Integrating Consumption and Asset Surveys: A Framework for Unified Household Balance Sheets

Vivek Gupta
Shiv Nadar Institution of Eminence

Fiorella Pizzolon
Hamilton College

Aditi Singh*
CAFRAL

Click here for latest version

Abstract

We propose a general framework for constructing complete balance-sheet data in environments where no single survey captures the full set of financial variables. The methodology aligns overlapping demographic and consumption variables, reconciles differences in survey timing, and imputes missing income using a suite of econometric and machine-learning models. The resulting unified dataset provides household-level measures of liquid and illiquid wealth, liabilities, income, and consumption, enabling analyses that would be infeasible using either survey alone. As an illustration, we apply the harmonized balance sheet to measure the prevalence and composition of financially constrained households following Kaplan et al. (2014) in India, documenting a sizable group of wealthy hand-to-mouth households who appear asset-rich but remain liquidity constrained. More broadly, the harmonization framework offers a scalable template for constructing balance sheets in other data-scarce settings and supports a wide range of applications in portfolio choice, financial vulnerability, and the distributional effects of macroeconomic policy.

*We gratefully acknowledge the excellent support provided by our research assistants, Ken Lam and Ashwath Damle, throughout this project. The views expressed here are solely the responsibility of the author and should not be interpreted as reflecting the views of the Center for Advanced Financial Reserach and Learning (CAFRAL) or the Reserve Bank of India.

1 Introduction

Understanding household balance sheets is central to analyzing consumption behavior, credit constraints, and the transmission of fiscal and monetary policy. These questions require simultaneous information on income, consumption, and assets. Yet in most developing economies, no single dataset provides such a unified view. Instead, researchers must rely on separate surveys that capture different components of the balance sheet, making it difficult to characterize household financial positions or to integrate micro data to inform macroeconomic models. In this paper, we address this challenge by constructing what is, to our knowledge, the first fully harmonized household balance sheet for a developing economy. Our approach combines the strengths of two nationally representative surveys and develops a replicable methodology for generating complete balance-sheet information in data-scarce environments.

We then demonstrate one application of this unified balance sheet: measuring the share of households that are financially constrained. This statistic is central to the calibration of modern macroeconomic models and plays a key role in policy and welfare analysis. Using the harmonized data, we identify financially constrained households, popularly called hand-to-mouth (HtM) households, and further characterize their liquidity positions following the framework of Kaplan et al. (2014). In contrast to existing work on emerging economies, which typically relies on coarse, single-survey proxies for liquidity and thus obscures substantial heterogeneity in household portfolios, our approach enables a precise distinction between liquid and illiquid wealth. By showing how two disparate surveys can be integrated into a coherent micro-level balance sheet, we illustrate the broader value of our scalable methodology for both descriptive analysis and macro-modeling applications in settings where unified household data are scarce.

India provides a compelling setting for implementing this approach. Despite high aggregate household savings, the vast majority of wealth is held in illiquid forms (land, housing, and gold), while access to liquid savings instruments and formal credit remains limited. These structural features make liquidity measurement particularly salient: a household's ability to smooth consumption depends less on its net worth and more on the accessibility and composition of its assets. A harmonized balance sheet allows us to capture this distinction rigorously and to examine forms of financial constraint that are invisible in datasets reporting only consumption or total assets.

To construct the harmonized balance sheet, we combine two complementary, nationally representative datasets: the All-India Debt and Investment Survey (AIDIS) and the Consumer Pyramids Household Survey (CPHS). AIDIS provides detailed information on household portfolios and liabilities but lacks income, while CPHS reports income and consumption but contains only binary indicators of asset ownership. We integrate these datasets by harmonizing shared demographic and consumption variables, and imputing missing income in AIDIS using a suite of econometric and machine-learning techniques. This procedure preserves the joint distribution of income, consumption, and wealth while ensuring consistency across

surveys. The resulting dataset offers, for the first time in a developing-country context, a unified view of household portfolios that is suitable for studying liquidity constraints, savings behavior, and policy transmission.

With the harmonized balance sheet in hand, we illustrate its usefulness by estimating the share of households that are financially constrained. Using the framework of Kaplan et al. (2014), we classify households into poor hand-to-mouth, wealthy hand-to-mouth, and unconstrained groups based on their liquid and illiquid asset positions relative to income. These classifications are widely used in the calibration of heterogeneous-agent macroeconomic models and are central to understanding the distributional transmission of fiscal and monetary policy. Applying our methodology to the Indian context reveals that poor HtM households account for 2–5% of households, while wealthy HtM households—those with substantial illiquid wealth but limited liquidity—comprise 15–27%. The harmonized balance sheet also allows us to examine consumption behavior. We document that average propensities to consume (APCs) are similar across HtM and non-HtM households, suggesting that financial constraints may be pervasive even among households with positive net worth. These estimates demonstrate the empirical value of harmonizing fragmented surveys by showing us patterns that would be obscured in single-survey analyses that do classifications based solely on net worth.

A harmonized household balance sheet of this kind is valuable far beyond the specific application to hand-to-mouth measurement. First, by jointly observing households' liquid and illiquid portfolios alongside income and liabilities, the dataset allows researchers to quantify the distributional channels of fiscal and monetary policy. Measures such as interest-rate exposure, cash-flow sensitivities, and borrowing constraints can be computed directly, enabling assessments of how policy shocks propagate across heterogeneous balance-sheet positions.

Second, the data allow for a granular characterization of financial vulnerability. Simple indicators—such as liquid assets relative to monthly consumption, debt-service ratios, and shock-absorption capacity—can be measured throughout the distribution, offering new insight into household resilience to unemployment, health shocks, or macroeconomic volatility. These measures can also be mapped across regions, caste groups, or occupations, highlighting pockets of vulnerability or financial fragility.

Third, the harmonized balance sheet supports empirical work on portfolio choice, credit frictions, and entrepreneurship. The dataset makes it possible to study how households allocate wealth between cash, deposits, gold, land, housing, and business assets, and how these allocations interact with access to formal and informal credit markets. This is particularly relevant in India, where illiquid wealth dominates and where gold and land play important roles in self-insurance, collateral, and informal finance.

Finally, the harmonized balance sheet enables structural macroeconomic work. The distribution of liquid and illiquid wealth, income dynamics, and borrowing constraints provides the necessary micro-foundations for calibrating Heterogeneous-Agent New Keynesian (HANK) models or other life-cycle models suited to emerging economies. This opens the

door to quantitative evaluation of fiscal stimulus, transfer programs, or monetary tightening in environments characterized by limited liquidity and high illiquid wealth. Together, these applications underscore that the harmonized balance sheet is not only a tool for measuring hand-to-mouth behavior; it is a general empirical resource that can support a broad research agenda in macroeconomic policy analysis, household finance, and development economics.

This paper makes three main contributions. First, it develops a novel, scalable methodology for constructing micro-level household balance sheets in data-scarce environments by integrating two nationally representative surveys using demographic harmonization and machine-learning–based income imputation. To our knowledge, this is the first implementation of such an approach in a developing-country context. Second, it provides the first unified household balance sheet for India, enabling precise measurement of liquid and illiquid wealth, liabilities, and income within a single framework. Third, using this dataset, we generate new evidence on the prevalence and composition of financially constrained households in developing countries, revealing a substantial group of wealthy hand-to-mouth households whose liquidity positions differ sharply from their net-worth positions. These contributions establish a foundation for future work on policy transmission, financial vulnerability, and heterogeneous-agent modeling in emerging economies.

Review of Literature: Our work contributes to three literatures. The first is the literature that develops methods for combining multiple household surveys or imputing missing balance-sheet components when no single dataset is comprehensive. Traditional approaches rely on statistical matching or regression-based imputations, but these methods often impose strong linearity assumptions and may fail to capture nonlinear relationships between income, consumption, and assets. Recent work has turned to machine-learning techniques to address such limitations, demonstrating improvements in predicting income, wealth, and demographic characteristics in data-scarce settings Wan (2023); Wang (2022); Kose et al. (2024). Nonetheless, applications of these tools to the construction of full household balance sheets remain rare, especially in developing countries where survey design differences complicate integration. Our approach builds on and extends this methodological literature by combining demographic harmonization with a suite of supervised learning algorithms (OLS, decision trees, random forests, neural networks, and XGBoost) to impute income and generate a unified micro-level balance sheet suitable for both descriptive analysis and macroeconomic modeling.

A second strand of literature in development and household finance examines how households in low and middle income countries allocate assets across liquid and illiquid forms, and how these portfolio choices shape vulnerability and inequality. Research has documented that households in developing economies tend to hold a large share of their wealth in illiquid assets—such as land, housing, and gold—with limited use of formal saving instrumentsRBI (2017); Dupas and Robinson (2013); Karlan et al. (2014). This portfolio structure reflects both institutional frictions, such as high transaction costs and shallow credit markets, and behavioral motives, including informal insurance and precautionary saving. At the same time, studies on wealth inequality emphasize that aggregate measures obscure substantial heterogeneity in asset composition and liquidity. However, because existing surveys typi-

cally measure only total assets or only consumption and income, they cannot jointly capture liquid buffers, illiquid holdings, and liabilities within the same household. As a result, empirical analyses of portfolio choice, financial vulnerability, or the distributional consequences of macroeconomic shocks often rely on incomplete balance-sheet information. By integrating two complementary surveys into a unified balance sheet, our dataset overcomes these limitations and enables a more accurate assessment of how asset composition, liquidity, and leverage interact in shaping household financial behavior.

A third strand of research applies the hand-to-mouth framework to countries outside the United States, including work on Japan Hara et al. (2016), China Cui and Feng (2017), Korea Song (2020), and several European economies Cherchye et al. (2023); Arroyo and Tisnés (2023). Studies of developing countries, such as Bracco et al. (2021), often find even higher prevalence of liquidity-constrained households, but these analyses typically rely on single surveys with limited information on portfolio composition or liquid wealth. Because these datasets conflate illiquid and liquid assets or measure only net worth, they provide an incomplete picture of household liquidity and may systematically misclassify households whose net worth is high but whose accessible resources are limited. The absence of unified balance-sheet microdata has therefore constrained the ability of researchers to apply modern heterogeneous-agent macro frameworks in emerging economies or to generate reliable cross-country comparisons. Our harmonized dataset addresses this gap by combining complementary surveys to produce a detailed balance sheet that captures precisely the liquid versus illiquid distinction needed for these analyses.

The remainder of the paper is structured as follows. Section 2 describes the harmonization strategy that can be used to integrate two surveys into a unified household balance sheet. Section 3 illustrates one application of the harmonized dataset by identifying financially constrained households and examining their liquidity positions in India. Section 4 discusses the broader set of analyses that the balance sheet enables and outlines several avenues for future research, and section 5 concludes.

2 Harmonization Methodology

Across low, middle, and even some high-income countries, household data systems are commonly built around two distinct types of surveys: consumption or income surveys, which measure household expenditures, earnings, and basic demographics; and asset or wealth surveys, which record the composition and value of financial and physical assets along with outstanding liabilities. For example, the United States relies on the Consumer Expenditure Survey (CEX) for consumption and the Survey of Consumer Finances (SCF) for wealth; Japan's National Survey of Family Income and Expenditure is complemented by the Keio Household Panel Survey for assets; China fields the China Family Panel Studies (CFPS) alongside the Household Finance Survey (CHFS); and many developing countries—including Indonesia, South Africa, and Ghana—conduct separate Living Standards Measurement Studies (LSMS) and specialized asset or agricultural modules. Because these surveys are designed

for different statistical purposes, they rarely contain overlapping information on all components of the household balance sheet.

The absence of a unified dataset poses challenges for research on portfolio choice, financial vulnerability, and the distributional transmission of macroeconomic policy. Consumption surveys measure flows but not the composition of wealth, whereas asset surveys measure stocks but typically omit income. Harmonizing these survey types offers a way to overcome this fragmentation by reconstructing a complete financial profile for each household. In this paper, we develop a systematic approach for merging a consumption survey and an asset survey into a unified balance sheet. Our framework harmonizes overlapping demographic and consumption variables, aligns asset information with income and expenditure data, and imputes missing income components using econometric and machine-learning methods. The result is a complete, internally consistent dataset containing liquid assets, illiquid wealth, liabilities, income, and consumption for each household, providing a general template that can be applied in any setting where survey systems are similarly fragmented.

The conceptual goal of harmonization is to combine these surveys in a way that preserves the underlying economic relationships among income, consumption, assets, and liabilities. Achieving this requires addressing three key sources of inconsistency. First, definitional inconsistencies must be minimized by aligning variables that appear in both surveys but differ in their coding, level of aggregation, or reference period. Second, temporal inconsistencies must be reconciled to ensure that asset stocks and income or consumption flows correspond to comparable points in time, particularly when one survey collects retrospective information. Third, because the value of the unified balance sheet depends on its ability to reflect realistic financial behavior, the harmonization procedure must preserve the joint distribution of balance-sheet variables rather than matching only unconditional means or totals.

Our approach follows these principles by harmonizing overlapping demographic and consumption variables, aligning survey reference periods, and imputing missing income using supervised learning techniques that exploit the rich covariate structure shared across surveys. Although our empirical implementation focuses on India, where the Consumer Pyramids Household Survey (CPHS) provides income and consumption and the All-India Debt and Investment Survey (AIDIS) provides asset and liability data, the underlying methodology is general. The subsections that follow describe how we operationalize definitional alignment, temporal matching, and income imputation to construct a unified household balance sheet.

2.1 Aligning Common Variables Across Surveys

The first step in constructing a harmonized household balance sheet is to identify and align the set of variables that are common to both the consumption survey and the asset survey. Although these surveys are typically designed for different statistical purposes, they often share a core set of demographic and socioeconomic characteristics that can serve as the basis for harmonization. These commonly include the age, gender, education, and employment status of the household head; household size and composition; and broad indicators of resi-

dence such as rural or urban location. Ensuring that these variables are defined consistently across surveys is essential for producing a coherent merged dataset.

Harmonizing common variables involves standardizing definitions, units of measurement, and categorical groupings. In many countries, consumption surveys record detailed expenditure categories, while asset surveys include only a subset of these items or report them in more aggregated form. Similarly, demographic variables may differ in their coding conventions or level of detail. The harmonization process requires constructing comparable categories across surveys, collapsing or expanding variable definitions as needed, and ensuring that missing or undefined responses are treated consistently.

In addition to demographics, consumption expenditures play an important role in harmonization because they are often the only monetary variable observed in both survey types. However, the scope of consumption differs systematically between surveys: consumption surveys typically include both purchased and imputed items such as home-produced goods, in-kind payments, or freely collected resources, whereas asset surveys may record only actual expenditures. To maintain comparability, we restrict attention to expenditure categories that are measured consistently across surveys and exclude items that are captured in only one instrument.

By aligning these common variables, we create a set of harmonized features that serve two purposes. First, they enable us to ensure that the samples drawn from each survey are comparable and reflect similar underlying populations. Second, they provide the conditioning variables necessary for imputing missing income or wealth components in later steps of the harmonization process. The next subsection discusses how we reconcile the temporal structure of the surveys to ensure that asset stocks and income flows correspond to the same underlying reference period.

2.2 Reconciling Temporal Differences

A central challenge in harmonizing consumption and asset surveys is that the two instruments often refer to different underlying time periods. Consumption or income surveys typically collect monthly or annual flow information based on short recall windows or high-frequency interviews. In contrast, asset surveys record the value of stocks at a specific point in time, often using retrospective questions that reference a fixed date or a particular agricultural or fiscal year. If these temporal differences are not addressed, the resulting merged dataset may pair income and consumption flows that do not correspond to the asset positions they are assumed to support.

To reconcile these inconsistencies, we align the reference periods used in each survey as closely as possible. When asset surveys report retrospective stock values, we match them to consumption or income information collected in the survey waves that are temporally closest to the reference date. When the consumption survey contains multiple waves per year, we use data from the wave or set of waves that bracket the asset reference period, thereby

reducing the mismatch between the timing of stocks and flows. In cases where the recall windows differ across surveys, we construct harmonized variables that reflect comparable time horizons, such as converting monthly expenditures into annual equivalents or adjusting asset values to the appropriate reference period.

Temporal alignment is particularly important for variables that change rapidly over time, such as liquid financial assets, outstanding debt, or short-term income fluctuations. By synchronizing the timing of stock and flow variables, we ensure that the resulting balance sheet represents a coherent financial snapshot of each household. This alignment also improves the performance of subsequent imputations by reducing measurement error arising from temporal mismatches.

2.3 Imputing Income

Even after harmonizing common variables and aligning reference periods, a key component of the household balance sheet typically remains unobserved in asset surveys: income. Because income is essential for constructing liquidity measures, computing consumption-to-income ratios, and assessing financial constraints, its absence must be addressed in a systematic and transparent manner. We adopt a supervised learning approach in which income reported in the consumption survey serves as the target variable, and the set of harmonized demographic, socioeconomic, and consumption variables serves as the predictor space.

We begin with a set of econometric models, such as ordinary least squares regressions of log income on observed household characteristics. These models provide interpretable benchmarks and capture linear relationships between income, consumption, and demographics. However, income determination often reflects nonlinearities, interactions, and threshold effects that are difficult to model parametrically. To account for these complexities, we complement traditional regressions with a suite of machine-learning algorithms, including decision trees, random forests, gradient-boosted decision trees, and neural networks. These methods are well suited to capturing flexible functional forms and identifying heterogeneous predictors of income across the population.

For each model, we train the supervised learner on the consumption survey data, using cross-validation to tune hyperparameters and evaluate predictive performance. The resulting models generate out-of-sample predictions of income for households in the asset survey, conditional on the harmonized covariates shared across the two surveys. To avoid relying on any single specification, we compare the performance of the full set of models using standard metrics such as root mean squared error.

The imputed income values produced through this procedure complete the set of variables necessary to construct a full household balance sheet. By relying on supervised learning, rather than ad hoc ratios or coarse consumption bins, we preserve the conditional relationships between income, consumption, and wealth—an essential feature for studying liquidity constraints, portfolio choice, and the distributional impacts of macroeconomic policy.

2.3.1 Consumption Bins

Our first approach to imputing household income in the asset survey 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 to income, that is, $\text{APC} = C/Y$. This measure reflects the share of current income that households devote to consumption over a given reference period.

In many countries, consumption and asset surveys overlap in their reporting of household expenditures, even if only one of the surveys contains income information. We exploit this overlap by sorting households in both surveys into a common set of consumption-based bins. Specifically, we rank households by total consumption within each survey and partition the distributions into one hundred bins that are defined identically across surveys. In the consumption survey, where both income and consumption are observed, we compute the APC for each bin by averaging the APCs of all households belonging to that bin.

To recover income in the asset survey, which reports consumption but lacks income, we assign to each household the APC corresponding to its consumption bin and apply the inverse relationship $Y = C/\text{APC}$. This procedure provides a simple, fully nonparametric imputation of income that relies only on the stable association between income and consumption within similar expenditure groups. While this approach abstracts from richer covariate information, it serves as a useful benchmark against which more flexible imputation methods can be compared.

2.3.2 Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) regression Greene (2018); Wooldridge (2019) provides a baseline econometric approach for imputing income in the asset survey. OLS estimates the conditional expectation of income given observable household characteristics by fitting a linear model that minimizes the sum of squared residuals between observed and predicted values. The resulting coefficient estimates capture the marginal association between each predictor and the dependent variable, holding other factors constant.

OLS is widely used due to its transparency and interpretability. Each regression coefficient has a clear economic meaning, and the framework allows for straightforward hypothesis testing and diagnostic evaluation. However, the approach relies on key assumptions, including linearity in parameters and additive separability of effects. When the true relationship between income and household characteristics is nonlinear or involves complex interactions, OLS may yield biased or inefficient predictions. Additionally, multicollinearity among regressors can inflate standard errors, reducing the precision of estimates. For these reasons, OLS serves as a useful benchmark for income imputation, while more flexible machine-learning models may better capture heterogeneous and nonlinear income dynamics in practice.

2.3.3 Decision Trees

A decision tree Mienye and Jere (2024) is a nonparametric supervised learning method that recursively partitions the predictor space based on feature values. At each internal node, the algorithm selects a splitting rule that best separates the data according to a chosen criterion, such as minimizing prediction error or impurity. The final leaves of the tree correspond to predicted income values for households that share similar characteristics. Because the method relies on simple threshold-based rules, it can flexibly capture nonlinear relationships and interactions among variables without requiring prior feature scaling or transformation.

Decision trees are attractive for their interpretability: the resulting tree structure provides a transparent sequence of decision rules that illustrate how predictions are generated. However, they also have notable limitations. A single tree can be highly sensitive to outliers, may exhibit strong preferences for variables with many possible split points, and often overfits the training data, leading to poor generalization performance. For these reasons, pruning strategies or restrictions on tree depth are commonly used to limit model complexity. When the number of predictors is large, dimensionality reduction or feature-selection techniques can be employed to improve computational efficiency and model stability.

To evaluate the performance of the decision tree model, we use cross-validation and tune hyperparameters such as tree depth, minimum samples required for splits, and minimum leaf size. Hyperparameters are selected using procedures such as grid search, random search, or Bayesian optimization, with model performance assessed using metrics such as root mean squared error (RMSE) and the coefficient of determination (R^2). We also compute feature importance measures to assess the relative contribution of different predictors to the model's performance, which aids interpretability in applications where nonlinearities and interactions are important.

In practice, decision trees serve as a flexible benchmark for income imputation in the harmonization framework, offering a balance between interpretability and the ability to capture nonlinear structures in the data.

2.3.4 Random Forest

Random forests Biau (2012) are an ensemble learning method that extends the decision-tree framework by aggregating the predictions of many trees. A single decision tree recursively partitions the predictor space using threshold-based rules, but such trees are prone to overfitting and can be sensitive to small changes in the data. Random forests address these limitations by training multiple trees on bootstrapped samples of the data and, at each split, considering only a random subset of predictors. The final prediction is obtained by averaging (for regression) or voting (for classification) across trees, yielding models that are substantially more robust and accurate than individual trees.

By combining the strengths of bagging and feature subsampling, random forests can capture

complex, nonlinear relationships in the data while reducing sensitivity to outliers, multicollinearity, and sampling variability. The method requires no prior standardization of predictors and provides measures of feature importance that summarize the contribution of each variable to predictive performance. A limitation of random forests is reduced interpretability relative to single decision trees, as the ensemble structure makes it difficult to trace individual prediction paths. In addition, training large ensembles can be computationally demanding, particularly when the number of predictors is high.

We evaluate the performance of the random forest model using k-fold cross-validation, in which the data are partitioned into k folds and the model is trained repeatedly on $k-1$ folds while being validated on the remaining fold. This procedure provides stable estimates of out-of-sample performance and helps guard against overfitting. Hyperparameters—including the number of trees, the maximum depth of each tree, the minimum number of observations required to split a node or form a leaf, and the number of predictors considered at each split—are tuned using search procedures such as grid search, random search, or Bayesian optimization. The latter uses a probabilistic surrogate model to guide the search efficiently toward high-performing configurations.

Model performance is assessed using metrics such as root mean squared error (RMSE) and the coefficient of determination (R^2). We also compute feature importance measures derived from the ensemble of trees. These measures summarize how often and how effectively each predictor contributes to splitting rules across the forest, offering insight into the relative importance of different household characteristics for income prediction. Random forests therefore provide a flexible and powerful tool for imputing income within the harmonized household balance-sheet framework.

2.3.5 Neural Networks

Neural networks Warner and Misra (1996) are flexible supervised learning models composed of interconnected layers of nodes (or neurons) that transform input variables through successive nonlinear operations. A typical feedforward neural network consists of an input layer that receives the predictors, one or more hidden layers that apply nonlinear transformations, and an output layer that produces a prediction. The network is trained by comparing predicted values to actual outcomes using a loss function and iteratively adjusting the weights on each connection through backpropagation to minimize this loss.

Neural networks are well suited for capturing complex, nonlinear relationships and interactions among covariates that may be difficult to model parametrically. Their flexibility, however, comes with several challenges. Because the functional form is learned rather than specified, neural networks are often less interpretable than traditional regression-based approaches. Training can also be computationally intensive, particularly when the number of predictors or hidden layers is large. Achieving good predictive performance typically requires careful selection of hyperparameters such as the number of layers and neurons, the choice of activation and loss functions, regularization parameters, and optimization algorithms.

Techniques such as dropout, batch normalization, and early stopping are commonly used to prevent overfitting and improve generalization.

In the context of income imputation, a fully connected feedforward neural network provides a flexible alternative to linear or tree-based models. The network takes the harmonized household characteristics as inputs and learns nonlinear mappings from these covariates to income. Model performance is evaluated using metrics such as the root mean squared error (RMSE), which summarizes the average magnitude of prediction error, and cross-validation procedures are used to assess robustness. Although neural networks require more computational resources and tuning effort, they can yield substantial gains in predictive accuracy when income depends on complex patterns in the data.

2.3.6 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.

2.4 Constructing Measure of Liquid and Illiquid Wealth

With income imputed and common variables harmonized, the next step is to construct a set of liquid and illiquid wealth measures that reflect the full structure of the household balance sheet. Asset surveys typically record a detailed breakdown of financial and physical assets, along with outstanding liabilities. However, these categories vary across countries in terms of coverage and granularity, and the distinction between liquid and illiquid assets is not always explicit in the underlying survey design. Our goal is to classify assets in a way that is economically meaningful and consistent with the literature on liquidity constraints and portfolio behavior.

Liquid assets include resources that can be readily used to finance current consumption or

buffer short-term income shocks. These typically consist of cash holdings, demand deposits, money-market instruments, short-term savings accounts, and easily tradable financial securities. Illiquid assets, by contrast, include those that are costly or time-consuming to convert into consumption. Common examples are housing, land, business equity, durable goods, long-term financial instruments, and pension or provident fund balances. Liabilities are classified symmetrically according to whether they are short-term obligations, such as consumer loans, or long-term commitments, such as mortgage or business debt.

To ensure comparability across countries, we adopt a two-step classification procedure. First, we map each asset category in the survey to a standardized taxonomy that distinguishes financial from non-financial assets, and within financial assets, separates short-term liquid instruments from long-term or restricted accounts. Second, we verify these classifications using institutional knowledge about the liquidity of each asset type in the relevant context, including withdrawal restrictions, transaction costs, and the depth of secondary markets. This approach allows the harmonized dataset to reflect both theoretical definitions and country-specific institutional features.

The resulting liquid and illiquid wealth measures, combined with the imputed income and harmonized consumption variables, provide a complete household balance sheet. These measures serve as the basis for analyzing liquidity constraints, assessing financial vulnerability, and studying portfolio choice within and across countries.

2.5 Validating the Harmonized Balance Sheet

A crucial step in the harmonization process is validating that the resulting balance sheet is internally consistent and externally credible. Because the unified dataset combines information from two independent surveys and relies on imputed components, it is essential to assess the extent to which the merged variables align with known aggregates, reproduce key empirical relationships, and reflect plausible household financial behavior.

We implement three layers of validation. The first focuses on internal consistency. We verify that the constructed balance sheet satisfies basic accounting identities, such as the additive decomposition of total wealth into liquid and illiquid components and the consistency between asset and liability definitions. We also examine whether income, consumption, and wealth exhibit expected correlations within the harmonized dataset, ensuring that the imputation procedure preserves underlying economic relationships rather than generating mechanical artifacts.

The second layer of validation assesses the external credibility of the harmonized variables by comparing their distributions to national aggregates and benchmarks from administrative or survey-based sources. For example, we compare the distribution of total assets, liabilities, and net worth to national wealth accounts or financial sector reports; similarly, we check that the consumption and income distributions are broadly consistent with those reported in other large-scale surveys. While differences across data sources are inevitable due to survey

design and sampling variation, close alignment in broad patterns provides assurance that the harmonization has not introduced systematic distortions.

The third layer of validation examines the robustness of the imputed income values. We compare predictions across the full set of econometric and machine-learning models, analyze the sensitivity of results to alternative model specifications, and evaluate stability across subsamples defined by demographic or socioeconomic characteristics. This robustness analysis ensures that the merged dataset does not depend critically on any single imputation method and that key results are not driven by model-specific features.

Together, these validation exercises confirm that the harmonized balance sheet provides a reliable and coherent representation of household financial positions. This foundation allows the dataset to be used confidently for applications ranging from the measurement of liquidity constraints to the analysis of portfolio choice, financial vulnerability, and the distributional effects of macroeconomic policy.

3 Application: Measuring Financially Constrained Households

One advantage of constructing a unified household balance sheet is that it allows researchers to study financial constraints using measures that require joint information on income, consumption, and the liquidity of household portfolios. In this section, we illustrate the usefulness of the harmonized dataset by applying it to the classification of financially constrained, or hand-to-mouth, households. Following the framework of Kaplan et al. (2014), we distinguish between households with low or negative liquid asset positions relative to income and those whose illiquid wealth masks underlying liquidity constraints. This application provides a natural demonstration of how the merged balance sheet enables richer and more accurate measurement of financial vulnerability than would be possible with either survey alone.

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. We follow the conceptual framework proposed by Kaplan et al. (2014), which classifies households according to their holdings of liquid and illiquid wealth. Liquid wealth refers to assets that can be readily accessed at low cost, such as cash holdings, demand deposits, and other short-term financial instruments. Illiquid wealth, by contrast, includes assets that are costly or time-consuming to convert into consumption, such as housing, land, business equity, and long-term financial accounts.

A household is considered financially constrained, or hand-to-mouth, if its liquid asset posi-

tion is insufficient to buffer even small deviations in income or expenditure. Formally, this occurs when liquid wealth falls below a threshold proportional to current income, or when the household holds negative liquid balances and is at or near its effective credit limit. Within this group, the framework distinguishes two types of households. Poor hand-to-mouth households hold little or no illiquid wealth and have limited long-term resources to draw upon. Wealthy hand-to-mouth households, in contrast, possess substantial illiquid assets but remain liquidity constrained because these assets cannot be easily accessed to finance short-run consumption needs.

Mathematically, 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.

Implementing this classification requires joint information on liquid assets, illiquid wealth, liabilities, income, and consumption—precisely the set of variables reconstructed in the harmonized balance sheet. By combining the flow information from the consumption survey with the stock information from the asset survey, we can measure liquidity constraints in a way that would not be possible using either data source alone. The next subsection describes the specific thresholds and empirical procedures used to operationalize these definitions.

3.2 Operationalizing Liquidity Thresholds and Credit Limits

To implement the hand-to-mouth classification empirically, we must translate the conceptual definitions of liquidity constraints into a set of operational thresholds. Following Kaplan et al. (2014), we assess whether a household's liquid asset holdings are sufficient to finance short-term consumption needs or small income shocks. A common benchmark in the literature is to compare liquid wealth to a fraction of current disposable income, reflecting the notion that

households tied to a short pay cycle may be unable to smooth even temporary fluctuations in earnings. Households with liquid wealth below this threshold are classified as constrained. For households with negative liquid balances, we additionally consider whether they appear to be at or near their effective credit limit, defined relative to income or typical borrowing amounts in the relevant institutional environment.

These thresholds require assumptions about the pay frequency or budgeting horizon used to evaluate liquidity. For example, a monthly pay cycle implies that liquid asset holdings must cover approximately one month of income in order to avoid short-run liquidity stress, while a biweekly or weekly cycle implies a correspondingly lower threshold. Because these assumptions may vary across countries or survey settings, we adopt a flexible approach in which alternative pay-period conventions and credit-limit assumptions can be used in robustness exercises.

The distinction between poor and wealthy hand-to-mouth households is determined by the level of illiquid wealth. Once households are identified as liquidity constrained based on their liquid asset position, we classify them as poor hand-to-mouth if their illiquid wealth is negligible or zero, and as wealthy hand-to-mouth if their illiquid holdings exceed this threshold. This distinction is important for understanding the broader financial position of constrained households, as those with substantial illiquid assets may still be vulnerable to short-term shocks despite appearing financially secure on the basis of net worth.

By applying these rules to the harmonized balance sheet—which jointly measures liquid wealth, illiquid assets, liabilities, and income—we obtain a consistent classification of financially constrained households. The next subsection presents the empirical results, including the prevalence of poor and wealthy hand-to-mouth households and the sensitivity of these measures to alternative threshold assumptions.

3.3 Combining Consumption and Asset Data

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 3 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.

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

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

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.

3.4 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.

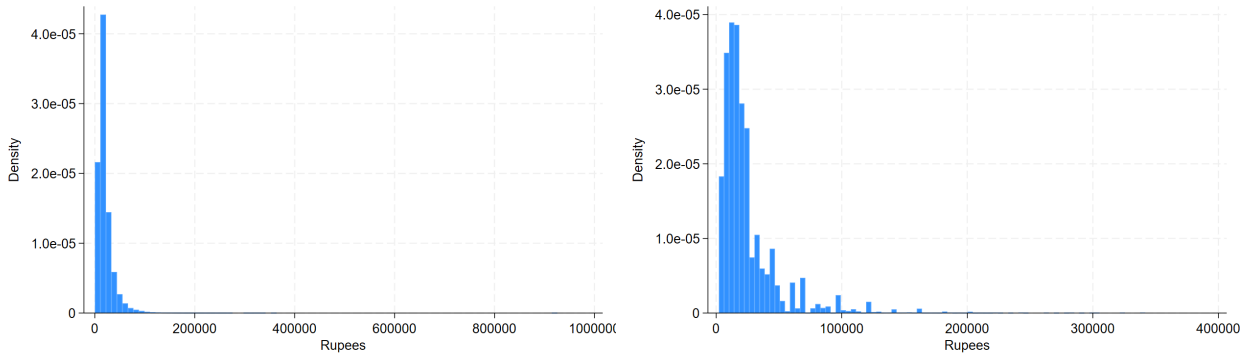
3.4.1 Consumption Bins

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.

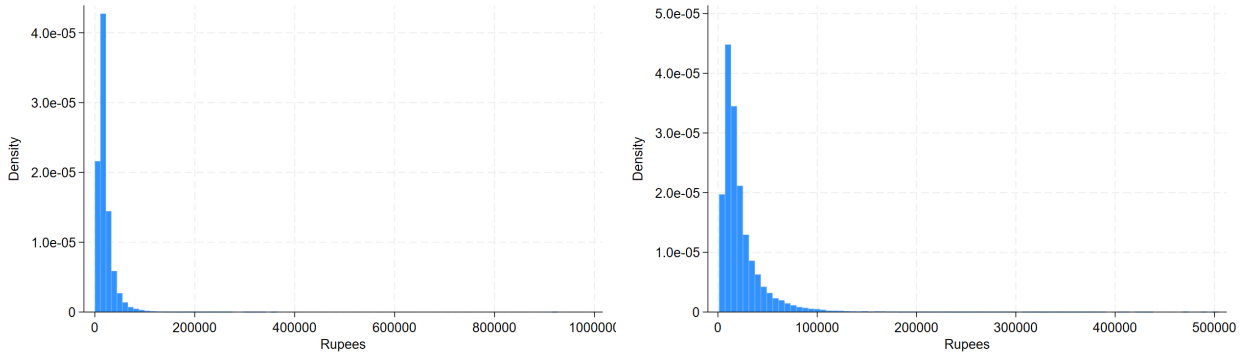
Figure 1: Monthly Household Income Distributions using Consumption Bins



(a) CPHS (b) Predicted in AIDIS using consumption bins

3.4.2 Ordinary Least Squares (OLS)

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.

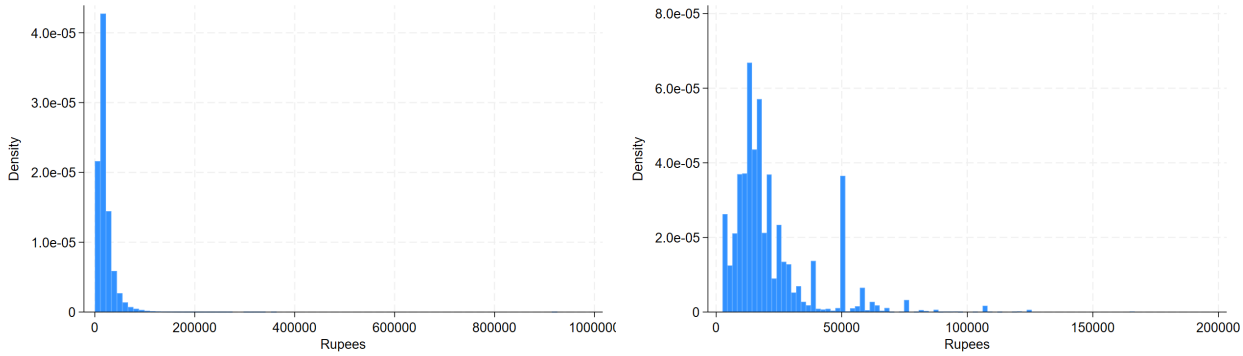
The effects of financial variables on income are more mixed. Ownership of savings is generally associated with higher income, although savings in fixed deposits and gold are exceptions, showing negative or negligible effects. Borrowing from various sources is typically positively associated with income, 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.

3.4.3 Decision Tree

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

3.4.4 Random forest

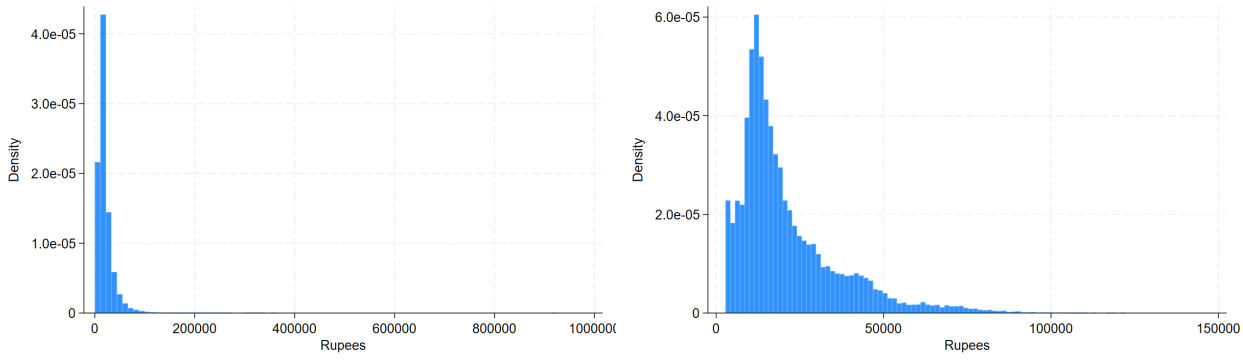
Model evaluation for the random forest income imputation was conducted using five-fold cross-validation on the CPHS sample. The CPHS dataset was partitioned into five equally sized folds; in each iteration, four folds were used to train the model and the remaining fold was used for validation. This procedure ensures that every household in CPHS contributes to both training and validation, providing a reliable estimate of out-of-sample predictive performance and reducing the risk of overfitting to idiosyncrasies in the CPHS data.

Hyperparameters were selected using 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 root mean squared error (RMSE) and the coefficient of determination (R^2), both computed on the CPHS validation folds. We also examined feature importance scores from the fitted random forest, which highlight the predictors most informative for income determination in the Indian context.

Figure 4 compares the distribution of actual monthly household income in CPHS with the corresponding income predictions generated by the tuned random forest model. The left panel shows the observed CPHS income distribution; the right panel shows the predicted distribution for the same households. The RMSE of 0.32 indicates that, on average, predicted incomes deviate from true incomes by roughly 32 percent, demonstrating that the model achieves a high degree of predictive accuracy in the Indian setting.

Figure 4: Monthly Household Income Distributions using Random Forest



(a) CPHS

(b) Predicted in AIDIS using random forest

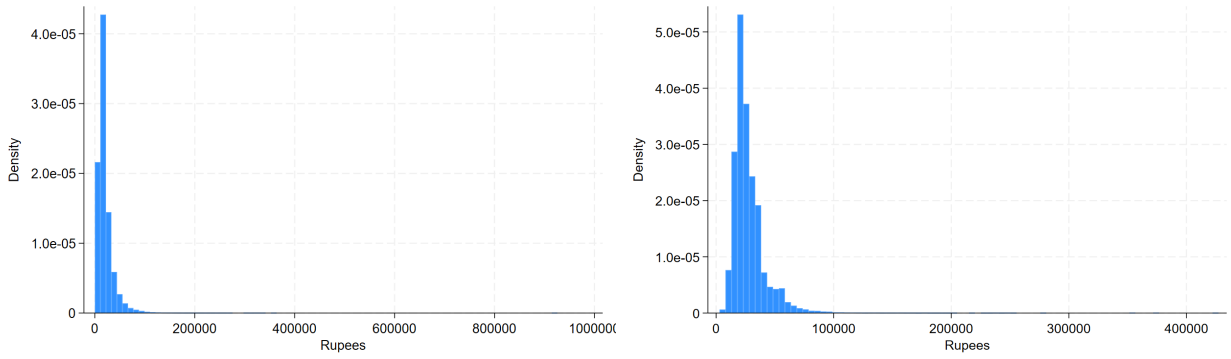
3.4.5 Neural Network

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

3.4.6 Extreme Gradient Boosting (XGBoost)

In our India-specific implementation, the XGBoost model was trained on the CPHS sample using five-fold cross-validation to ensure that the income predictions would generalize well beyond the households observed in the training folds. CPHS was partitioned into five approximately equal subsets; in each iteration, four folds were used to train the model and the remaining fold served as the validation set. This procedure is particularly important in the Indian context, where household income distributions are highly skewed and exhibit substantial regional and occupational heterogeneity, making robustness to sample variation essential.

Hyperparameter tuning combined Bayesian optimization and random search. Bayesian optimization was used to efficiently explore the hyperparameter space by constructing a surrogate probabilistic model, typically based on a Gaussian process or tree-structured Parzen estimator, that identifies regions of the parameter space with high expected improvement. This approach is well suited to the nonlinear and high-dimensional relationships present in CPHS income data. Complementing this, random search drew hyperparameter combinations uniformly from predefined ranges, providing broad coverage and reducing the likelihood of missing high-performing configurations in parts of the space that Bayesian optimization might explore less frequently.

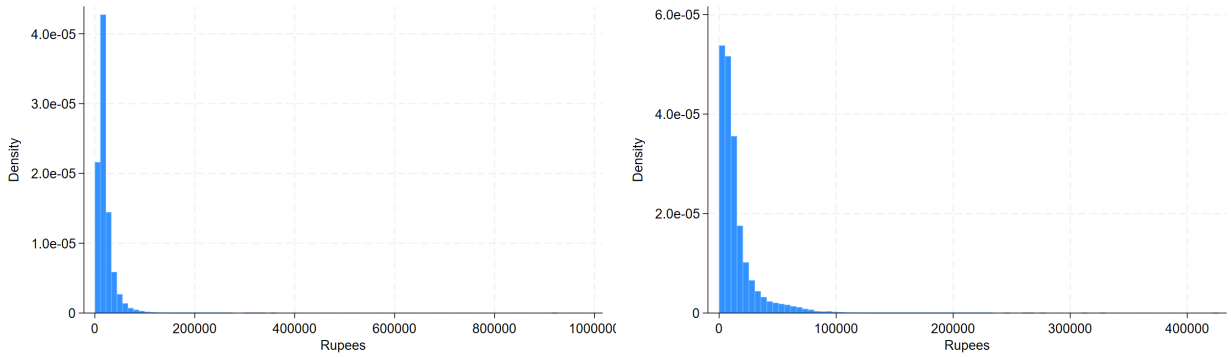
Model performance was evaluated using two standard metrics: the root mean squared error (RMSE), which captures the average magnitude of prediction error in the same units as household income, and the coefficient of determination (R^2), which measures the proportion of income variation explained by the model. These metrics provide a comprehensive evaluation of predictive accuracy in the CPHS sample. To aid interpretability, we also examined XGBoost's feature importance scores, which summarize how frequently and how effectively each predictor contributes to splits across the ensemble of boosted trees. In the Indian case, variables related to consumption expenditure, education, household size, and urban location emerged as key drivers of predicted income, consistent with well-established patterns in Indian labor-market and household-finance data. The resulting tuned XGBoost model was then used to impute income for households in the AIDIS survey, enabling us to complete the harmonized balance sheet for the Indian population.

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

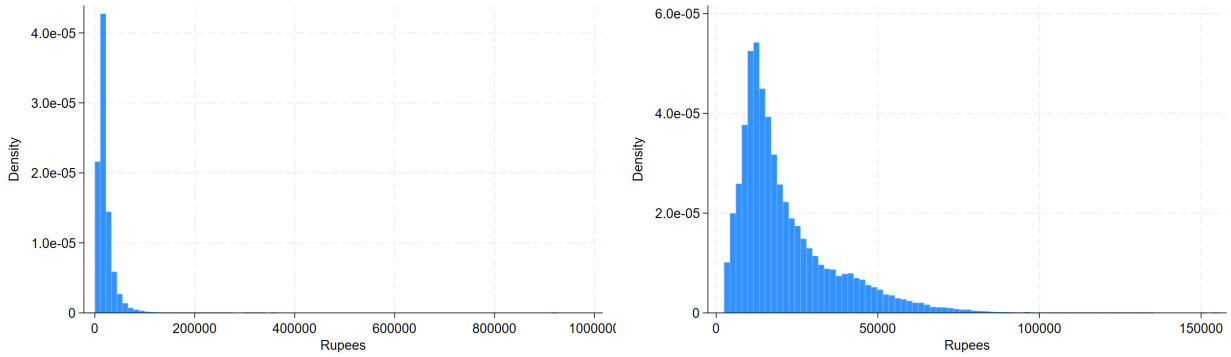
CPHS, while the right panel illustrates the distribution of predicted income using the XGBoost model described above. Figure 6 corresponds to the default model, while Figure 7 corresponds to the fine tuned model. The RMSE of 0.32, in both cases, indicates that, on average, the predicted incomes deviate from actual incomes by approximately 32%, 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

3.5 Results: Prevalence and Composition of Financially Constrained Households

Applying the liquidity-threshold rules to the harmonized household balance sheet yields a clear picture of the prevalence and composition of financially constrained households. Because the classification relies jointly on liquid assets, illiquid wealth, income, and liabilities, it captures heterogeneity that would be invisible in either the consumption survey or the asset survey alone. Across the full sample, we find that a nontrivial share of households have liquid

asset positions that fall short of even modest short-run smoothing needs, placing them in the hand-to-mouth category under standard threshold assumptions.

Disaggregating this group in table 1 reveals substantial variation in underlying financial positions. 2% to 5% of households hold little or no illiquid wealth and are therefore classified as poor hand-to-mouth. These households face both short-term and long-term financial limitations, lacking liquid buffers as well as meaningful asset holdings that could support future consumption or borrowing. In contrast, a larger fraction of households, between 15% and 27%, possess significant illiquid assets—such as housing, land, or business capital—but remain liquidity constrained because these assets cannot be accessed quickly or without substantial transaction costs. These wealthy hand-to-mouth households highlight the importance of distinguishing between net worth and liquidity when measuring financial vulnerability.

Table 1: 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 2 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 2: 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

3.5.1 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.

4 Additional Applications of the Harmonized Balance Sheet

While the hand-to-mouth classification illustrates one useful application of the harmonized balance sheet, the unified dataset supports a much broader set of analyses in household finance and macroeconomic policy. Because the harmonized data jointly observe income, consumption, liquid and illiquid assets, and liabilities, they enable researchers to pose questions that cannot be addressed using either survey alone. This section outlines several classes of applications to highlight the versatility of the harmonization framework.

A first set of applications concerns portfolio choice and wealth inequality. The balance sheet allows for a detailed decomposition of household wealth into its liquid and illiquid components, enabling analyses of how households allocate resources among cash, deposits, housing, land, business assets, and long-term financial instruments. This decomposition provides a basis for studying the determinants of portfolio composition, the role of illiquid assets in informal insurance or precautionary saving, and the contribution of different asset types to overall wealth inequality. Because the balance sheet captures both stocks and flows, it also allows researchers to study how incomes, liabilities, and asset choices interact over the life cycle or across demographic groups.

A second set of applications relates to the measurement of financial vulnerability and shock exposure. Liquid-wealth-to-consumption ratios, debt-service burdens, and other balance-sheet indicators can be used to assess households' ability to withstand income shocks, health shocks, or macroeconomic volatility. Spatial or demographic mapping of these indicators can highlight pockets of financial fragility within a country and inform the design of social protection programs. Moreover, the distinction between liquid and illiquid assets allows for a more nuanced understanding of households that appear financially secure on the basis of total wealth but remain highly exposed to short-term shocks.

A third set of applications centers on the distributional transmission of macroeconomic policy. The harmonized balance sheet contains the components needed to estimate interest-rate exposure, cash-flow sensitivities, and the distribution of liquid buffers—all of which are key inputs in modern heterogeneous-agent macroeconomic models. These models emphasize how monetary and fiscal policies affect households differently depending on the structure of their balance sheets. The harmonized dataset provides the empirical foundation for calibrating such models and for assessing how policy interventions propagate through households with varying liquidity positions and asset portfolios.

Finally, the harmonization framework itself is portable and can be applied in any setting where consumption and asset surveys are collected separately. Many low- and middle-income countries field survey systems similar to those described in this paper, making the methodology relevant beyond a single context. By outlining a general approach to integrating fragmented microdata, the framework offers a template for constructing unified balance sheets that support cross-country comparisons and a wide range of policy-relevant analyses.

In the next section, we present a series of robustness checks that demonstrate the stability of the hand-to-mouth classification and the reliability of the harmonized balance sheet under alternative modeling choices and definitions.

5 Conclusion

This paper develops a general framework for harmonizing consumption and asset surveys in order to construct unified household balance sheets in settings where no single dataset con-

tains all relevant financial information. By aligning common variables, reconciling temporal differences, imputing missing income using supervised learning, and constructing consistent measures of liquid and illiquid wealth, we show how two distinct surveys can be combined into a coherent dataset that preserves the joint distribution of income, consumption, and wealth. Although our empirical implementation draws on a particular pair of surveys, the methodology is broadly applicable to other countries where consumption and asset data are collected separately.

We illustrate the usefulness of the harmonized balance sheet through an application to the measurement of financially constrained, or hand-to-mouth, households. This exercise highlights how unified balance-sheet data reveal patterns of liquidity constraints that would be obscured in either survey alone, including the presence of households with substantial illiquid wealth but limited liquid buffers. The results demonstrate the value of distinguishing liquidity from net worth when assessing financial vulnerability, calibrating macroeconomic models, or designing targeted policy interventions.

More broadly, the harmonized dataset supports a wide range of applications beyond the hand-to-mouth classification. It enables detailed analyses of portfolio choice, wealth inequality, and financial fragility; provides the inputs required to study the distributional transmission of fiscal and monetary policy; and offers a foundation for calibrating heterogeneous-agent macroeconomic models in environments where administrative microdata are limited. The harmonization framework is portable across countries and can serve as a template for constructing comparable balance-sheet datasets in other settings.

Future work may build on this framework in several directions. One avenue is to extend the harmonized balance sheet longitudinally by incorporating additional survey rounds or panel structures, enabling the study of dynamic portfolio adjustments and income shocks. Another is to link the balance sheet to administrative, credit bureau, or geospatial data to deepen analyses of financial behavior and vulnerability. By providing a transparent and flexible methodology for integrating fragmented household surveys, this paper aims to support a broader agenda on the measurement and analysis of household financial positions in developing economies.

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.

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).

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

Dupas, P. and Robinson, J. (2013). Why don't the poor save more? evidence from health savings experiments. *American Economic Review*, 103(4):1138–1171.

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

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

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

Karlan, D., Osei, R., Osei-Akoto, I., and Udry, C. (2014). Agricultural decisions after relaxing credit and risk constraints. *The Quarterly Journal of Economics*, 129(2):597–652.

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 Details of the Data

This section describes the details of the two datasets used in our application - AIDIS and CPHS.

A.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 3: 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."

A.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.

B 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).

C 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

C.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

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

C.2 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

C.3 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