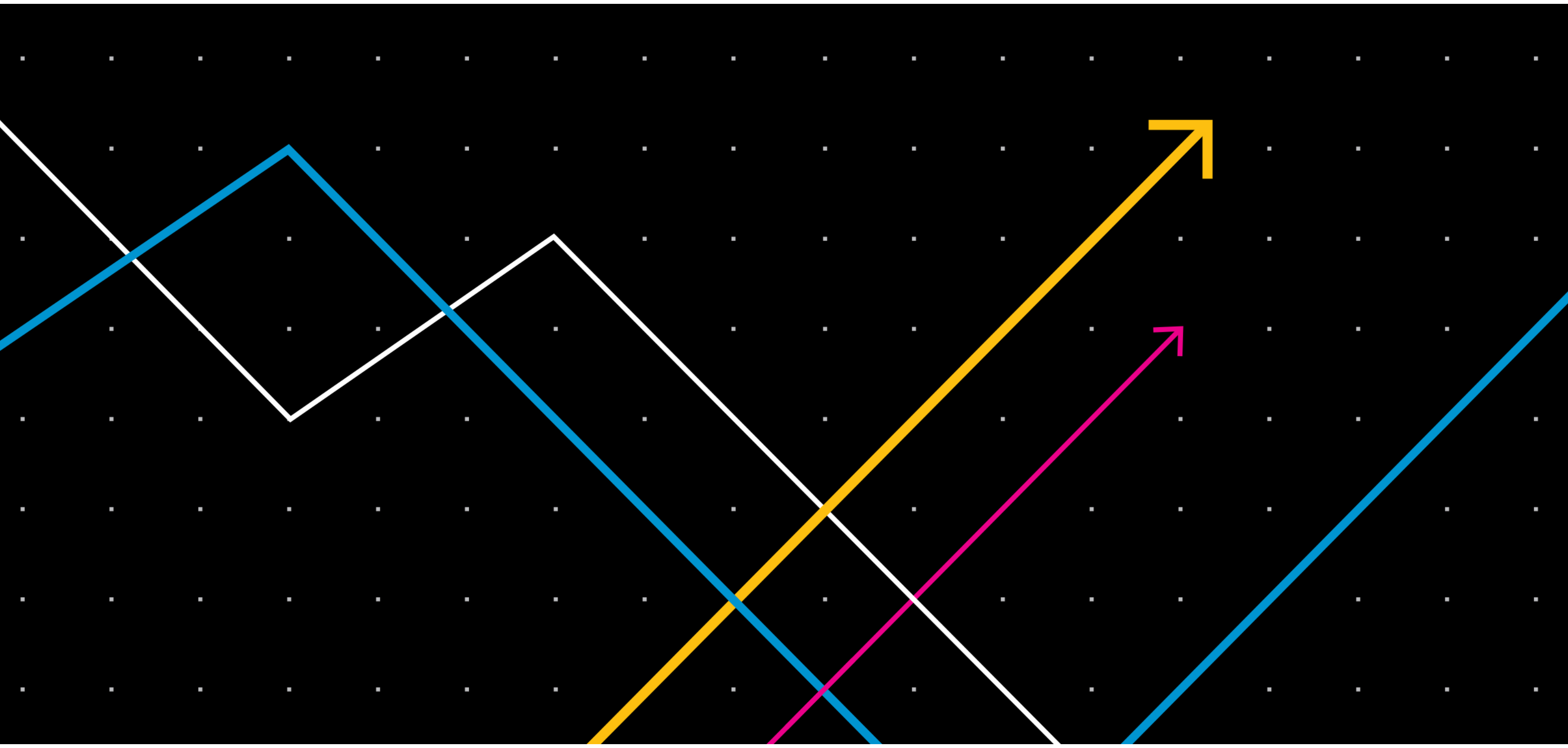
WEALTH AND FINANCIAL WELL-BEING



RESEARCH REPORT

Using Machine Learning to Estimate Racially Disaggregated Wealth Data at the Local Level

Aaron R. Williams *Mingli Zhong* *Breno Braga*

February 2023

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

Contents

Acknowledgments iv

Using Machine Learning to Estimate Racially Disaggregated Wealth Data at the Local Level 1

Data Sources 2

2018 Survey of Income and Program Participation 3

2019 American Community Survey 3

Methodology 4

Analytical Approach 4

Analytical Process 6

Generating Statistics by Race and Ethnicity 13

Results

Liquid Assets 14

Net Worth 19

Unequal Opportunities by Race and Ethnicity 26

Conclusion and Discussion 28

Appendix A. CFED (Corporation for Enterprise Development) Imputations 32

Appendix B. Model Descriptions 34

Notes 36

References 37

About the Authors 38

Statement of Independence 39

Acknowledgments

This report was funded by a grant from JPMorgan Chase & Co.. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

We are grateful to Signe-Mary McKernan, Alena Stern, Oriya Cohen, and Courtney Jones for their input; Alex Carther for excellent technical assistance; Madeline Pickens and Vincent Pancini for their thorough code review; and Liza Hagerman for her copyedits.

Using Machine Learning to Estimate Racially Disaggregated Wealth Data at the Local Level

Household wealth data at the local level are generally not widely available, especially statistics disaggregated by race and ethnicity. Using machine learning, we estimate net worth and emergency savings data at the local, city, state, and national levels. We also disaggregate our estimates by racial and ethnic groups at the city, state, and national levels. We use machine-learning models trained on the Survey of Income and Program Participation (SIPP), a survey with detailed wealth information but too few observations for local estimates, to estimate emergency savings and net worth using the American Community Survey (ACS), which allows for state and local estimation. The imputations allow us to estimate the proportion of households with more than $2,000 in emergency savings and median net worth at the household level. We then aggregate household-level data to different geographic levels.

Household wealth is a safety net. It protects families from unexpected expenses such as replacing a water heater and income shocks such as a jobless spell. Household wealth is also a springboard. It gives families the opportunities to invest in wealth-building opportunities like starting a small business or moving to an area with more opportunity.

Understanding wealth is necessary for uncovering the barriers to wealth generation and designing policies that unlock opportunity for everyone. Two datasets, the Federal Reserve Board's Survey of Consumer Finances and the US Census Bureau's Survey of Income and Program Participation (SIPP), have detailed information about wealth and are typically used for national or state estimates (Bhutta et al. 2020). Unfortunately, little information is widely available about wealth at the local level, especially disaggregated by different racial and ethnic groups. This is because household surveys with adequate sample sizes for local estimation, such as the American Community Survey (ACS), do not include detailed information about wealth.

We employ machine learning to leverage a smaller nationally representative survey with detailed questions about household wealth to impute household wealth onto the ACS, which allows for small area estimates. We released this novel dataset with the Urban Institute's "Financial Health and Wealth Dashboard" in 2022.[1]

In particular, we use machine learning to estimate median net worth and the proportion of households with at least $2,000 emergency savings at the local level defined as Public Use Microdata Areas (PUMAs). We provide financial measures not only for a city, but also for subregions (Public Use Microdata Areas, or PUMAS) within many cities and for different racial and ethnic groups. City leaders and policymakers are able to use the dashboard to identify wealth disparities within cities. Because household wealth measures for subregions within cities are not typically available in public survey data, we use a machine-learning method to fill the gap in the existing survey data.

We use household measures including net worth and emergency savings to better understand households' wealth. Liquid assets—the sum of assets in checking accounts, stocks, bonds, and other liquid savings accounts—can indicate households' resilience and ability to bounce back from financial shocks. Net worth—total assets minus total debt—can provide an overview of households' economic well-being and their ability to pursue new opportunities.

Overall, the 2022 financial health dashboard provides firsthand data and analysis to city leaders and policymakers for a better understanding of financial health, financial resilience, wealth, and debt within local regions in their cities and counties. They will also be able to compare their cities and counties with neighboring cities and counties.

This works builds on previous asset and wealth imputation research. Using 2014 data, the Prosperity Now Asset Scorecard estimated the asset poverty rate, liquid asset poverty rate, and share of households with liquid assets for many cities in the US. The Federal Reserve Bank of "St. Louis's Real State of Family Wealth"[2] provides national-level average wealth for various demographic groups overtime. Using 2013 data, Ratcliffe and colleagues (2017) estimated the shares of unbanked and underbanked populations for different PUMAs in New York City. This project advances this literature by (1) looking at a broader set of asset outcomes; (2) estimating assets at the PUMA level throughout the country; (3) using more recent data than the literature; and (4) disaggregating the estimates by race and ethnicity when possible. Because PUMA boundaries are often smaller than city boundaries, our analysis sheds light on wealth disparities within cities.

Data Sources

We use detailed wealth data and socioeconomic information from the SIPP to train machine-learning models. The models' predictors include socioeconomic and financial variables from the SIPP that are also available in the ACS. The models' outcomes are liquid assets and net worth. Taking the models as given, we then estimate the liquid assets and net worth for all households in the ACS. The ACS has more granular geographic information than the SIPP but lacks detailed wealth outcomes. The ACS sample covers all the PUMAs in each city and county. Given the predicted wealth measures for the ACS sample, we finally provide summary statistics for net worth and liquid assets for each subregion in cities and counties.

2018 Survey of Income and Program Participation

The SIPP is a household survey of demographic, economic, and government program participation information administered by the US Census Bureau. The survey is nationally representative and is representative of some, but not all, states. The SIPP is a panel dataset with monthly questions asked once a year. The survey is an important source of information about assets and its questionnaire asks detailed questions about government securities, checking accounts, savings accounts, money market accounts or funds, certificates of deposit, municipal and corporate bonds, stocks, mutual funds, retirement accounts, annuities, trusts, and rental properties.

We construct liquid assets and net worth from about a dozen individual questions about assets. Liquid assets include transaction accounts and interest-earning accounts such as checking accounts, savings accounts, certificates of deposit, money market accounts, government securities, municipal and corporate bonds, mutual funds, and stocks. Net worth is the sum of asset values minus the sum of liabilities for a household. We include 26,548 heads of households as our core sample from the 2018 SIPP.

2019 American Community Survey

The ACS is a household survey of demographic, social, economic, and housing characteristics administered by the US Census Bureau on a rolling basis. Survey responses are pooled from all 12 months in a year to create an annual dataset. The survey does not include information about liquid assets or net worth, but it shares many common variables with the SIPP that predict wealth.

The Bureau receives more than 2 million responses to the survey each year, which makes it a useful source of data for small area estimation. We use a Public Use Microdata Sample (PUMS) accessed through IPUMS, which contains 1,217,716 households (Ruggles 2021). While published ACS tables contain information about census tracts with about 4,000 to 8,000 people each, PUMS data are limited to Public Use Microdata Areas (PUMAs). PUMAs are census-generated geographic areas that do not overlap, do not cross state lines, and contain 100,000 or more people each. In some places, PUMAs are larger than entire counties, but in large cities like New York or Chicago that contain many PUMAs, PUMAs can be thought of as large neighborhoods.

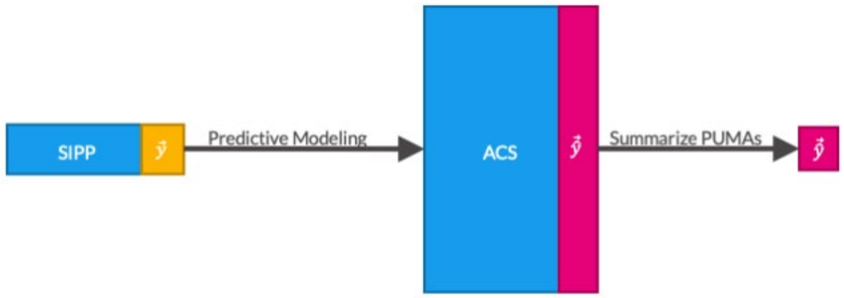
Methodology

Analytical Approach

Our general approach is to leverage the detailed variables in the SIPP to impute values onto the ACS, which contains enough households for small area estimation (figure 1). The approach of using a smaller, more detailed survey to impute variables onto a larger, less detailed dataset is also employed by Blumenstock, Cadamuro, and On (2015), who impute wealth onto administrative cell phone metadata in Rwanda. We estimate models on the SIPP, which has detailed information about household demographics, income, state of residence, and wealth and then predict wealth information using the ACS, which has detailed information about demographics and income. After imputing household microdata onto the ACS, we summarize liquid assets as the proportion of households in a PUMA with more than $2,000 in liquid assets, and we summarize net worth as median net worth for households in a PUMA. We also summarize this information at the city, state, and national levels.

FIGURE 1

The Big Idea for Our Imputation Approach



Source: Authors' calculations.

Our project builds on “Generating Household Wealth and Financial Access Estimates for Local Geographies” (CFED 2015). The authors imputed the share of households in asset poverty and the share of households in liquid asset poverty using the SIPP and ACS. We build on their work in a few ways. First, we use a different measure for liquid assets, and we add median net worth. Second, we use flexible machine-learning models that often outperform parametric models like linear regression. Third, we leverage a robust literature about missing data and imputation to create better imputations. Overall, their imputations attenuated toward the national average, with low-wealth states having too much imputed wealth and high-wealth states having too little imputed wealth (see appendix A).

In the following sections, we offer a summary of concepts from machine learning and missing data imputation that are key features of our analysis.

IMPUTE MICRODATA; THEN AGGREGATE

We have household-level ACS data, but we are interested in quantities estimated for Public Use Microdata Areas (PUMAs). PUMAs are census-generated geographic areas that do not overlap, do not cross state lines, and contain 100,000 or more people each. One approach would be to build models that make PUMA-level estimates. Instead, we opt to impute values at the household level and then aggregate the imputed information to the PUMA level. This better uses the detailed information available at the household level for the SIPP and ACS.

MULTIPLE IMPUTATION

We use multiple imputation (Rubin 1978), an approach where multiple values are predicted for each unit with missing data, to observe the variation in our prediction. However, we do not combine these estimates into formal standard errors for our quantities of interest. The final values in the dashboard are the means of the multiple imputations for each geography.

STOCHASTIC IMPUTATION

Stochastic imputations are imputations with a random error. In our case, we sample from a conditional distribution instead of using conditional mean imputation, such as the predicted value from a linear regression model. This offers two advantages. First, stochastic imputation is necessary for multiple imputations; otherwise, every imputation for a given record would have the same imputed value.

Second, stochastic imputation is more likely to have adequate sample variance. Little and Rubin (2020) note that conditional mean imputation distorts sample variances and covariances and leads to bias when the tails of distributions are being studied. “For example, an imputation method that imputes conditional means for incomes tends to underestimate the percentage of units in poverty” (page 73).

This can be seen in the CFED results in appendix A, where the states in the tails of the distribution are too close to the overall mean.

SEQUENTIAL MULTIPLE IMPUTATION

We are interested in imputing two variables: liquid assets and net worth. We would sacrifice valuable information if we independently imputed the variables. We employ a fully conditional specification—like Raghunathan and colleagues (2001) and van Buuren and colleagues (2006) do—to impute a joint distribution such that

$$f(X_1, X_2 \mid \theta_1, \theta_2) = f_1(X_1 \mid \theta_1) \cdot f_2(X_2 \mid X_1, \theta_2)$$

where X_i are the variables to be imputed and θ_i are model parameters such as regression coefficients. Here, the left side is a joint distribution, and the right side is a marginal distribution times a conditional distribution. We impute liquid assets and then impute net worth conditional on liquid assets and other predictors. We only iterate the model estimation once.

NONPARAMETRIC MACHINE-LEARNING MODELS

The missing data literature focuses on parametric modeling and embraces Bayesian modeling because posterior predictive distributions are natural for stochastic imputation because posterior predictive distributions represent a conditional distribution instead of a conditional mean. Instead, we use nonparametric models from machine learning, including classification and regression trees (Breiman et al. 1984) and random forests (Breiman 2001). We focus on tree-based models because they work well with sampling weights, are easy to tune, and are easy to sample. Nonparametric models sometimes sacrifice global interpretability—how easy it is to understand how a trained model generally makes predictions—but they can outperform parametric models for prediction. We modify algorithms so they generate stochastic predictions.

Analytical Process

This section briefly outlines the process we use to estimate models on the SIPP data and impute liquid assets and net worth onto the ACS.

STEP 1: PREPARE THE DATA

We pull SIPP data from the US Census Bureau website. We drop 38 observations where the head of household is younger than age 18. Also, we create an ordinal education–level variable.

We then pull ACS data from IPUMS. We create an ordinal education variable that matches the SIPP. We simplify race and ethnicity categories and resolve missing values for the foreign-born indicator. We also convert the top-coded category for "ages 90+" to numeric age 90 so the entire age variable can be treated as numeric.

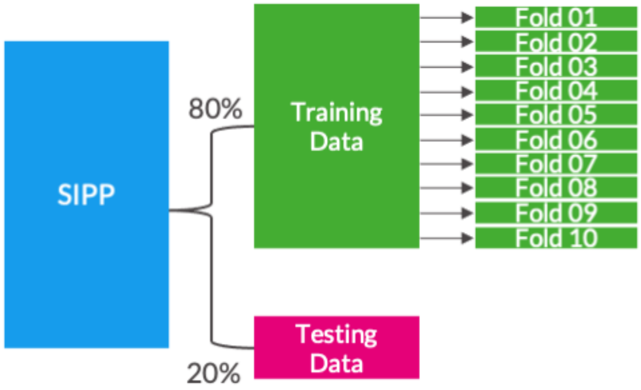
STEP 2: SPLIT DATA

We want to estimate models on the SIPP that make good predictions on the ACS. In machine learning, it is easy to overfit the data used to estimate models and end up with estimated models that do not generalize to the prediction task of interest. We partition our data into a training dataset with 80 percent of observations from the SIPP and a testing dataset with 20 percent of observations from the SIPP. The datasets are disjointed and exhaustive. We do not touch the testing data until model selection is complete, at which point we can estimate the out-of-sample error.

For model selection, we further split our training data into 10 disjointed and exhaustive samples (or "folds") and employ v-fold cross validation to avoid overfitting our models (Blum, Kalai, and Langford 1999; Burman 1989). For each outcome variable and model specification, we estimate the model 10 times using 9 of the folds and then estimate our error metric on the remaining fold. We call the folds used to train a model "analysis" data and the remaining fold used to test a model "assessment" data. This gives us estimates of out-of-sample error rates that can be used to pick the "best" model. Figure 2 shows how we split the data, and figure 3 shows the v-fold cross-validation process.

FIGURE 2

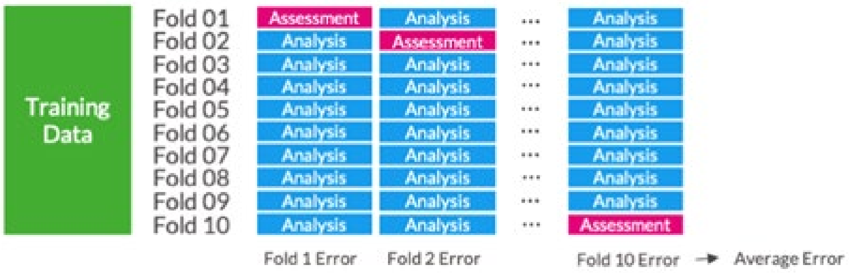
How We Use Our Data



Source: Authors' calculations.

FIGURE 3

V-Fold Cross Validation



Training Data					
	Fold 01	Assessment	Analysis	...	Analysis
	Fold 02	Analysis	Assessment	...	Analysis
	Fold 03	Analysis	Analysis	...	Analysis
	Fold 04	Analysis	Analysis	...	Analysis
	Fold 05	Analysis	Analysis	...	Analysis
	Fold 06	Analysis	Analysis	...	Analysis
	Fold 07	Analysis	Analysis	...	Analysis
	Fold 08	Analysis	Analysis	...	Analysis
	Fold 09	Analysis	Analysis	...	Analysis
	Fold 10	Analysis	Analysis	...	Assessment
		Fold 1 Error	Fold 2 Error		Fold 10 Error → Average Error

Source: Authors' calculations.

STEP 3: CHOOSE ERROR METRICS

The usefulness of an error metric depends on the application of a model. We release household data aggregated to different geographies. Therefore, we care about summary statistics of our household-level data. We use several error metrics to pick the "best" model:

- **Weighted threshold error:** The difference between the true and predicted weighted proportions of households with a chosen asset greater than a chosen threshold. For example, this metric estimates how well the predicted data recreate the proportion of households with more than $2,000 in liquid assets.
- **Weighted 25th percentile error:** The difference between the true and predicted weighted 25th percentile for a chosen asset.
- **Weighted 50th percentile error:** The difference between the true and predicted weighted median for a chosen asset.
- **Weighted 75th percentile error:** The difference between the true and predicted weighted 75th percentile for a chosen asset.

We use weighted threshold error and weighted 50th percentile error because they connect directly to the statistic we include in the dashboard. We also look at weighted 25th percentile error and weighted 75th percentile error to see if it is feasible to release other statistics.

All error metrics are weighted using the household weights from the SIPP. We use the average metric from the 10 iterations of the cross validation to pick the best model, but we allow ourselves to override this decision if error metrics vary a lot from fold-to-fold in cross validation.

STEP 4: ESTIMATE THE FIRST ROUND OF MODELS

We estimate two models for step 4—one for each outcome variable: (1) liquid assets and (2) net worth. We estimate three important considerations for each model:

1. **The potential predictors included in the model:** Each outcome variable needs its own predictive model. Both models have the following potential predictors:
   a. household income
   b. education level of the head of household (unordered): less than high school (HS), HS or GED, some college, bachelor's degree, and a postgraduate degree
   c. education level of the head of household (ordered): integers where 1 = less than HS, 2 = HS or GED, 3 = some college or a bachelor's degree, and 4 = a postgraduate degree
   d. race/ethnicity for the head of household: Asian, non-Hispanic; Black, non-Hispanic; other, non-Hispanic; and White, non-Hispanic
   e. gender of the head of household
   f. age of the head of household
   g. indicator for if the head of household is single
   h. indicator for if the head of household is foreign born
   i. indicator for metro area
   j. state of residence
   k. number of adults in household
   l. number of dependents in household
   m. home value
   n. indicator for homeownership
   o. imputed liquid assets (only for the net worth model in step 5)
2. **The algorithm and its hyperparameters:** We test decision trees and random forest models for each outcome variable. Decision trees are nonparametric models for classification and regression that include feature selection and work well where linear models fall short. Decision trees are quick to estimate and easy to understand, but they tend to overfit the data. See appendix B for a technical explanation of decision trees and random forests.

Random forest models are ensembles of decision trees that take longer to estimate but tend to not overfit the data and work well “out-of-the-box.” A random forest is an ensemble of **ntree** decision trees where each tree is estimated on a bootstrap sample of the training data and each split in the decision tree is selected by only considering **mtry** of the predictors included in the model. The bootstrap sampling and limitation on predictors considered introduces randomness and decorrelates the individual decision trees. The depth of each tree is determined by **min_n**.

We experiment with several variations of unweighted and weighted random forests. We use traditional random forests. Random forests typically make conditional mean predictions by averaging the ntree predictions from the forest. We use quantile random forests, which make conditional quantile predictions. In our case, we focus on the median, or 50th percentile, predicted value.

Conditional mean imputation results in imputed values with distorted sample variance and too few values in the tails of the distribution (Little and Rubin 2020). Conditional median imputation suffers the same limitations. So we randomly sample a predicted value from the vector of **ntree** predictions available in the random forest model. This approximates a posterior predictive distribution and leads to better results. For the best model, which is a type of random forest model, we hyperparameter tune ntree, mtry, and min_n, a process of trying different parameter values, to achieve the best model performance.

3. **Feature and target engineering:** We run through a set of preprocessing steps each time we estimate a model. We repeat this process for each v-fold cross-validation step, instead of once at the beginning of the machine-learning process, to avoid data leakage because some preprocessing steps are sample dependent. Data leakage is any time data outside of the training data are used to estimate a model. Data leakage results in underestimates of out-of-sample error metrics. For example, if we need to center a variable, the overall SIPP mean would include observations from the testing data and cause data leakage. Our preprocessing steps are as follows:

    a. drop zero or near-zero variance predictors

    b. impute missing values for predictors—we use K Nearest-Neighbors imputation for missing values for Metro, the only variable with missing values; because imputation is sample-dependent, we repeat this process within cross validation to prevent data leakage

c. both dependent variables are right skewed—we apply Yeo-Johnson transformations to deemphasize outliers (Yeo and Johnson 2000); the lambda parameter for Yeo-Johnson transformations is sample-dependent; we estimate lambda during cross validation to prevent data leakage

d. dummy encode categorical features

In addition to the model-selection metrics outlined above, we examine variable importance to assess whether the most important predictors are sensible given our subject-matter knowledge. Variable importance is a measure of how much a predictor reduces the heterogeneity of nodes. Variable importance can be biased and can overstate the impact of continuous predictors. The variable importance matches our subject-matter knowledge. For example, other measures of wealth and household income are the most important predictors for wealth. We now discuss each model.

Liquid assets is the most straightforward of the two variables to model. Random forest models with sample imputed values outperform the other models of liquid assets for most parameterizations. The imputed weighted 75th percentile is systematically low, but the error is minor and the only metric of interest is the proportion of households with more than $2,000 in liquid assets.

Net worth can be negative and has larger outliers than the other variables. The models without liquid assets as a predictor are generally poor. They systematically overestimate the weighted median and the weighted 75th percentile.

Based on these results, we decide to impute liquid assets and then net worth.

STEP 5: ESTIMATE THE SECOND MODEL

We are interested in modeling the joint distribution of liquid assets and net worth. In addition to resulting in a sensible relationship between liquid assets and net worth, modeling the joint distribution means we can leverage predictions from the first model to improve the second model.

We use a sequential imputation approach for liquid assets and net worth. In sequential imputation methods, the order of imputations usually goes from the variable with the lowest share of values that are missing to the variable with the highest share of values that are missing (Rubin and Schafer 1990). In our case, all values are missing for liquid assets and family wealth on the ACS.

We impute liquid assets first because the initial model for liquid assets far outperforms the initial model for net worth. We then repeat the model estimation process from step 4 for net worth with imputed liquid assets as a predictor. The addition of liquid assets as a predictor in the model for net worth improves the model for net worth.

STEP 6: ESTIMATE THE GENERALIZATION ERROR (OUT-OF-SAMPLE ERROR)

We use these two models—the liquid assets model from step 4 and the net worth model from step 5— to make predictions on the testing data held out in step 1. We recompute the error metrics for each model on the testing data and get an estimate of the generalization error or out-of-sample error. These error metrics give us a sense of the error of our predictions for the ACS.

STEP 7: IMPUTE VALUES ONTO THE ACS

We have an imputation model for liquid assets and net worth. We use these models to generate stochastic imputations for all households in the 2019 ACS. We repeat this process fifteen times and then average the imputations.

In general, the optimal number of implicates for multiple imputation is an open question (van Buuren 2018). We look at running mean plots and determine that the average of imputations is stable after only a few imputations, and 15 is more than enough.

STEP 8: SUMMARIZE TO THE PUMA LEVEL

Finally, we aggregate household-level imputed liquid assets and net worth at the PUMA level. For liquid assets, we calculate the weighted proportion of households with at least $2,000 in liquid assets. For net worth, we calculate the weighted median net worth. For some cities, we further aggregate these PUMAs to the city level.

Generating Statistics by Race and Ethnicity

We have imputed liquid assets, imputed net worth, and the observed race/ethnicity for the head of the household for each household in the ACS. We calculate the proportion of households with at least $2,000 in liquid assets and median net worth for each PUMA conditional on the observed race/ethnicity of the head of household to generate statistics by race and ethnicity. We generate estimates for households led by people who are Hispanic, non-Hispanic AAPI, non-Hispanic Black, and non-Hispanic other races.

Many PUMAs lack sufficient observations to report statistics disaggregated by race and ethnicity. Imputed asset metrics (median net worth and emergency savings) are not reported when the sample size is less than 50 or the coefficient of variation is more than 0.4. In addition, for net worth, we did not suppress values when the standard error was less than $5,000 when the standard error was less than $15,000, and when the CV was less than 0.5.

Results

In this section, we present the following results for each outcome variable:

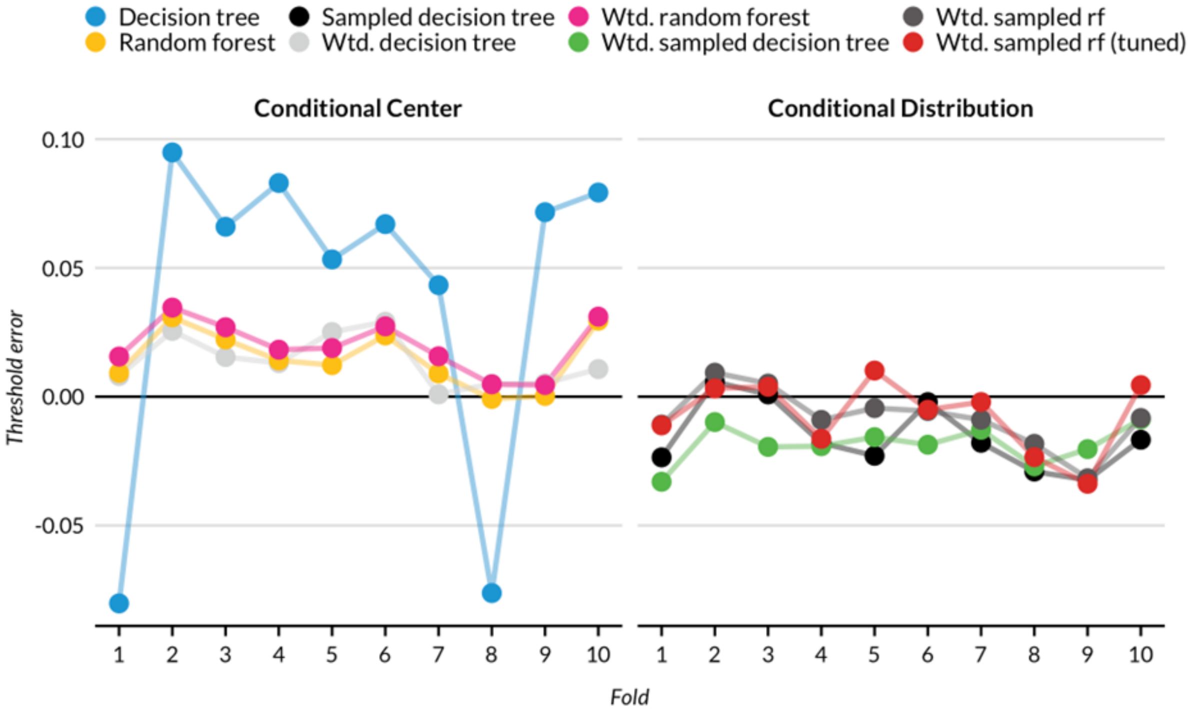
- the algorithm that performs best on the training data with cross validation
- the hyperparameterization that performs best on the training data with cross validation
- variable importance
- the out-of-sample error rate estimated on the testing data
- validation of the imputations against known estimates from other data sources

Liquid Assets

The weighted random forest model with sampling performs best when evaluated with cross validation on the training. Our outcome of interest is the proportion of households with at least $2,000 in liquid assets (figures 4 and 5), but we also assess performance on the median liquid assets for completeness (figure 6). Sampling from a conditional distribution dramatically outperforms using a conditional mean or conditional median (conditional center). The four conditional distribution models are comparable, and we opt for the hyperparameter-tuned model.

FIGURE 4

Conditional Distribution Imputation Outperforms Conditional Center Imputations for Liquid Assets

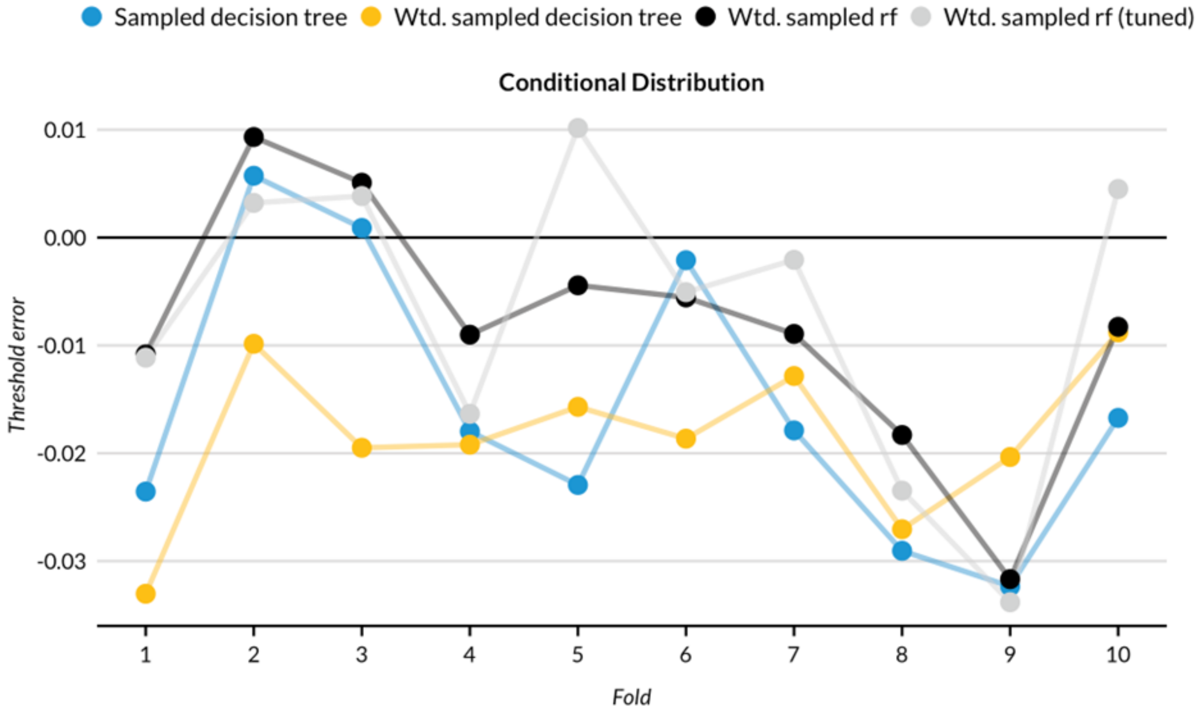


Source: Authors' calculations using the 2018 SIPP.

Notes: Each fold is a different analysis, or holdout, dataset from v-fold cross validation. Results from several models are excluded for clarity.

FIGURE 5

Conditional Distribution Imputation Methods Are Comparable for Imputing Liquid Assets

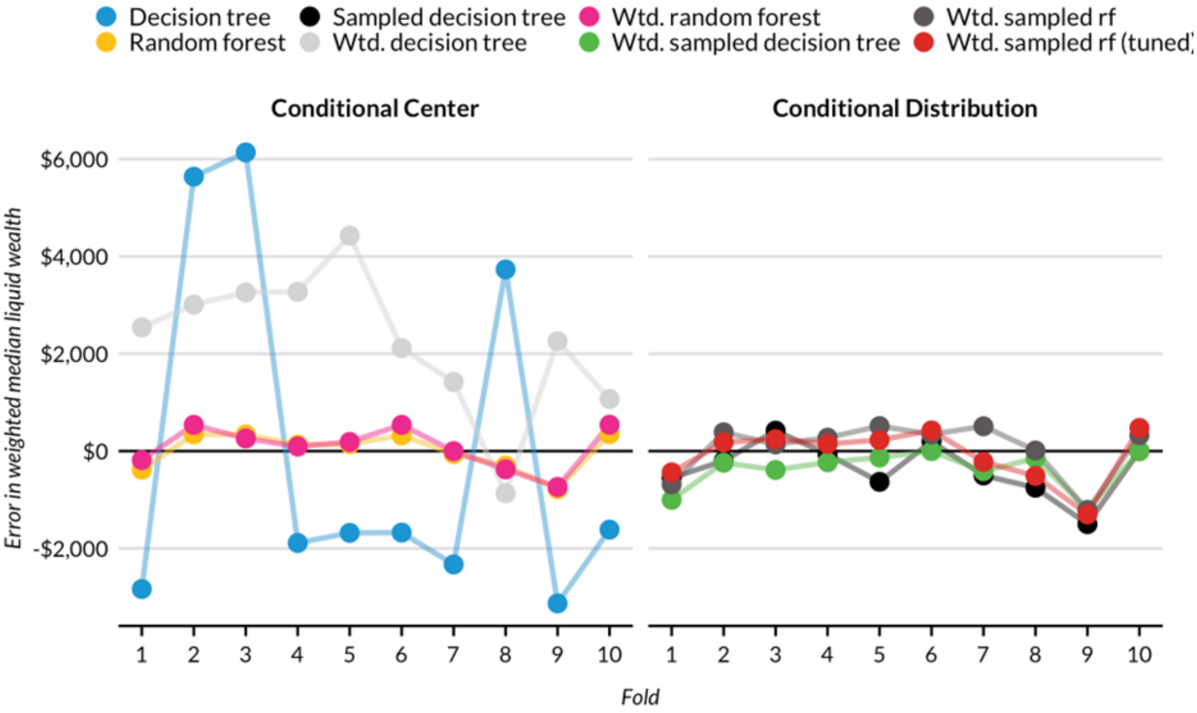


Source: Authors' calculations using the 2018 SIPP.

Note: Each fold is a different analysis, or holdout, dataset from v-fold cross validation.

FIGURE 6

Conditional Distribution Imputation Methods Are Comparable for Imputing Median Liquid Assets



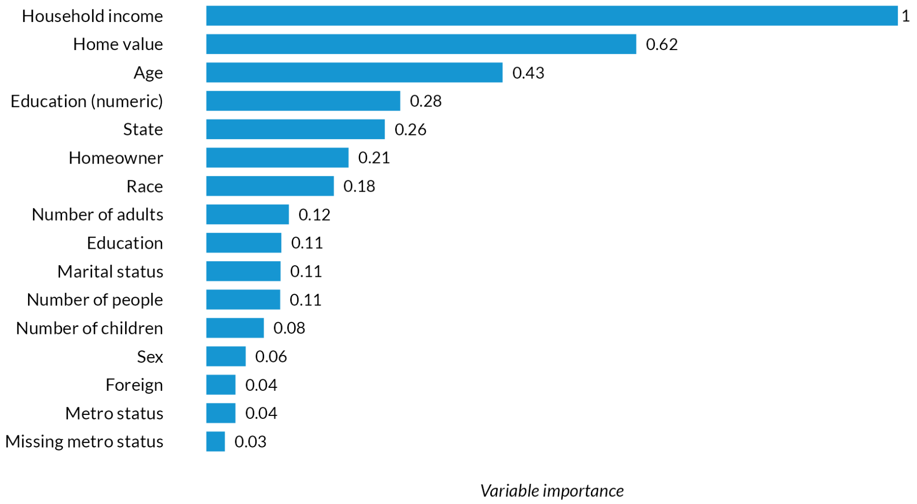
Source: Authors' calculations using the 2018 SIPP.

Notes: Each fold is a different analysis, or holdout, dataset from v-fold cross validation. Results from several models are excluded for clarity.

Figure 7 shows that household income, home value, and age are the most important predictors for predicting liquid assets.

FIGURE 7

Household Income and Home Value Are the Most Important Predictors for Liquid Assets



Source: Authors' calculations using the 2018 SIPP.
Notes: Variable importance is measured by the reduction in the sum of squared errors attributable to each variable. Larger bars indicate more importance. All values are divided by the maximum variable reduction and represent relative importance.

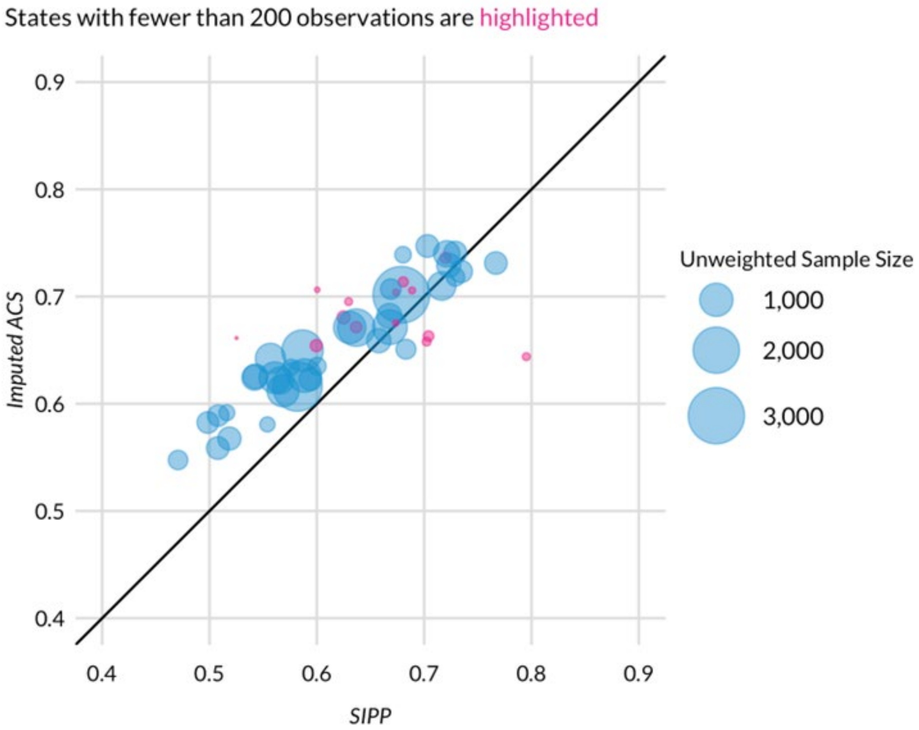
The liquid assets results are not very sensitive to hyperparameter tuning. We choose 500 trees, at least 5 values in each final node, and consider three variables at each split.

The final model overestimates the number of households with $2,000 in liquid savings by 0.2 percentage points and overestimates the weighted median liquid assets by $331 when measured against the testing dataset.

Finally, figure 8 compares the weighted proportion of households with more than $2,000 in liquid assets at the state level calculated on values reported in the SIPP and on the imputed ACS values. Our imputation methodology follows the trend, but the imputed ACS values are systematically higher than the SIPP values. The SIPP is not representative off all states, so state plots with fewer than 200 observations in the SIPP are colored magenta (see figure 8).

FIGURE 8

Imputed Liquid Assets Closely Matches the SIPP at the State Level



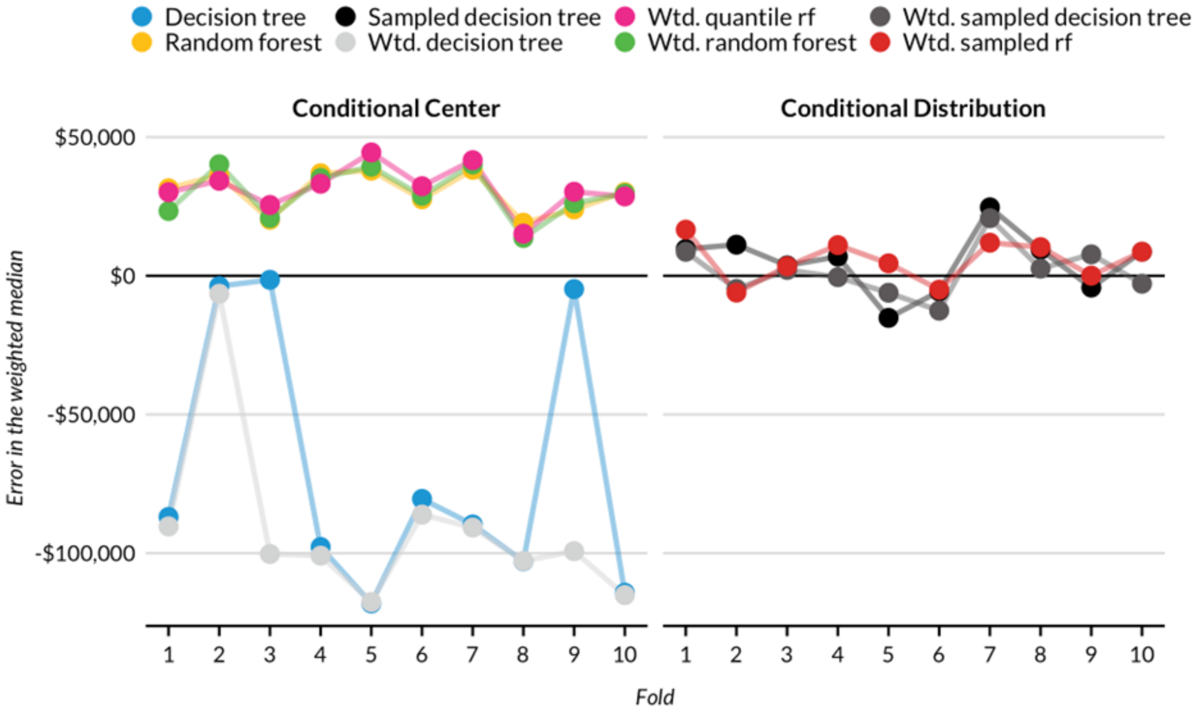
Source: Authors' calculations using the 2018 SIPP and 2019 ACS.

Net Worth

The weighted random forest model with sampling performs best when evaluated with cross validation on the training data. Our outcome of interest is median net worth (figures 9 and 10), but we also evaluate our models on the proportion of households with at least $360,000 in net worth for completeness (figure 11). Sampling from a conditional distribution dramatically outperforms using a conditional mean or conditional median (conditional center). The four conditional distribution models are comparable, and we opt for the hyperparameter-tuned model.

FIGURE 9

Conditional Distribution Imputation Outperforms Conditional Center Imputation for Net Worth

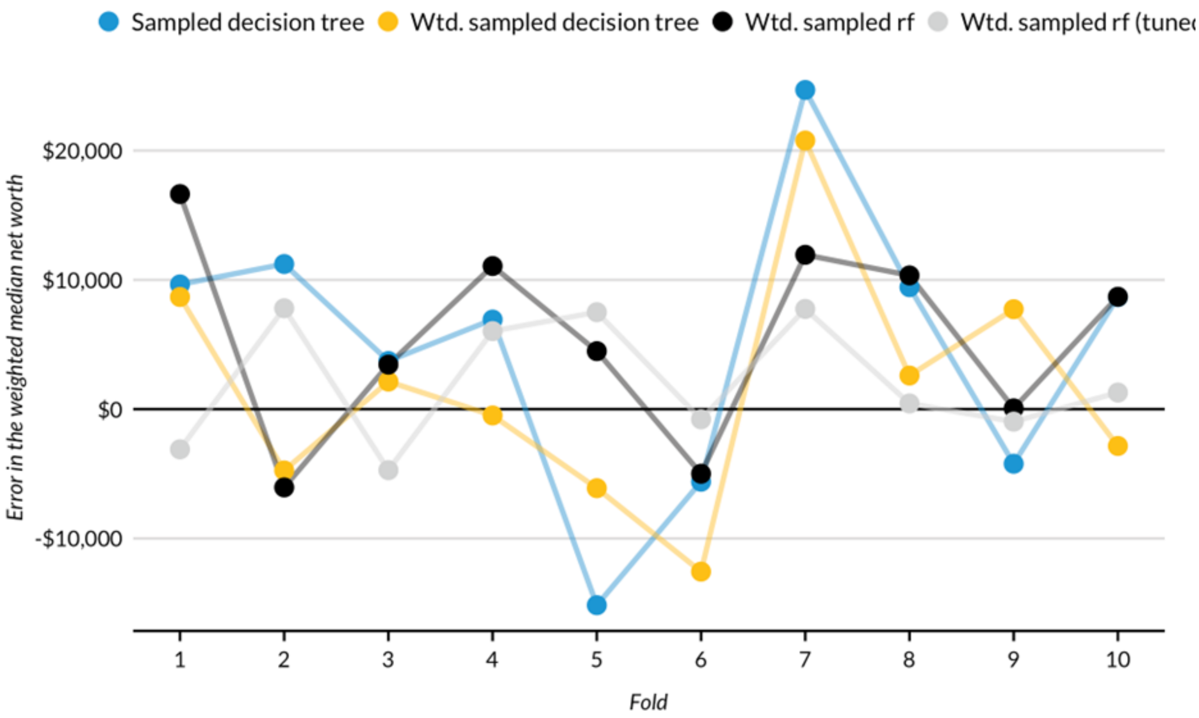


Source: Authors' calculations using the 2018 SIPP.

Notes: $360,000 in net worth is a threshold comparable with $2,000 in liquid assets. Each fold is a different analysis, or holdout, dataset from v-fold cross validation.

FIGURE 10

Conditional Distribution Imputation Methods Are Comparable for Net Worth

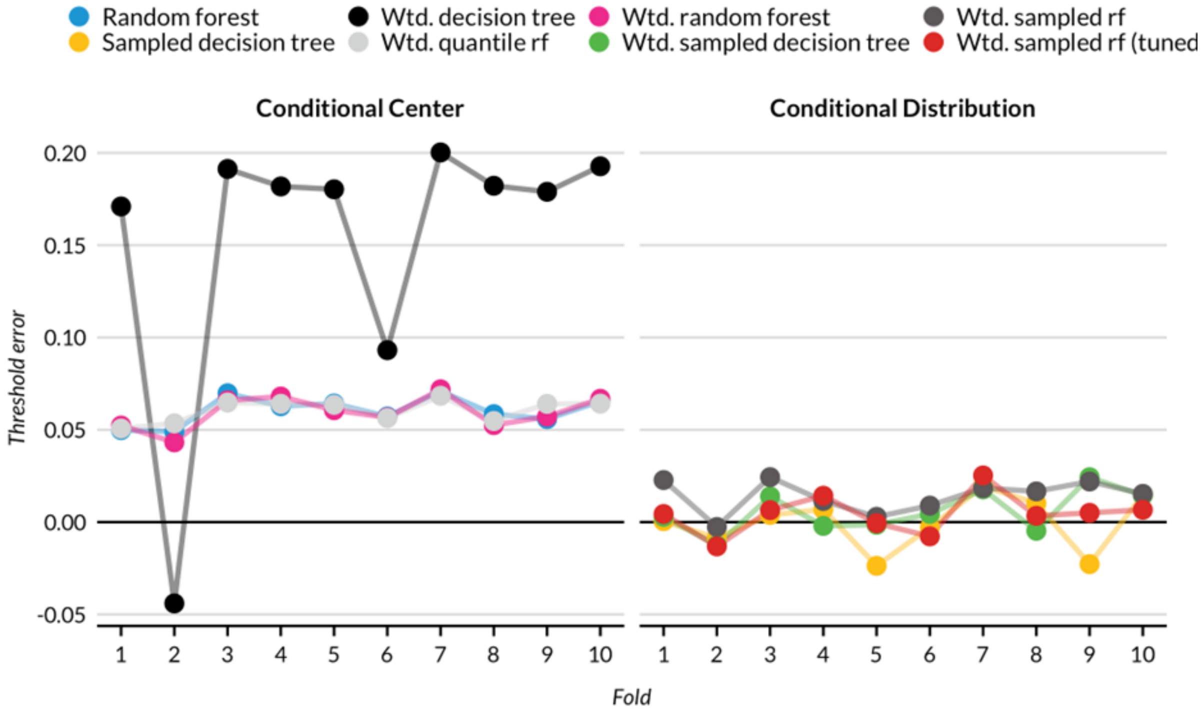


Source: Authors' calculations using the 2018 SIPP.

Note: Each fold is a different analysis, or holdout, dataset from v-fold cross validation.

FIGURE 11

The Net Worth Imputation Model Performs Well for the Threshold Measure

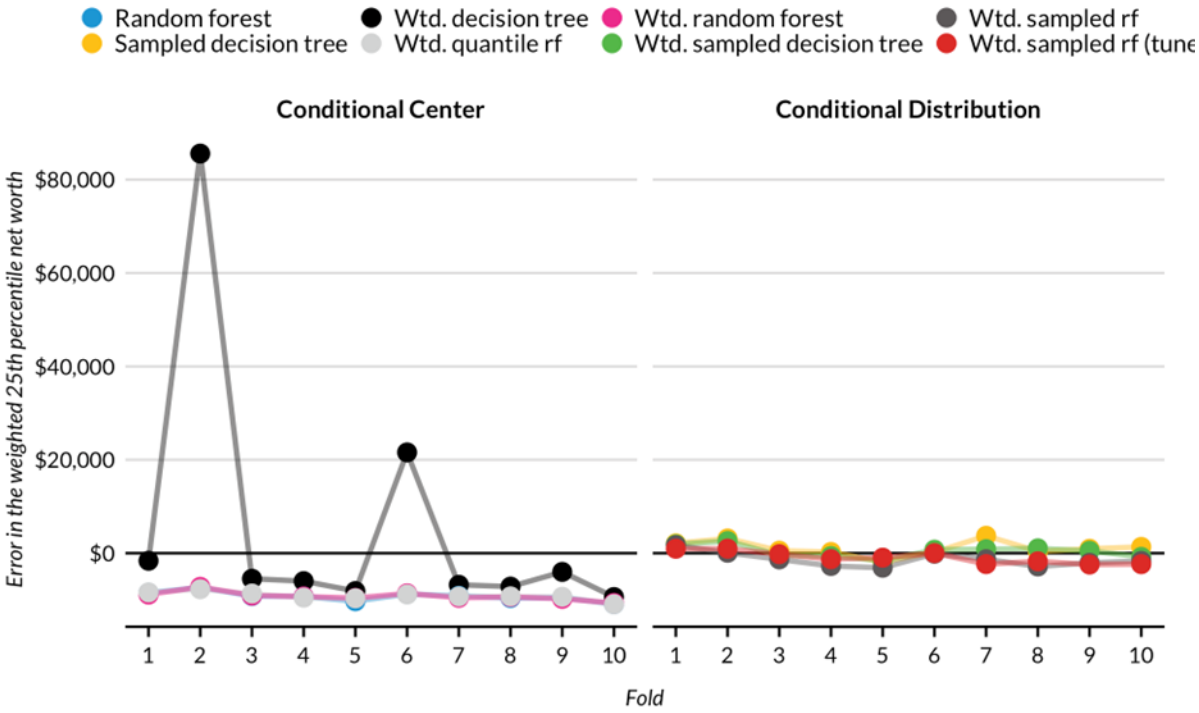


Source: Authors' calculations using the 2018 SIPP.
Note: Each fold is a different analysis, or holdout, dataset from v-fold cross validation.

We also consider the weighted 25th percentile net worth (figure 12) and weighted 75th percentile net worth (figure 13). In both cases, sampling from weighted random forests outperforms other approaches.

FIGURE 12

Conditional Distribution Imputation Outperforms Conditional Center Imputation for 25th Percentile Net Worth

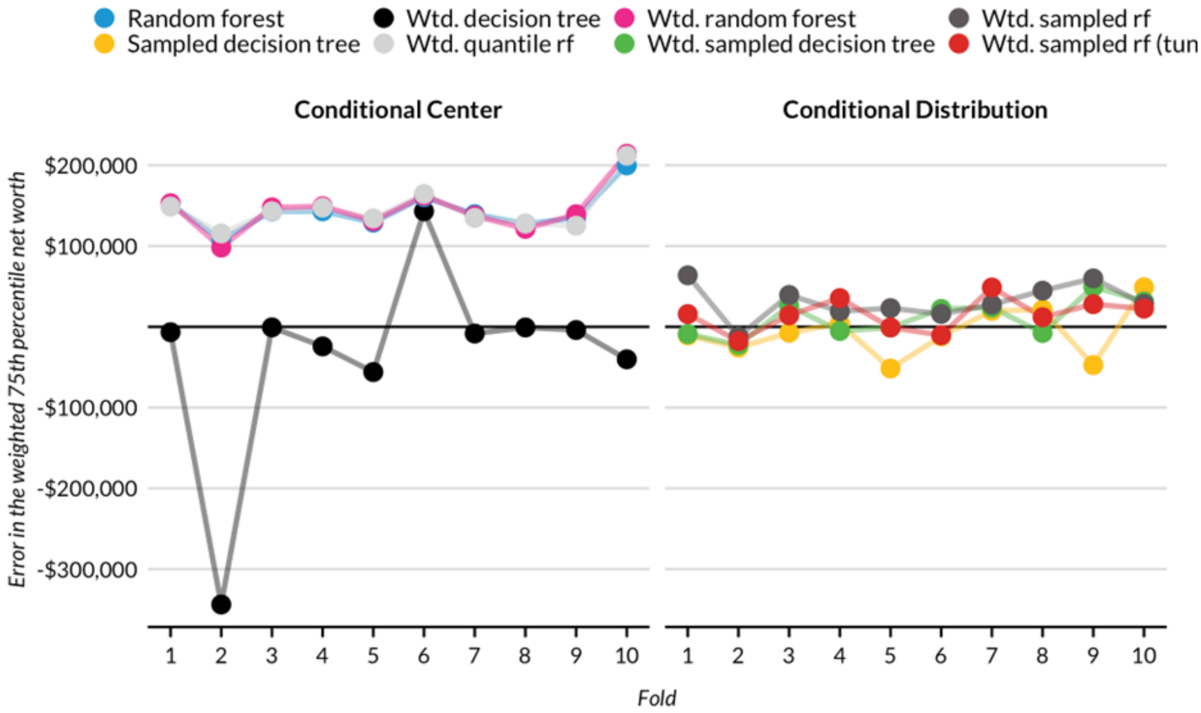


Source: Authors' calculations using the 2018 SIPP.

Note: Each fold is a different analysis, or holdout, dataset from v-fold cross validation.

FIGURE 13

Conditional Distribution Imputation Outperforms Conditional Center Imputation for 75th Percentile Net Worth



Source: Authors' calculations using the 2018 SIPP.
Note: Each fold is a different analysis, or holdout, dataset from v-fold cross validation.

For net worth, the hyperparameterization affects model performance when measured by cross validation on the training data. Table 1 shows different hyperparameterizations ranked by the average error in weighed median. We choose 500 trees, no minimum number of observations in the final node, and consider three variables at each split.

TABLE 1

Aggregated Results from Hyperparameter Tuning Sampling from a Random Forest Model

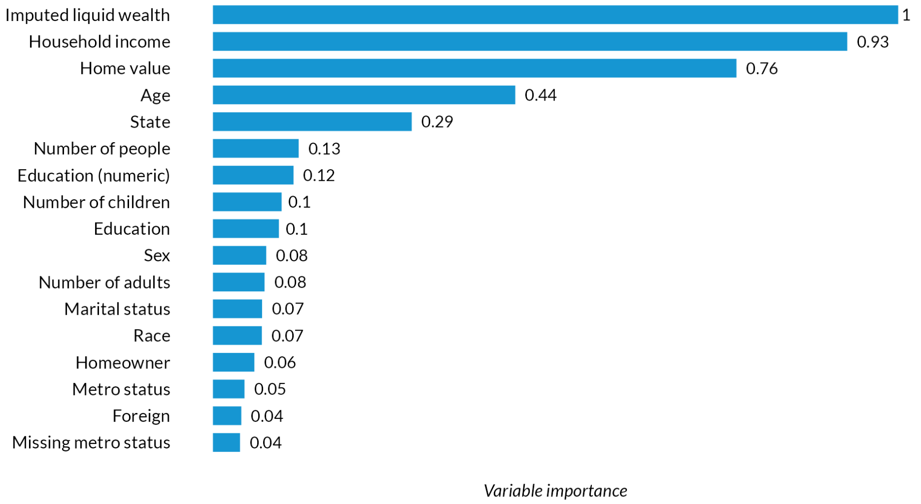
Rank	trees	min_n	mtry	Threshold error	Median error	25th percentile error	75th percentile error
1	500	1	4	-0.00974	219	-831	9,411
2	500	2	6	-0.0101	-397	-617	19,630
3	500	1	6	-0.0118	1,154	-932	9,751
4	500	2	5	-0.0165	-1467	-1323	9,938
5	500	3	5	-0.0125	2,683	-821	24,495
6	500	3	6	-0.0125	2,808	-993	21,429
7	500	1	3	-0.0254	2,882	-2006	26,421
8	500	2	4	-0.0127	3,263	-845	19,145
9	500	1	5	-0.0135	3,267	-1061	13,402
10	500	3	4	-0.0153	6,714	-946	20,680
11	500	2	3	-0.0204	7,982	-1504	18,933
12	500	3	3	-0.0265	10,637	-1965	30,907

Source: Authors' calculations.

Figure 14 shows that the predicted value of liquid assets, household income, home value, and age are the most important predictors for predicting net worth.

FIGURE 14

Imputed Liquid Assets and Household Income Are the Most Important Predictors of Net Worth



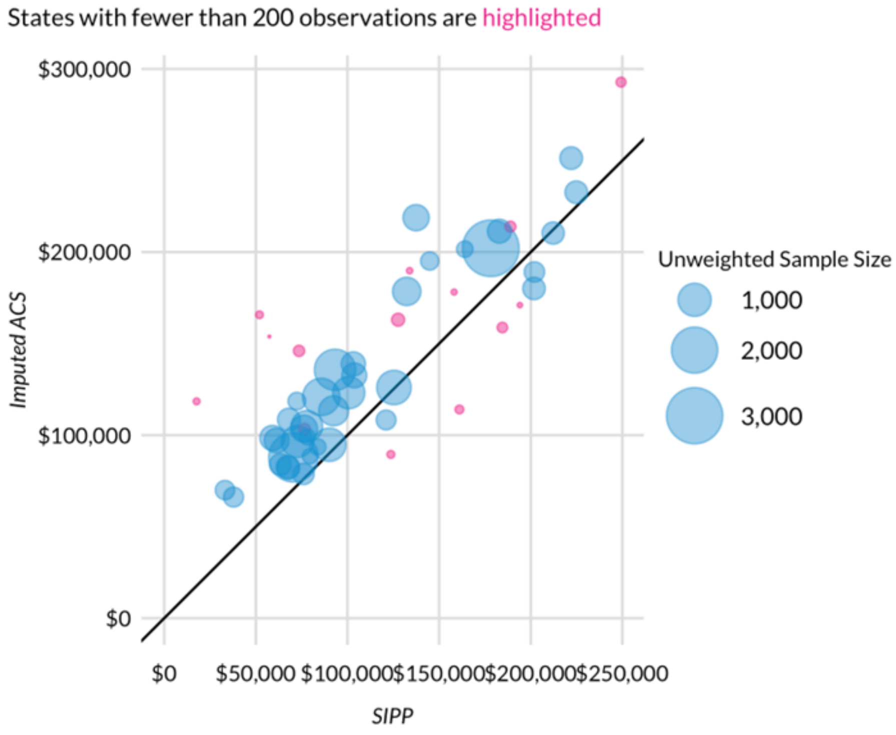
Source: Authors' calculations using the 2018 SIPP.
Notes: Variable importance is measured by the reduction in the sum of squared errors attributable to each variable. Larger bars indicate more importance. All values are divided by the maximum variable reduction and represent relative importance.

The final model underestimates the number of households with $360,000 in net worth by 1.3 percentage points and underestimates the weighted median net worth by $2,386 dollar when measured against the testing dataset.

Finally, figure 15 compares the weighted median wealth at the state level calculated from values reported in the SIPP and on the imputed ACS values. Our imputation methodology follows the trend, but the imputed ACS values are systematically higher than the SIPP values. The SIPP is not representative of all states, so state plots with fewer than 200 observations in the SIPP are colored magenta (see figure 15).

FIGURE 15

Imputed Median Net Worth Matches the SIPP at the State Level



Source: Authors' calculations using the 2018 SIPP and 2019 ACS.

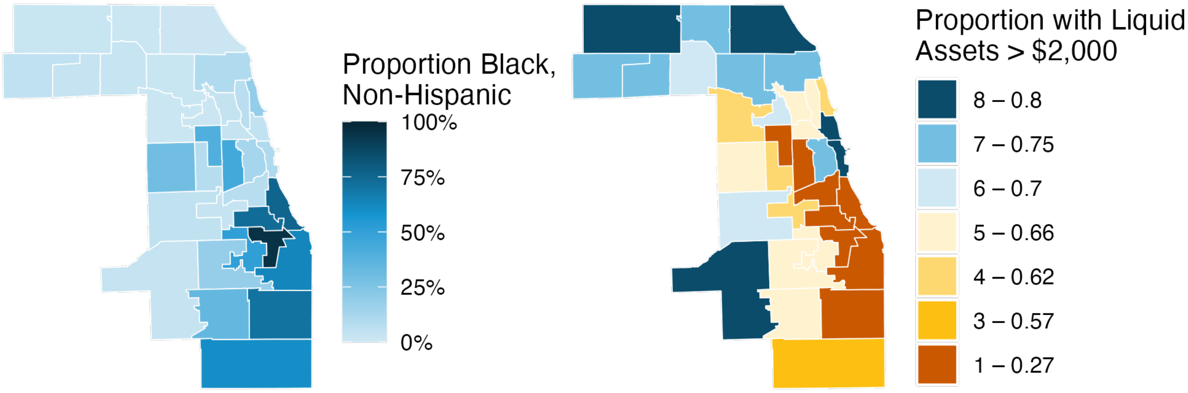
Unequal Opportunities by Race and Ethnicity

The imputed data can be used to explore the stark differences in liquid assets and net worth in different communities. Many of these differences are driven by structural racism and compounded disadvantages faced most acutely by Black Americans. Figure 16 shows the proportion of people who are Black (left) and the proportion of households with more than $2,000 in liquid assets (right) in

PUMAs that touch Cook County, Illinois. The liquid assets variable is grouped into eight groups. The pattern is clear. PUMAs with the highest proportion of households with at least $2,000 in liquid assets are the PUMAs with the lowest Black, Non-Hispanic population proportions. The PUMAs with the lowest proportion of households with at least $2,000 in liquid assets have the highest Black, Non-Hispanic population proportions.

FIGURE 16

Lack of Liquid Assets Is Concentrated in Black Communities in Chicago and Cook County

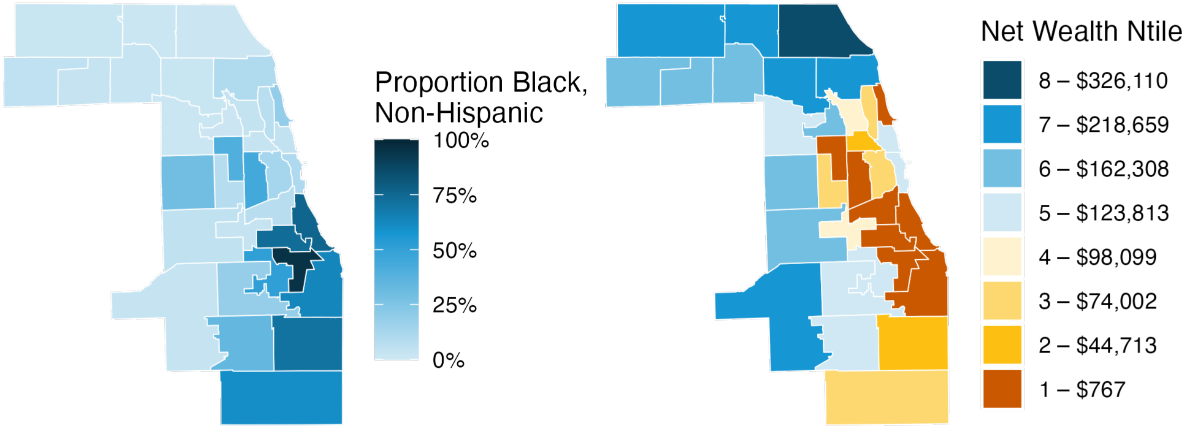


Source: Authors' calculations using the 2018 SIPP and 2019 ACS.

Figure 17 shows the proportion of people who are Black (left) and median net worth (right) in PUMAs that touch Cook County, Illinois. The net worth variable is grouped into eight groups. The pattern is clear. PUMAs with the highest median net worth are the PUMAs with the lowest Black, Non-Hispanic population proportions. The PUMAs with the lowest median net worth have the highest Black, Non-Hispanic population proportions.

FIGURE 17

Lack of Net Worth Is Concentrated in Black Communities in Chicago and Cook County



Source: Authors' calculations using the 2018 SIPP and 2019 ACS.

Conclusion and Discussion

This report demonstrates a feasible methodology for using detailed wealth information from the SIPP and large sample sizes from the ACS to learn about wealth at the local level. This methodology works for coarse aggregates like a threshold or median but should not be used for making inferences about individual households or complex statistics calculated on the household values.

We demonstrate with cross validation on the training data that the accuracy of the machine-learning models is sensitive to modeling decisions. Sampling from weighted random forests provided the best results in all instances and was better than sampling from decision trees and much better than conditional mean random forests or conditional median random forests.

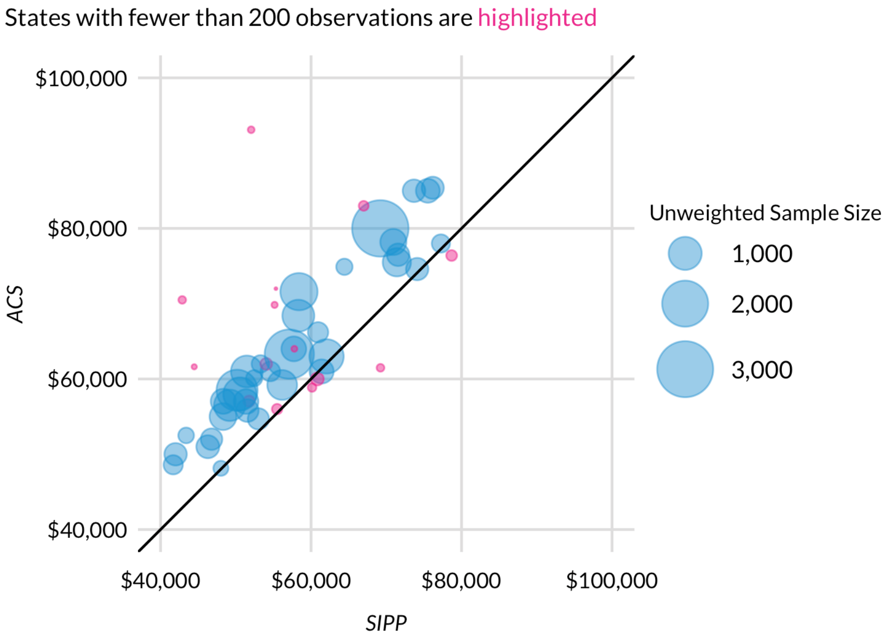
Much of the imputation literature focuses on parametric methods, and in particular, Bayesian methods. The machine-learning models tested in this methodology make accurate predictions, and while they are less interpretable than parametric regression models, we are able to understand their behavior with variables' importance. However, variable importance should not be thought of as causal in any way. Home wealth is predictive of liquid assets and net worth, but that does not mean that home wealth causes wealth accumulation or that policymakers should interpret these results as justification for encouraging homeownership.

Figures 8 and 15 demonstrate that there are nontrivial errors when validating the imputed ACS against the SIPP at the state level. First, caution should be exercised when looking at these visualizations because the SIPP is not reliable for all 50 states and the District of Columbia. The SIPP estimates have a sampling error. The imputed ACS estimates have a sampling error and an imputation error. Although the ACS has an additional type of error, it has much less of a sampling error than the SIPP. For example, Alaska has 47 observations in the SIPP data and 2,257 observations in the ACS data. Vermont has 53 observations in the SIPP data and 2,777 observations in the ACS data. We summarize data to the PUMA level for our analyses. In our data, PUMAs range from 183 to 1,961 observations with a median size of 506 observations. This suggests that while imputation error is a concern for the ACS, a sampling error is less of a concern with the ACS than the SIPP.

Second, the imputed ACS is systematically higher for the liquid assets outcome and the net worth outcome. This is likely because important predictors for liquid assets and net worth are systematically higher in the ACS than the SIPP. Figure 18 shows higher state medians for income, and figure 19 shows higher state medians for home values. In both examples, values above the black diagonal line have higher values for the ACS than the SIPP.

FIGURE 18

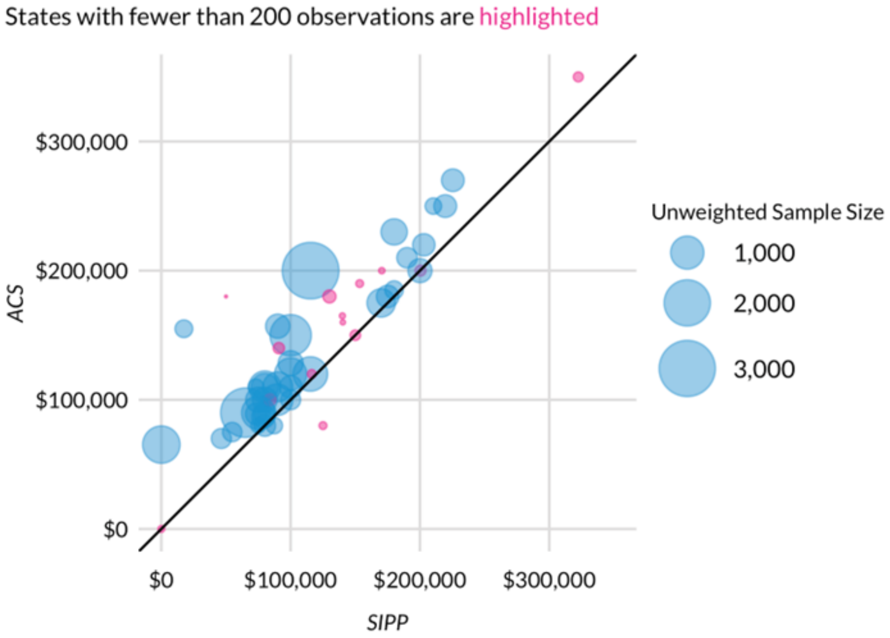
Median Income Is Systematically Higher in the ACS Than the SIPP



Source: Authors' calculations using the 2018 SIPP and 2019 ACS.

FIGURE 19

Median Home Value Is Systematically Higher in the ACS Than the SIPP



Source: Authors' calculations using the 2018 SIPP and 2019 ACS.

The modest increases in reported liquid assets and net worth are tolerable and maybe even desirable because the SIPP underreports assets.

We elect to not publish the 25th and 75th percentiles separately. First, this is because we care about differences between PUMAs more than the differences within PUMAs. Second, while the errors for 25th percentiles were small in dollar terms because the true values are close to zero, the errors for 75th percentiles were large. In the future, we could use quantile random forests to model the 25th, 50th, and 75th percentiles separately.

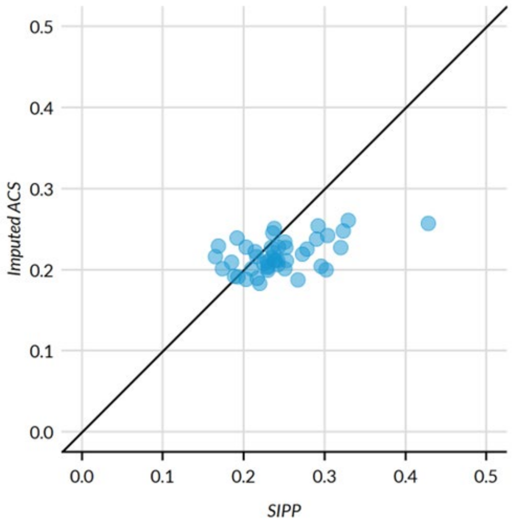
The SIPP and ACS generate different estimates for many variables. There are also known differences between the SIPP and the Federal Reserve Board's Survey of Consumer Finances, which is considered one of the best datasets about wealth (Chen et al. 2018). Calibration is an approach to adjusting sampling weights to hit known population totals. We could treat the ACS as known population totals and calibrate the SIPP to hit key predictors like income and home value. This would change the models estimated on the SIPP. Alternatively, we could treat the SCF as known population totals and calibrate the asset outcomes in the SIPP to hit the SCF. This would change the models estimated on the SIPP. Finally, we could treat the SCF as known population totals and calibrate the imputed asset outcomes onto the ACS to hit the SCF. This would change the statistics calculated using the ACS.

Finally, we focus on liquid assets and net worth in this analysis, but this methodology could be applied to other household finance–related variables with place-based policy considerations. It would be useful to apply this methodology with access to workplace retirements like 401(k) and 403(b) accounts because many states are considering state-mandated retirement accounts.

Appendix A. CFED (Corporation for Enterprise Development) Imputations

FIGURE A.1

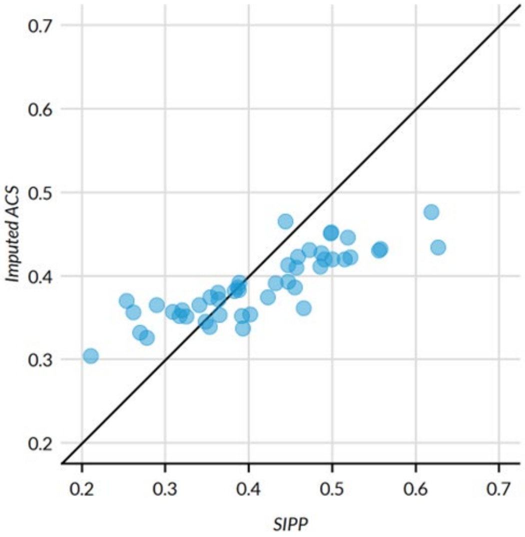
The Imputed Asset Poverty State Tabulations Are Too Close to the National Mean



Source: Table B1 in CFED (2015).

FIGURE A.2

The Imputed Liquid Assets Poverty State Tabulations Are Too Close to the National Mean



Source: Table B2 in CFED (2015).

Appendix B. Model Descriptions

Decision Trees

Decision trees are nonparametric models that use predictor variables to sort observations of an outcome variable into relatively homogeneous groups. Decision trees are often called classification trees for categorical outcome variables and regression trees for numeric outcome variables. Decision trees, or Classification and Regression Trees, were developed by Breiman and colleagues (1984). We use the rpart implementation of decision trees. Therneau and Atkinson (2019), the authors of rpart, describe how a tree is built as follows:

- find the variable that best splits the data into two groups; split the data
- for each subgroup, find the variable that best splits the data into two groups; split the data
- continue this process until the subgroups reach a user-specified minimum size or until no improvement can be made; and
- optionally, use cross validation to reduce the full tree to avoid overfitting; this is known as "pruning"

The best splits for regression typically minimize the mean-squared error of the resulting nodes after each split. Predictions come from the node that corresponds to the observed predictors for an observation. For regression trees, the mean of the corresponding final node is typically used, but this results in two issues for imputation. First, the mean predictions usually result in imputed values with too little variance within an implicate. Second, the mean imputations are deterministic and do not allow for multiple imputation. Instead, we sample the predicted value from the corresponding final node to impute liquid assets or net worth.

Random Forests

Random forest models (Breiman 2001) ensemble—meaning to combine the predictions from many models—decision trees. Typically, the process outlined above is repeated 100 to 1,000 times without pruning and with two important changes. First, each tree is estimated using a bootstrap sample of the training data. This adds a random component into model estimation, but the estimated trees are often highly correlated. Second, before each split, a random sample of predictors is considered. Sampling decorrelates the trees.

The main hyperparameter forests are the number of trees ensembled, the proportion of values bootstrapped sampled, and the number of predictors sampled before each split. We determine these values through hyperparameter tuning. Bootstrap sampling observations and randomly sampling predictors before each split reduces the predictive accuracy of each tree, but the ensemble of many imperfect trees often outperforms one decision tree.

Predictions are traditionally the mean of the predicted values from the ntree decision trees. For stochastic imputation, we randomly sample one value from the ntree decision trees.

Notes

[1] See the Urban Institute's "Financial Health and Wealth Dashboard," where local-level asset data are released: https://apps.urban.org/features/financial-health-wealth-dashboard/ (October 6, 2022).

[2] Ana Hernández Kent and Lowell R. Ricketts, "The Real State of Family Wealth: Quarterly Trends in Average Wealth and Demographic Wealth Inequality," Federal Reserve Bank of St. Louis, November 29, 2022, https://www.stlouisfed.org/institute-for-economic-equity/the-real-state-of-family-wealth.

References

Bhutta, Neil, Jesse Bricker, Andrew C. Chang, Lisa J. Dettling, Sarena Goodman, Joanne W. Hsu, Kevin B. Moore, Sarah Reber, Alice Henriques Volz, and Richard A. Windle. 2020. "Changes in U.S. Family Finances from 2016 to 2019: Evidence from the Survey of Consumer Finances." *Federal Reserve Bulletin* 106 (5): 1–42.

Blum, Avrim, Adam Kalai, and John Langford. 1999. "Beating the Hold-Out: Bounds for K-Fold and Progressive Cross-Validation." In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 203–08. Santa Cruz: University of California, Santa Cruz.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264): 1073–76.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45:5–32.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* New York: Chapman and Hall.

Burman, Prabir. 1989. "A Comparative Study of Ordinary Cross-Validation, V-Fold Cross-Validation and the Repeated Learning-Testing Methods." *Biometrika* 76 (3): 503–14.

CFED (formerly Corporation for Enterprise Development). 2015. "Technical Documentation: Generating Household Wealth and Financial Access Estimates for Local Geographies." Washington, DC: CFED.

Chen, Anqi, Alicia H. Munnell, and Geoffrey H. Sanzenbacher. 2018. "How Much Income Do Retirees Actually Have? Evaluating the Evidence from Five National Datasets." Working paper #2018-14. Chestnut Hill, MA: Center for Retirement Research at Boston College.

Little, Roderick J. A., and Donald Rubin. 2020. "Statistical Analysis with Missing Data." 3rd ed. Hoboken, NJ: Wiley.

Raghunathan, Trivellore E., James M. Lepkowski, John Van Hoewyk, and Peter Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27 (1): 85–96.

Ratcliffe, Caroline, Cary Lou, Diana Elliott, and Signe-Mary McKernan. 2017. "Financial Health of Residents: A City-Level Dashboard." Urban Institute.

Rubin, Donald B. 1978. "Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse." *Proceedings of the survey research methods section of the American Statistical Association* (Vol. 1, pp. 20-34). Alexandria, VA, USA: American Statistical Association.

Rubin, Donald B., and J. L. Schafer. 1990. "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data." In *ASA 1990 Proceedings of the Statistical Computing Section*, 83–88. Alexandria, VA: American Statistical Association.

Ruggles, Steven, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2021. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V11.0.

Therneau, Terry M., and Elizabeth J. Atkinson. 2019. "An Introduction to Recursive Partitioning Using the RPART Routines." Rochester, MN: Mayo Foundation.

van Buuren, Stef, Jaap PL Brand, Catharina GM Groothuis-Oudshoorn, and Donald B. Rubin. 2006. "Fully Conditional Specification in Multivariate Imputation." *Journal of Statistical Computation and Simulation* 76 (12): 1049–64.

van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Yeo, In-Kwon, and Richard A. Johnson. 2000. "A New Family of Power Transformations to Improve Normality or Symmetry." *Biometrika* 87 (4): 954–59.

About the Authors

Aaron R. Williams is a senior data scientist in the Income and Benefits Policy Center at the Urban Institute, where he works on retirement policy, microsimulation models, data privacy, and data imputation methods. He has worked on Urban's Dynamic Simulation of Income (DYNASIM) microsimulation model, the Social Security Administration's Modeling Income in the Near Term (MINT) microsimulation model, and the Tax Policy Center's synthesis of individual tax records. He holds a BS in economics from Virginia Commonwealth University, a BA in music from Virginia Commonwealth University, and an MS in mathematics and statistics from Georgetown University, where he is currently an adjunct professor in the McCourt School of Public Policy.

Mingli Zhong is a research associate in the Center on Labor, Human Services, and Population at the Urban Institute. She is also a visiting scholar at the Wharton School of the University of Pennsylvania. Zhong received the Equity & Inclusion Young Professionals Fellowship from the Association for Public Policy Analysis and Management (APPAM) in 2022. Her doctoral dissertation received the Social Security Administration Dissertation Fellowship Program in Retirement and Disability Research and Robert R. Nathan Fellowship. Zhong's research focuses on retirement savings, wealth, and debt. She received a PhD from the Wharton School of the University of Pennsylvania.

Breno Braga is a labor economist and principal research associate in the Center on Labor, Human Services, and Population. His research has covered topics such as the role of local conditions in asset accumulation and the local factors associated with debt in collections. His articles have been published in academic journals, including the *Journal of Labor Economics*. Braga received his MA in economics from the Pontifical Catholic University of Rio de Janeiro and his PhD in economics from the University of Michigan.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

URBAN INSTITUTE · ELEVATE · THE · DEBATE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org