.0466	.2683	.1158
XGBoost-Default	.0233	.1518	.0312
XGBoost-Fine tuned	.0318	.1839	.0398

Table 3 reports the average propensity to consume (APC) by HtM category. Somewhat surprisingly, APCs appear relatively similar across all groups. One possible explanation is that, while average consumption behavior is comparable, the marginal propensity to consume (MPC) may differ across categories. Alternatively, the similarity in APCs may reflect the

broader prevalence of financial constraints in an emerging economy, where even non-HtM households face limited access to credit or savings instruments.

Table 3: Average Propensity to Consume

Income prediction method	P-HtM	W-HtM	Non-HtM	HtM-NW	Non-HtM-NW
Consumption Bins	0.3407	0.3416	0.3447	0.3389	0.3458
OLS	0.3383	0.3666	0.3721	0.3243	0.3833
Decision Tree	0.4250	0.4087	0.4124	0.4185	0.4035
Random Forest	0.4558	0.4202	0.4038	0.4481	0.4064
Neural Network	0.3465	0.3166	0.3123	0.3545	0.3111
XGBoost-Default	0.5041	0.5055	0.7443	0.5110	0.6040
XGBoost-Fine Tuned	0.4315	0.4124	0.3997	0.4295	0.4017

4.1 Robustness

To assess the reliability of our baseline estimates and the validity of the household classifications employed, we conduct a set of robustness checks addressing concerns related to measurement error, classification criteria, and national representativeness.

First, we evaluate the sensitivity of our hand-to-mouth (HtM) classification to alternative definitions. While our primary approach follows Kaplan et al. (2014), we reclassify households using a net-worth-based definition consistent with Zeldes (1989), based on wealth-to-income ratios. Table 4 compares these estimates, showing broad alignment with our baseline results while highlighting differences in the scope of liquidity constraints captured.

Second, we explore alternative specifications of liquid and illiquid asset categories. This includes reclassifying financial instruments such as equities, mutual funds, debentures, and business wealth as illiquid assets. We also modify credit access assumptions by increasing the allowable credit limit from one month's income to one year's income. Additionally, we vary the assumed income pay frequency (weekly vs. biweekly) to assess how it affects HtM status. Under these alternative scenarios, the proportion of households identified as HtM remains relatively stable, reinforcing the robustness of our results.

Third, we examine a subgroup of financially fragile households—defined as those whose liquid assets fall below a threshold of their regular consumption plus ₹2,000. This definition captures marginally more constrained households and results in a higher share of HtM classification, underscoring the sensitivity of estimates to liquidity thresholds.

Finally, we validate the national representativeness of the AIDIS and CMIE-CPHS datasets by comparing aggregate statistics on assets, debt, and investment with those reported by the Reserve Bank of India (RBI). We also benchmark CPHS consumption data against the

Household Consumption Expenditure Survey (HCES) conducted by the National Sample Survey Office (NSSO). These comparisons confirm that both datasets align well with official sources, supporting their use for nationally representative analysis.

5 Conclusion

This paper presents, to the best of our knowledge, the first economy-wide measurement of liquidity constraints in India, based on a novel harmonized dataset that combines the asset detail of AIDIS with the income, consumption, and demographic granularity of CPHS. By employing multiple income imputation techniques including consumption binning, ordinary least squares (OLS), and several machine learning models, we offer a robust framework for identifying hand-to-mouth (HtM) households. In doing so, the paper also lays the groundwork for future applications of machine learning methods to handle missing variables in large-scale household datasets.

Our results reveal several key findings. First, roughly one in three Indian households is classified as HtM, and over half of these are wealthy hand-to-mouth (W-HtM), a group characterized by significant illiquid wealth but insufficient liquid assets to buffer short-run shocks. Second, both poor and wealthy HtM households exhibit average propensities to consume (APCs) that are approximately the same as those of unconstrained households.

Our analysis underscores the urgent need for more frequent and detailed household balance sheet data in India. Such data would allow researchers to track liquidity positions over time, analyze vulnerability to macroeconomic shocks, and evaluate the long-run impact of policy interventions.

Several limitations of this study warrant mention. Our assumption regarding credit limits, though calibrated to Indian microfinance data, is necessarily stylized; access to administrative borrowing records would enhance the accuracy of our classifications. Additionally, measurement error in self-reported consumption may introduce bias into income imputations. Finally, the cross-sectional nature of AIDIS precludes analysis of life-cycle dynamics in wealth accumulation and liquidity access.

Future research can build on this work by using our harmonized dataset within a Heterogeneous-Agent New Keynesian (HANK) framework calibrated to India, to quantify the general equilibrium effects of liquidity constraints. The development of similar harmonized datasets in other developing countries could also advance the study of household behavior where administrative data is scarce. Linking future rounds of AIDIS with the expanding CPHS panel would allow researchers to trace household liquidity trajectories over time and examine the effects of financial innovations, such as the expansion of digital payments, on household resilience. For now, our findings make one thing clear: understanding India's macroeconomy requires acknowledging a vast population that sits atop substantial illiquid wealth, yet remains one short-lived shock away from cutting consumption.

References

Arroyo, C. and Tisnés, E. (2023). What drives cross-country differences in the share of hand-to-mouth households? Technical report, CEMFI Working Paper No. 2305.

Auclert, A. (2019). Monetary policy and the redistribution channel. *American Economic Review*, 109(6):2333–2367.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research 13 (2012) 1063-1095*, 13:1063–1095.

Bracco, J. R., Galeano, L. M., Juarros, P. F., Riera-Crichton, D., and Vuletin, G. J. (2021). Social transfer multipliers in developed and emerging countries: The role of hand-to-mouth consumers. Technical Report 9627, World Bank Policy Research Working Paper Series.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785 – 794.

Cherchye, L., De Rock, B., and Vermeulen, F. (2023). Identifying wealthy hand-to-mouth households in belgium. *National Bank of Belgium Working Paper*, (419).

Cloyne, J., Ferreira, C., and Surico, P. (2020). Monetary policy when households have debt: New evidence on the transmission mechanism. *Review of Economic Studies*, 87(1):102–129.

Cui, Q. and Feng, S. (2017). Hand-to-mouth households in china. *China Economic Review*, 44:1–15.

Greene, W. H. (2018). *Econometric Analysis*. Pearson, New York, 8th edition.

Hara, K., Kitao, S., and Kawaguchi, D. (2016a). Identification of hand-to-mouth households in japan. *Review of Income and Wealth*, 62(4):723–752.

Hara, K., Unayama, T., and Weidner, J. (2016b). The wealthy hand-to-mouth in japan. *RIETI Discussion Paper Series*, (16-E-073).

Kaplan, G., Moll, B., and Violante, G. L. (2018). Monetary policy according to hank. *American Economic Review*, 108(3):697–743.

Kaplan, G. and Violante, G. L. (2014). A model of the consumption response to fiscal stimulus payments. *Econometrica*, 82(4):1199–1239.

Kaplan, G., Violante, G. L., and Weidner, J. (2014). The wealthy hand-to-mouth. *Brookings Papers on Economic Activity*, 2014(1):77–153.

Kose, E., Manley, H., and Miller, D. (2024). Backcasting population data in the 1960s with supervised learning. *Demographic Research*, 50:1123–1154.

Mienye, I. D. and Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *Institute of Electrical and Electronics Engineers (IEEE)*, 12:86716–86727.

RBI (2017). Report of the household finance committee.

Song, M. (2020). Consumption responses and the hand-to-mouth households in korea. *Korean Economic Review*, 36(2):173–205.

Wan, Z. (2023). Performance evaluation of machine learning models on income forecasting. *Applied and Computational Engineering*, 27:24–29.

Wang, J. (2022). Research on income forecasting based on machine learning methods and the importance of features. EAI.

Warner, B. and Misra, M. (1996). Understanding neural networks as statistical tools. *The American Statistician*, 50:284–293.

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach*. Cengage Learning, Boston, MA, 7th edition.

Zeldes, S. P. (1989). Consumption and liquidity constraints: An empirical investigation. *Journal of Political Economy*, 97(2):305–346.

Appendix

A Alternative Definitions of HtM

In this section, we compare our baseline estimates of hand-to-mouth (HtM) households with an alternative classification strategy proposed by Zeldes (1989). This approach relies on a net-worth-based criterion, as outlined in the identification section. Table 4 presents the distribution of households across HtM categories using two distinct definitions: one based on net worth (HtM_{NW}), following Zeldes (1989), and the other based on liquid assets (HtM_{LIQ}), as proposed by Kaplan et al. (2014).

Table 4: Hand-to-mouth groups using Zeldes and KVW definitions

	Not-H2M	H2M_{NW}	H2M_{LIQ}
Shares	52.55%	3.65%	43.80%
		By LIQ (KVW)	
By NW (Zeldes)		Not-H2M	H2M
Not-H2M		52.55%	36.74%
H2M		3.65%	6.06%

Notes: Sample is from AIDIS 2019 with a total sample size of 101,481.Income is from xgboost.

The top panel presents the distribution of households across three groups: non-hand-to-mouth (non-HtM), hand-to-mouth based on net worth (HtM_{NW}), and hand-to-mouth based on liquid asset constraints (HtM_{LIQ}). A majority of households (52.55%) are classified as non-HtM, 3.65% as HtM_{NW}, and 43.80% as HtM_{LIQ}, indicating that the liquidity-based definition captures a substantially broader segment of financially constrained households.

The lower panel cross-tabulates the HtM_{NW} and HtM_{LIQ} classifications. Among households not identified as HtM_{NW}, 52.55% are also not HtM_{LIQ}, while 36.74% are liquidity constrained. Among those classified as HtM_{NW}, 3.65% are not liquidity constrained, while 6.06% meet both criteria. These results suggest that the HtM_{LIQ} definition identifies a broader group of constrained households than the net worth-based approach proposed by Zeldes (1989).

B Alternative classifications of liquid and illquid assets

Table 5: Summary of HtM Types Under Different Scenarios(consumption bins)

	P-HtM	W-HtM	N-HtM	HtM	HtM-NW
Baseline(biweekly-pay)	0.031	0.165	0.803	0.197	0.034
1-year income credit limit	0.045	0.264	0.691	0.309	0.046
Weekly pay period	0.016	0.065	0.919	0.081	0.019
Monthly	0.052	0.322	0.626	0.374	0.055
Higher illiquid wealth cutoff	0.032	0.165	0.803	0.197	0.034
Business as illiquid asset	0.025	0.172	0.803	0.197	0.033
Direct as illiquid asset	0.031	0.166	0.803	0.197	0.034
Other valuables as illiquid asset	0.010	0.186	0.803	0.197	0.013
Financially fragile household	0.057	0.371	0.572	0.428	0.059

B.0.1 Income Imputations Results Using OLS

Table 6: OLS (Ln) Income Regression Results

	Coefficient	t-stat
Ln(Consumption)	0.978***	(9575.81)
Region	0.175***	(1566.77)
Gender	0.0539***	(409.42)
Age	0.0295***	(1356.76)
Age^2	-0.000231***	(-1066.97)
Household size	0.00189***	(134.84)
Education (ref: Illiterate)		
Below Primary	0.0327***	(151.17)
Primary	0.0522***	(242.85)
Upper Primary/Middle	0.0337***	(152.72)
Secondary/Higher Secondary	0.0773***	(356.50)
Graduate	0.235***	(966.33)
Post-Graduate +	0.444***	(1730.93)
Caste (ref: Scheduled Tribe)		
Scheduled Caste	0.0217***	(180.87)
Other Backward Class	0.0493***	(429.67)
Other	0.106***	(871.81)
Religion (ref: Hindu)		
Muslim	-0.0698***	(-684.74)
Christian	0.0632***	(263.11)
Sikh	0.164***	(801.12)
Jain	0.204***	(278.59)
Buddhist	-0.00805***	(-18.55)
Other	-0.109***	(-43.08)
Employment Type (ref: Urban, Self-Employed)		
Urban, Regular Salary	0.0904***	(805.50)
Urban, Casual Labor	-0.128***	(-963.14)
Rural, Self-Employed in Agriculture	0.143***	(1530.31)
Rural, Self-Employed in Non-Agriculture	0.0834***	(691.16)
Rural, Regular Salary	0.238***	(1674.66)
Rural, Casual Labor in Agriculture	0.0288***	(217.08)
Savings		
Bank Account	0.150***	(645.44)
Life Insurance	0.104***	(1685.96)
Fixed/Recurring Deposit	-0.0223***	(-369.08)
Post Office	0.0542***	(756.36)
Gold	-0.0721***	(-274.59)
Business	0.0387***	(250.80)
Borrowing Source		
Bank	0.112***	(1052.36)
Employer	0.225***	(430.01)
Relative/Friend	0.0397***	(366.45)
NBFC/MFI	0.0744***	(426.52)
Self Help Group	0.123***	(752.87)
Chit Fund	0.120***	(307.13)
Shop	-0.0901***	(-673.25)
Money Lender	0.0732***	(507.84)
Borrowing Purpose		
Education	-0.0261***	(-119.98)
Medical	-0.0787***	(-340.39)
Repayment	0.104***	(579.68)
Housing	-0.0464***	(-328.59)
Household expenses	-0.0179***	(-149.15)
Business investment	0.0108***	(80.17)
Constant	-0.208***	(-198.55)
Observations	181,699,747	
R-squared	0.58	
Adjusted R-squared	0.58	
Root MSE	0.37	

t-statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B.1 HtM Based on Zero Kink

Table 7: Share of HtM households based on zero kink definition

	P-HtM	W-HtM	HtM-NW
Consumption Bins	.0292	.1264	.0307
OLS	.0460	.1694	.1095
Decision Tree	.0274	.1196	.0288
Random Forest	.0263	.1172	.0274
Neural Network	.0417	.2077	.1047
XGBoost-Default	.0185	.0857	.0196
XGBoost-Fine Tuned	.0272	.1202	.0284

Table 8: APC by HtM Household based on the zero kink definition

	P-HtM	W-HtM	HtM-NW	Non P-HtM or W-HtM	Non-HtM-NW
Consumption Bins	.3398	.3370	.3398	.3470	.3457
OLS	.3314	.3532	.3273	.3846	.3835
Decision Tree	.4187	.4042	.4138	.4037	.4039
Random Forest	.4542	.4187	.4516	.4055	.4069
Neural Network	.3418	.3159	.3685	.3135	.3118
XGBoost-Default	.4701	.4322	.4643	.6268	.6039
XGBoost-Fine Tuned	.4264	.4057	.4240	.4017	.4022

B.2 HtM Based on Credit Limit as 1 Month of Income

Table 9: Share of HtM Household based on credit limit (1 month of income) definition

	P-HtM	**W-HtM**	**HtM-NW**
Consumption Bins	.02919467	.1263884	.030656215
OLS	.045996409	.16937523	.10947938
Decision Tree	.027440542	.1196468	.028789083
Random Forest	.026250588	.11722678	.027381785
Neural Network	.041698761	.20771289	.10468184
XGBoost-Default	.018541427	.085691281	.019565299
XGBoost-Fine Tuned	.027183721	.12018552	.028358581

Table 10: Share of HtM Household based on credit limit (1 month of income) definition

	P-HtM	**W-HtM**	**HtM-NW**
Consumption Bins	0.0292	0.1264	0.0307
OLS	0.0460	0.1694	0.1095
Decision Tree	0.0274	0.1196	0.0288
Random Forest	0.0263	0.1172	0.0274
Neural Network	0.0417	0.2077	0.1047
XGBoost-Default	0.0185	0.0857	0.0196
XGBoost-Fine Tuned	0.0272	0.1202	0.0284