Methodological Advances in Microdata Imputation: Quantile Regression Forests for Wealth Imputation

María Juaristi[*] Nikhil Woodruff[*] Max Ghenis[*]

June 16, 2025

Abstract

Policy analysts are regularly faced with the difficulty of peforming policy analysis that relies on wealth wihout a comprehnsive datataset capturing representative demographic, income and asset data across records. To address this challenge, this paper evaluates the methodological advantages of Quantile Regression Forests (QRF) for wealth imputation from the Survey of Consumer Finances (SCF) to the Current Population Survey (CPS), two US microdata surveys. We demonstrate that QRF outperforms traditional imputation approaches by preserving conditional distributions rather than merely conditional means, a critical distinction that makes distributional analyses of policy reforms under microdata more accurate. Our empirical analysis, implemented through the `microimpute` package, provides evidence that QRF reduces bias in wealth imputations, achieving a 20.5% reduction in average quantile loss (a measure of error against the full conditional distribution instead of central estimates) compared to OLS, 14.8% compared to Hot Deck Matching and a 6% reduction compared to Quantile Regression. Using 5-fold cross-validation on 22,975 SCF households, QRF achieves an average quantile loss across all quantiles of 6.6m, demonstrating superior distributional accuracy, particularly in the 10th-80th percentile range. While we focus on wealth in our experimental approach, these results suggest QRF's advantages extend to any skewed variable requiring distributional preservation, including consumption, medical expenses, and other heavy-tailed economic measures. These technical improvements have substantial implications for microdata enhancement and fiscal policy research, enabling analysis from wealth tax values and asset-dependent SNAP or Medicaid qualifications to policy impact across wealth deciles. We release our open-source `microimpute` package to facilitate microimputation across the field, providing automated method comparison, hyperparameter tuning, and survey weight integration capabilities that streamline the imputation workflow for complex survey data.

[*] PolicyEngine

1 Introduction

Microsimulation models and detailed microdata analyses are essential tools for understanding the distributional impacts of policies and social changes. These analyses require data that accurately represent both the demographic composition of a population and its economic circumstances. However, available data sources, particularly large-scale surveys, often suffer from missing data due to item nonresponse (Dempster and Rubin, 1983) or as a result of survey design. If not appropriately addressed, missing data can introduce substantial bias, undermining the validity of research conclusions (Graham, 2009) or limiting analysis opportunities altogether, increasing the risk of ineffective policy decisions and unintended consequences.

This problem extends beyond item nonresponse to systematic underreporting of certain income types in surveys. Evidence from fiscal and financial surveys in the UK demonstrate how data imputation can improve data quality. Dividend income in the Family Resources Survey is severely underreported, with the survey not even collecting data about directors' dividend incomes until the 2021-2022 survey year (Department for Work and Pensions, 2023). The UK's "SPI adjustment" methodology addresses this by replacing survey responses with administrative tax data for high earners, revealing income distributions previously hidden by measurement error (Advani et al., 2023). These examples illustrate that imputation from administrative sources does not merely fill gaps but can provide superior data quality compared to self-reported survey responses, particularly for sensitive financial variables prone to underreporting.

Traditional imputation approaches struggle with wealth data's right-skewness, heavy tails, and non-linear relationships with demographic and economic predictors. Wealth's heterogeneous interaction with income complicates imputation. Advani and Summers (Advani and Summers, 2020) demonstrated that capital gains, a key component of wealth changes, are distributed across the income spectrum with substantial volatility, finding that even individuals at the 80th income percentile have only a 1% probability of realizing any taxable gains, while those who do receive gains show extreme variability in amounts. These characteristics fundamentally violate assumptions underpinning conventional methods like Ordinary Least Squares (OLS) and Quantile Regression, resulting in significant distortions that undermine policy analysis (Meinshausen and Ridgeway, 2006).

This paper demonstrates that Quantile Regression Forests (QRF) provides superior performance for wealth imputation between the Survey of Consumer Finances (SCF) and the Current Population Survey (CPS). By modelling entire conditional distributions rather than conditional means alone, QRF preserves critical distributional features of wealth data. We implement this approach through the `microimpute` package, a specialised tool developed for survey data imputation that provides a complete pipeline for imputation and analysis, tailored to the dataset at hand. This package automates the comparison of four imputation methods, namely QRF, OLS, Hot Deck Matching, and Quantile Regression, automatically selecting the one achieving lowest average quantile loss to perform the final wealth impuation.

The remainder of this paper is organized as follows: Section 2 reviews the statistical properties of wealth microdata and the evolution of imputation techniques in the literature, discussing the strengths and limitations of the four methods evaluated. Section 3 describes our data sources (SCF and CPS) and their characteristics. Section 4 presents the `microimpute` package in detail. Section 5 presents our empirical results. Section 6 discusses implications and limitations, and Section 7 concludes. Our analysis makes two key contributions:

- An open-source microimputation package that facilitates the evaluation of multiple imputation methods tailored to specific dataset needs
- A validation framework comparing novel methodological approaches to traditional imputation methods demonstrated on statistically challenging data like wealth distributions
- A demonstration of QRF's advantages for wealth imputation, achieving better distributional estimates and a reduction in average quantile loss compared to traditional methods

2 Background

Effective imputation requires understanding both the data's distributional properties and the techniques available to handle them. This section first explores the statistical properties of wealth microdata that challenge imputation. It then reviews the literature on microdata imputation methods, tracing their development and practical applications.

2.1 Statistical properties of wealth distributions and imputation challenges

Wealth microdata present unique statistical challenges that can render innacurate policy analyses when using traditional imputation methods.

1. **High skewness and concentration**: Wealth distributions are typically right-skewed, with a small percentage of households holding a large share of total net worth (Chen et al., 2020). This concentration means that imputation models assuming normality can perform poorly, biasing estimates of wealth aggregates and inequality (Lun and Khattree, 2019).
2. **Outliers and extreme values**: Legitimate extreme values are common and can unduly influence parametric imputation models. Robust methods or data transformations are often necessary (Chen et al., 2020).
3. **Non-linear relationships**: Wealth's relationship with predictor variables such as age, education, and income is highly non-linear (Zillow Group, 2024), requiring more flexible imputation methods.

2.2 Traditional microdata imputation methods

Among traditional imputation methods, we have selected three to examine, namely Ordinary Least Squares regression, Quantile regression, and Hot Deck Matching, due to their diverse approaches to imputation and relevance in the literature. We also study more novel approaches like Quantile Regression Forests, which provides an opportunity for more robust microdata imputation. In the following sections we discuss the methodological details of each method.

2.2.1 Ordinary Least Squares (OLS)

OLS imputation predicts missing values in a recipient dataset based on a linear regression model trained on a donor dataset. The model is specified as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i,$$

where y_i is the variable to be imputed for observation i in the recipient dataset, x_{i1},...,x_{ip} are predictor variables common to both donor and recipient datasets, β_0,...,β_p are coefficients estimated from the donor dataset, and ε_i is the error term (Bruch, 2023). In deterministic regression imputation, the imputed value is typically the expected value of y_i. An alternative, stochastic regression imputation, adds a randomly drawn residual (from the donor model's residuals or a normal distribution with estimated variance σ^2) to the predicted value: $y_{imputed} = y_i + e_i$, where e_i $N(0, \sigma^2)$. Stochastic imputation aims to preserve the variability of the original data better than deterministic imputation (Anil).

OLS assumes linearity, homoscedasticity (constant variance of errors), and normally distributed errors, all of which are typically violated by skewed wealth data (Von Hippel, 2007). While OLS imputation might yield consistent estimates for means and variances even with non-normal data, it can produce considerable bias for shape-dependent estimands like percentiles or skewness coefficients (Von Hippel, 2007). Furthermore, deterministic OLS imputation systematically underestimates the true variance of the completed data (Barceló, 2008).

2.2.2 Quantile Regression (QR)

Quantile regression (QR) models the conditional quantiles (e.g., median, 10th percentile, 90th percentile) of a response variable, Y, given a set of predictors, X (Koenker and Bassett, 1978). The model for the τ-th quantile is:

$$Q_Y(\tau|X) = \beta_0(\tau) + \beta_1(\tau)x_1 + \beta_2(\tau)x_2 + ... + \beta_p(\tau)x_p$$

When inputting from a donor to a recipient dataset, QR models for various quantiles τ are fitted on the donor dataset using common predictor variables. These fitted models are then applied to the recipient dataset to predict the conditional quantiles for observations with missing data (Parker). To generate a single imputed value, one might impute the conditional median ($\tau = 0.5$) or draw a value from an estimated conditional distribution constructed from multiple quantile predictions (Wei et al., 2014). For instance, a random quantile $\tau*$ can be selected from a uniform distribution, and the imputed value computed by interpolating between the estimated responses for quantiles directly above and below $\tau*$ (Chen and Yu, 2007).

QR is more robust to outliers and better at handling skewed distributions and heteroscedasticity than OLS because it does not make strong assumptions about the error distribution (Zhao et al., 2023). This makes it particularly suitable for economic variables like wealth, where relationships may vary across the distribution, better preserving its overall shape (Kleinke and Reinecke, 2020). However, while more robust than OLS, standard quantile regression still assumes linear relationships between predictors and the outcome at each specific quantile (Meinshausen and Ridgeway, 2006). It requires fitting separate models for different quantiles, which can increase complexity, and may struggle with high-dimensional data or very complex non-linear patterns (Meinshausen and Ridgeway, 2006).

2.2.3 Hot Deck Matching imputation

Hot Deck imputation replaces missing values in a recipient record with an observed value from a "similar" donor record. When imputing from a donor dataset to a recipient dataset, "similarity" is established using variables common to both datasets (D'Orazio et al., 2021). This often involves defining adjustment cells, which are groupings of records in both datasets based on shared categorical variables (e.g., gender, education level). Donors to be matched to a receiver record are then selected from the corresponding cell in the donor dataset (Chen and Shao, 2000). For continuous or mixed data, a distance metric (e.g., Euclidean, Mahalanobis) is calculated between a recipient record and potential donor records based on common variables. The donor record with the smallest distance is chosen (D'Orazio et al., 2021). Once the matching is done, the exact value from the selected donor record in the donor dataset is then used to fill the missing item in the recipient dataset (Andridge and Little, 2010).

Hot Deck methods are non-parametric and do not require explicit model specification, making them robust to weak distributional assumptions (D'Orazio et al., 2021). Since imputed values are actual observed values from the donor dataset, they are inherently plausible and can help preserve the marginal distribution of the imputed variable if donors are well-matched (Andridge and Little, 2010). Nonetheless, a critical challenge is ensuring an adequate and representative donor pool in the donor dataset for all types of recipients in the target dataset. This is particularly difficult for extreme wealth values, where suitable donors may be scarce or unrepresentative, leading to biased imputations or overuse of certain donors (Haziza, 2009). Additionally, it may struggle to maintain complex multivariate relationships, especially when imputing across datasets with different underlying structures or sampling designs (Siddique and Belin, 2008). The effectiveness is highly dependent on the choice of matching variables (Office of Tax Analysis, 2012), as well as "similarity" metrics. Poorly defined cells or metrics can lead to inappropriate donor selection and biased results (Andridge and Little, 2010). Most critically for policy analysis, Hot Deck methods provide only observed values rather than capturing the full conditional distribution, making it impossible to properly quantify imputation uncertainty and data variance. This affects winner-loser analysis and the evaluation of distributional impacts of policy changes where small differences in imputed values can dramatically affect conclusions about who benefits or loses (Rubin, 1987).

2.2.4 Quantile Regression Forests (QRF)

Quantile Regression Forests (QRF) (Meinshausen and Ridgeway, 2006) extend Random Forests (RF), which are ensemble learners that build multiple decision trees and excel at capturing non-linearities and interactions (Breiman, 2001), to estimate conditional quantiles. When imputing from a donor dataset to a receiver dataset, QRF models are trained on the donor dataset using predictor variables common to both. Instead of storing only mean values in terminal nodes (as in standard RF regression), QRF retains all observed outcome values for the training instances that fall into each terminal leaf of each tree (Meinshausen and Ridgeway, 2006).

For a given point x (representing the predictor values for an observation in the recipient dataset), the conditional distribution function $\hat{F}(y|X = x)$ of the target variable Y is estimated as:

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) \cdot \mathbf{1}_{Y_i \leq y},$$

where Y_i are the observed values from the donor dataset, $1_{Y_i} \leq y$ is an indicator function (1 if $Y_i \leq y$, 0 otherwise), and $w_i(x)$ represents the weight assigned to each donor observation i. This weight is derived from the forest structure; specifically, $w_i(x)$ is positive if observation i from the donor set falls into the same terminal node as x in any tree, and its magnitude reflects how often this co-occurrence happens across all trees in the forest (Kleinke and Fritsch, 2023).

The τ-th conditional quantile is then estimated by finding the minimum value y for which the estimated cumulative distribution function $\hat{F}(y|X = x)$ is greater than or equal to τ:

$$\hat{Q}_\tau(y|X = x) = \inf y : \hat{F}(y|X = x) \geq \tau$$

(Meinshausen and Ridgeway, 2006). This allows for the estimation of any quantile without retraining the model (Woodruff and Ghenis, 2024). For imputation, particularly multiple imputation, values can be drawn randomly from this estimated conditional distribution for each observation in the recipient dataset requiring imputation. This process helps reflect the uncertainty inherent in the imputation.

This approach offers several critical advantages for microdata imputation, and wealth imputation more specifically:

1. **Distribution preservation**: By modeling the entire conditional distribution, QRF is adept at capturing and preserving the right-skewness and heavy tails characteristic of wealth distributions (Meinshausen and Ridgeway, 2006).

2. **Non-linear relationship handling**: The tree-based structure of RF, and thus QRF, automatically handles complex non-linear relationships between predictors (e.g., demographic variables) and the imputation target (e.g., wealth) without requiring explicit transformation or pre-specification of these functional forms (Tang and Ishwaran, 2017).

3. **Automatic interaction detection**: QRF naturally incorporates interactions between predictor variables, as tree splitting rules inherently consider combinations of features (Tang and Ishwaran, 2017).

4. **Robustness to outliers**: The ensemble nature of random forests and the focus on quantiles (rather than just the mean) make QRF less sensitive to extreme outliers in the donor data that might distort parametric models (Tang and Ishwaran, 2017).

5. **Single model for all quantiles**: Unlike standard QR, which requires fitting separate models for different quantiles, QRF produces an estimate of the entire conditional distribution from a single trained model, making it computationally more efficient (Meinshausen and Ridgeway, 2006).

When imputing from a survey with specific design features to a more general survey with somewhat different distributional properties, QRF's ability to learn localized relationships in the predictor space can be advantageous. If survey weights from the donor are incorporated during the QRF training (e.g., by influencing tree construction or the sampling of observations for bootstrap aggregation), the model can learn to represent the oversampled segments appropriately. The subsequent prediction onto the receiver dataset, which has its distinct sample structure, then relies on the learned conditional distributions. The challenge lies in ensuring that the relationships learned from the donor are transportable and applicable to the recipient, and that the resulting imputations in the recipient dataset, when combined with its own survey weights, yield valid population estimates. While QRF itself doesn't explicitly model survey design features like clustering or stratification in a formal statistical sense unless specifically adapted, its flexibility in capturing complex data structures can implicitly handle some of the heterogeneity introduced by such designs (Hao and Naiman, 2007), making it much stronger than other more limited imputation approaches.

Nonetheless, QRF has its own limitations. Given the data-splitting nature of a tree, certain terminal nodes may receive a single or very few extreme training samples. When imputing, all the data points from the receiver dataset that land on those leaves will likely receive the same or very similar values for the imputed variable, even if there are differences in predictor values between them. In practice, this means that if the donor and receiver datasets have distributional differences, imputations at the extreme tails of the receiver dataset may suffer. Data points that are not necessarily unusual or extreme might receive extreme imputations, while truly extreme values in the receiver dataset are at risk of not being regarded as so if the donor dataset had a narrower range of values for the training variable.

3 Data

3.1 Survey of Consumer Finances

The Survey of Consumer Finances (SCF), sponsored by the Federal Reserve Board, is a triennial survey providing detailed information on U.S. households' assets, liabilities, income, and demographic characteristics. Its dual-frame sample design includes a standard national area-probability sample and a list sample deliberately oversampling wealthy households to better capture the skewed wealth distribution (Barceló, 2006). The SCF is a benchmark for wealth imputation research due to its detailed financial data and the known complexities arising from its design and the nature of wealth. Item nonresponse in public-use SCF datasets is addressed by the Federal Reserve through a multiple imputation approach that generates five complete datasets with different imputed values, using sequential regression-based procedures that incorporate range constraints, logical data structures, and empirical residuals to preserve the complex multivariate relationships inherent in wealth data (Kennickell, 1998).

Specifically, we use the 2022 summarized SCF as our donor dataset.

3.2 Current Population Survey

The Current Population Survey (CPS), conducted by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics, is a monthly survey primarily focused on labor market information. The Annual Social and Economic Supplement (ASEC) collects detailed annual income data and some information on assets and liabilities, though far less comprehensively than the SCF. The CPS uses a national probability sample and is a key source for income and poverty statistics. Missing data, particularly for income items, is also a feature of the CPS.

3.3 Comparative analysis and characteristics for imputation

Beyond wealth data's inherent challenges, imputing between SCF and CPS presents additional complications due to their differences in scope, design, and wealth data measurement. These complications include:

1. **Sampling approach**: The SCF employs a dual-frame sample design, deliberately oversampling wealthy households through a list sample derived from tax returns. The CPS uses a more standard probability sample that does not effectively capture the upper tail of the wealth distribution (Bryant, 2023).

2. **Sample size and frequency**: The SCF typically includes about 4,500-6,000 households and is conducted triennially, while the CPS surveys approximately 60,000 households monthly.

3. **Wealth variable coverage**: The SCF collects extremely detailed information on financial assets and liabilities, while the CPS survey design does not request most of this data from its respondents, making direct matching of asset categories difficult.

These structural differences create challenges for transferring wealth information between surveys through traditional imputation methods. The predictors available in both datasets may not involve linear relationships with wealth, and the surveys may have vastly different sample sizes at various points along the wealth distribution.

4 Methodology

4.1 `microimpute` package implementation

`microimpute`[1] is PolicyEngine's specialized Python framework that enables variable imputation through multiple statistical methods, providing a consistent interface for comparing and benchmarking different imputation approaches using quantile loss calculations.

4.1.1 Core capabilities

The package currently supports four primary imputation methods: Hot Deck Matching, Ordinary Least Squares Linear Regression, Quantile Regression Forests (QRF), and Quantile Regression. This approach allows researchers to systematically evaluate which technique provides the most accurate results for their specific dataset and research objectives.

4.1.2 Key features for microimputation

`microimpute` addresses imputation challenges between complex survey datasets through several specialized features:

1. **Survey data weights integration**: Handles survey data weights through sampling to ensure that models are trained on a donor data set representative of the true distribution.

2. **Method comparison and benchmarking**: Allows researchers to easily compare different approaches and automatically determine the method providing the most accurate results.

3. **Flexible methodological set-up**: Enables advanced usage through specified hyperparameter setting and tuning.

4. **Quantile-based evaluation**: Uses quantile loss calculations to assess imputation quality across different parts of the distribution.

5. **Autoimputation**: Provides an integrated imputation pipeline that tunes method hyperparameters to the specific datasets, compares methods, and selects the best-performing to conduct the requested imputation in a single function call.

4.1.3 Implementation details

`microimpute`'s QRF implementation extends `scikit-learn`'s Random Forest to provide full conditional quantile estimation, enabling stochastic imputation that preserves distributional properties rather than relying solely on point estimates. OLS and QuantReg methods are implemented using `statsmodels`, while Matching uses the R `StatMatch` package's Hot Deck Matching capabilities. The `microimpute` package is designed to be modular, allowing for easy extension with additional imputation methods in the future.

[1] Complete documentation, implementation details, and usage examples are available at https://policyengine.github.io/microimpute/.

5 Results

5.1 Evaluation metrics

To properly assess imputation quality for wealth data, we employ quantile loss to evaluate performance across quantiles. Quantile (or pinball) loss evaluates how well a model predicts a chosen conditional quantile by assigning an asymmetric penalty that depends on the sign of the forecast error and the target quantile (Koenker and Bassett, 1978). Formally, for residual $e_i = y_i - \hat{y}_i$ and quantile level $\tau \in (0, 1)$, the loss is

$$L_\tau(e_i) = \max\bigl(\tau\, e_i,\ (\tau - 1)\, e_i\bigr);$$

thus, under-prediction of an upper-tail quantile (large positive e_i) is penalized at rate τ, whereas over-prediction is penalized at rate $\tau - 1$. Minimising this piece-wise linear function ensures that, in expectation, exactly τ percent of true outcomes fall below the model's prediction, giving the estimator its distribution-calibrating property first formalised by Koenker and Bassett (Koenker and Bassett, 1978). Because the weights differ for positive versus negative errors, quantile loss captures directional bias. This is a crucial feature when imputing highly skewed variables like wealth, where large under-estimates in the right tail must be discouraged more strongly than equal-sized over-estimates. Compared with mean-squared or mean-absolute error, which optimise the conditional mean or median and penalise errors symmetrically, quantile loss directly targets the entire conditional distribution and remains robust to outliers, making it the appropriate metric for assessing the performance of imputation models when distribution-awareness is a priority (Ghenis, 2018).

Additionally, we ensure that the resulting wealth distributions resemble that of the donor wealth distribution in the SCF. The SCF includes household weights to quantify how representative each recorded household and its data are. Thus, by sampling the SCF distribution according to its weights, we can get an accurate estimate of the true US wealth distribution, which simultaneously serves as a visual check for imputed wealth distributions.

Lastly, by evaluating the average wealth by disposable income decile that results from each method's imputations, we can compare whether and how sensibly wealth imputations correlate with income as expected for each method's results, with households that have the highest disposable income also having the highest wealth.

5.2 Experimental setup

Using Microimpute's `autoimpute` function, we evaluate QRF against OLS, standard Quantile Regression, and hot Deck Matching in net worth imputation. To create a ground truth for evaluation, we employ cross-validation, splitting the SCF into 5 folds, or subsets. For each model at a time, we:

1. Mask the wealth values of one holdout subset at a time
2. Tune the hyperparameters of those models that have a flexible set-up to the SCF data set, namely Matching[2] and QRF[3]

[2] hyperparameters: `dist_fun`, `constrained`, `constr_alg`, `k`

[3] hyperparameters: `n_estimators`, `min_samples_split`, `min_samples_leaf`, `max_features`, `bootstrap`

    (a) Hyperparameter tunning is done through the `optuna` package, which paralelizes hyperparameter searches to identify the combination of hyperparameters that minimizes the average quantile loss across all quantiles for the SCF dataset.

3. Impute wealth (`networth`) onto the SCF holdout set using each method with the remaining four folds as the training data

    (a) SCF survey data weights (captured in the variable `household_weight`) are passed into `autoimpute` for sampling

    (b) Demographic and financial variables available in both surveys behave as predictors, after being preprocess for consistent variable names and encoding: `is_female`, `age`, `race`, `own_children_in_household`, `employment_income`, `interest_dividend_income`, `pension_income`

    (c) Each model performs imputation at 20 equally spaced quantiles

4. Compare the imputed values to the SCF training values, measuring quantile loss at each of the imputation quantiles

5. Average quantile loss across all folds and quantiles

This approach allows for direct comparison between methods on a common testing framework, providing a robust assessment of relative performance. The model with the lowest average quantile loss is chosen as the best performing and automatically selected to perform the final imputation onto the CPS data set, for which the full SCF will be used for training.

Next, the imputation from the SCF onto the CPS is replicated for the other three models that were not selected, with an identical setup that includes training on the full SCF, the same predictor variables, and imputations at the median quantile, avoiding introducing any bias toward a specific side of the distribution.

5.3 Imputation results

The cross-validation results demonstrate QRF's superior performance across the wealth distribution. QRF achieved an average quantile loss of 6.6m across all quantiles, outperforming OLS (8.3m), Hot Deck Matching (7.75m), and standard Quantile Regression (7.05m). This 20.5% improvement over OLS and 14.8% over Matching and 6.4% improvement over Quantile Regression represents a substantial gain in imputation accuracy.

The performance advantage was particularly pronounced in the middle portions of the distribution (10th-80th percentiles), where QRF maintained consistently lower quantile loss values. While Quantile Regression showed competitive performance at extreme quantiles (above the 85th percentile), QRF's overall consistency across the entire distribution made it the optimal choice for wealth imputation.

Microimpute's automated hyperparameter tuning contributed significantly to these results. The optimal QRF configuration identified through cross-validation included 200 trees and a minimum of 5 samples per leaf node for a split. These parameters balanced model complexity with generalization capability, preventing overfitting while capturing the nuanced relationships between demographic predictors and wealth outcomes.

5.3.1 Quantile loss

Comparing all four models to each other, we can evaluate performance through quantile loss, as well as comparing the overlap of the imputed wealth distribution with the original SCF data.

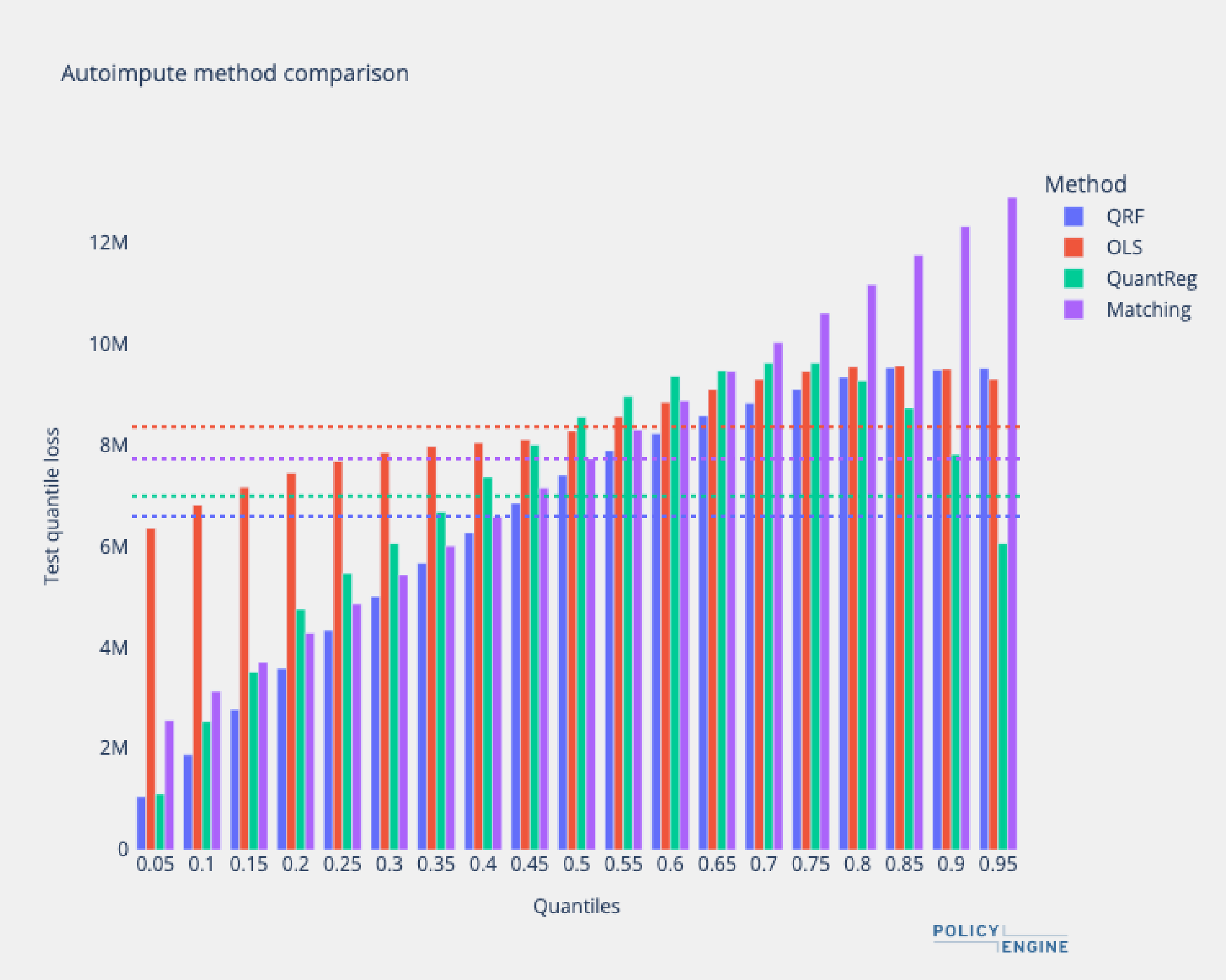

Figure 1: Cross-validation method performance at 20 equally spaced quantiles. Test quantile loss at each quantile is averaged across 5 cross-validation folds, and the dashed lines represent average quantile loss across all quantiles.

We observe how QRF proves to be the best-performing model out of the four. It significantly outperforms Matching and OLS on average, while only being outperformed by QuantReg past the 80th quantile. Matching's performance presents the perfect opportunity to see the quantile loss's directional bias at work. Because Matching does not incorporate quantile information into its imputation process in any way, it will match donor and receiver units, and thus impute values identically regardless of the quantile at work. The laddering behavior observed above results from the fact that Matching's wealth imputations consistently underpredict relative to the true wealth values, and this property is increasingly favored for quantiles below the median, and increasingly penalized as quantiles approach the right end of the distribution.

5.3.2 Imputed wealth distribution comparisons

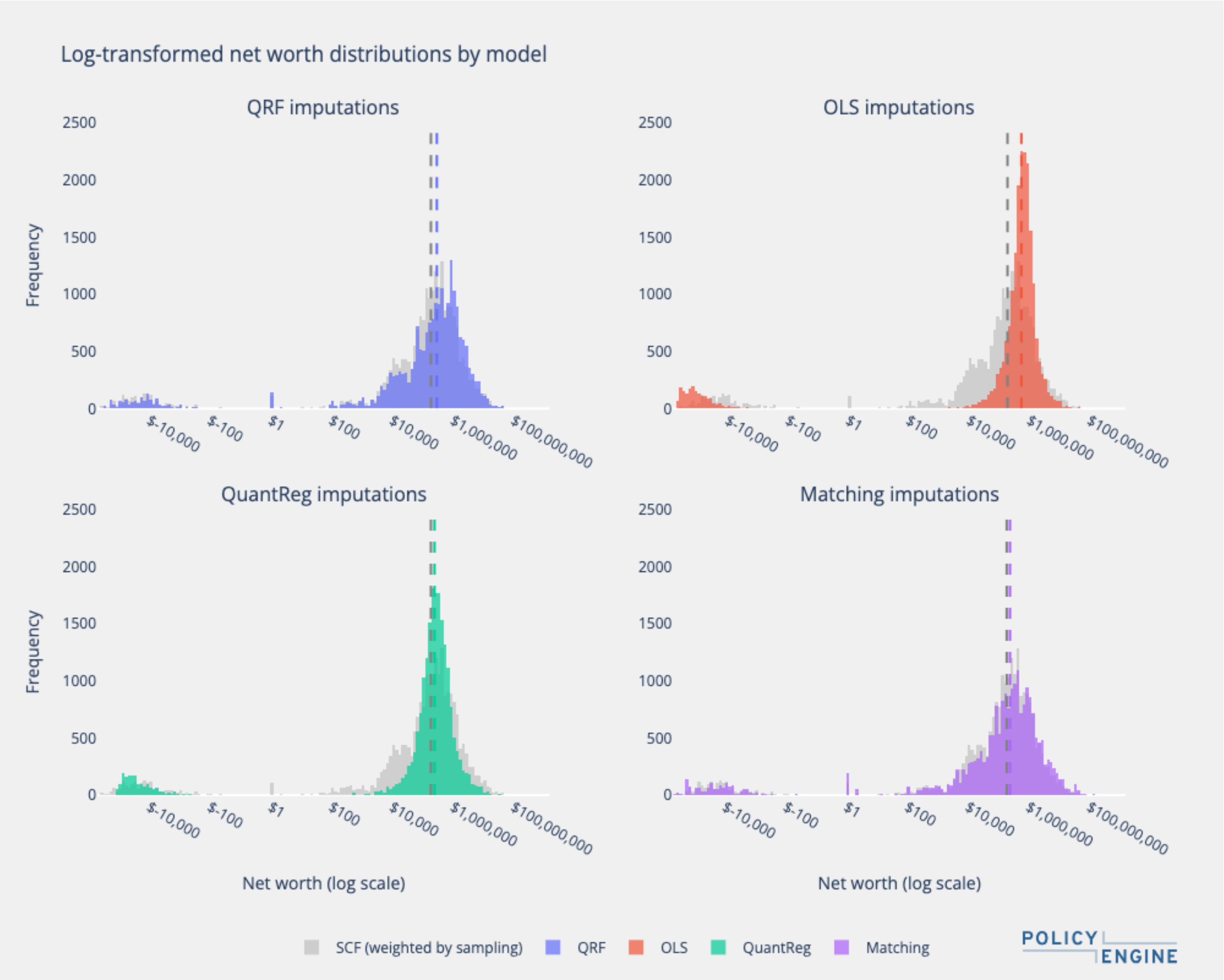

Figure 2: Weight-adjusted imputed wealth distributions model comparisons, relative to the donor SCF wealth distribution. Dashed lines represent median values.

By visually comparing the wealth distributions resulting from imputing with each method, and comparing them to the weighted donor distribution, we gain a more comprehensive understanding of imputation performance, moving past the test quantile loss average measured on the SCF. The distribution of wealth values imputed by QRF closely resembles the original SCF distribution, suggesting that QRF is not only the best method in terms of quantile loss but also strong when achieving distributional estimates close to the true wealth distribution in the United States. Because Matching directly takes values from the donor distribution and imputes them onto the receiver, it is unsurprising that its distribution closely resembles the donor distribution. However, Matching's performance in quantile loss confirms that even if it takes the exact values from the donor, it may struggle to uncover the non-linear relationship between the different predictors and wealth values, resulting in a seemingly correct data distribution but innacurate imputation given certain household characteristics. Meanwhile, OLS and QuantReg fail to capture the variability across the distribution and impute most values at the lower or higher ends of the distribution.

5.3.3 Distribution of wealth by disposable income decile

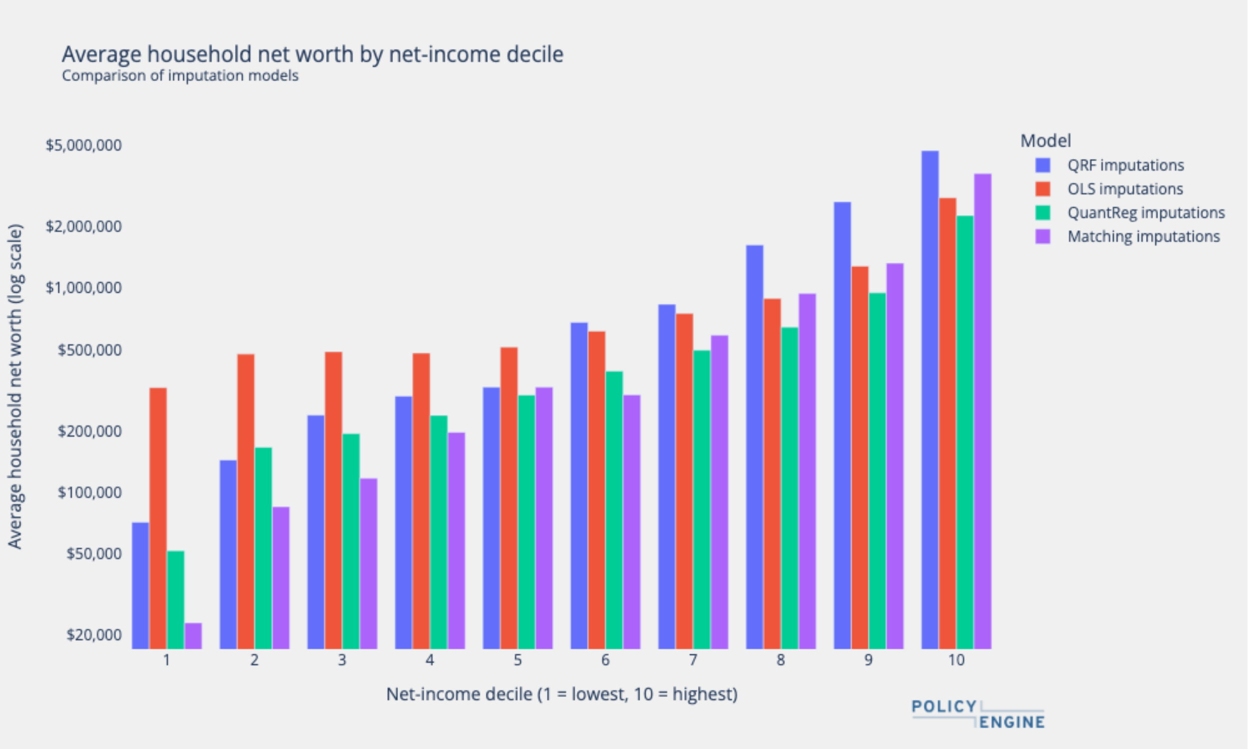

Figure 3: Average household net worth by disposable income decile model comparison. Households are divided into 10 equally sized groups based on their disposable income.

These results support the observations above, with QRF presenting the most consistent and plausible relationship to disposable income, with a gradually increasing average as the deciles increase. This plot also demonstrates the caveats of the other models, for example showing the lower variability in OLS's predictions, and Matching's consistent underprediction across deciles.

6 Discussion

This paper has demonstrated both theoretically and empirically that Quantile Regression Forests provide substantial advantages for microimputation, particularly through the example of wealth imputation from the SCF to the CPS. By preserving the full conditional distribution of wealth, QRF maintains the critical statistical properties of wealth data that traditional methods fail to capture.

6.1 Strengths powered by microimpute

The microimpute package's design philosophy and implementation choices provide several key advantages that contributed to the success of our wealth imputation analysis:

1. **Unified interface for method comparison**: microimpute's consistent API across all imputation methods enabled systematic benchmarking without implementation-specific biases. This standardization ensures that performance differences reflect genuine methodological advantages rather than implementation artifacts (PolicyEngine, 2025).

2. **Automated method selection**: The package's autoimpute function streamlines the imputation workflow by automatically comparing methods and selecting the best performer based on quantile loss metrics. This feature proved particularly valuable given the wealth data's complexity, as it removed subjective method selection and ensured optimal performance.

3. **Survey weight integration**: microimpute's native support for survey weights through stratified sampling ensures that imputation models properly represent population distributions. This capability is crucial when transferring information between surveys with different sampling designs, such as the SCF's oversampling of wealthy households.

4. **Quantile-aware evaluation**: By implementing quantile loss as the primary evaluation metric, microimpute directly addresses the challenges of skewed distributions. This metric's asymmetric penalty structure naturally prioritizes accurate imputation at distribution tails, where traditional metrics like Root Mean Squared Error (RMSE) often fail.

5. **Computational efficiency**: The package's optimized implementation enables processing of large microdata files while maintaining reasonable computation times. Cross-validation on the full SCF dataset, including QRF hyperparameter tuning, was completed in under 30 minutes on standard hardware, making iterative experimentation feasible.

6. **Open-source transparency**: As an open-source tool, microimpute allows full inspection and modification of imputation algorithms, promoting reproducibility and enabling custom extensions for specific research needs (PolicyEngine, 2025).

6.2 Limitations and future improvements

Despite the demonstrated advantages, several limitations warrant consideration for future development:

6.2.1 Current package limitations

While microimpute currently supports four imputation methods, expanding to include modern machine learning approaches such as neural networks, gradient boosting machines, or deep learning architectures could further improve performance, particularly for complex multivariate relationships (Alaa et al., 2024). The package would benefit from implementing ensemble methods that combine multiple imputation approaches, potentially leveraging the strengths of different methods across different parts of the distribution. Moreover, model selection and assessment could be enhanced with evaluation metrics additional to quantile loss, ensuring a thorough understanding of each model's behavior at every step.

6.2.2 QRF-specific challenges

The terminal node sparsity issue identified in Section 2.2.4 remains a fundamental limitation of tree-based methods. When few training samples reach certain terminal nodes, multiple observations in the recipient dataset may receive identical imputed values, potentially underestimating variability in extreme wealth categories. Future work could explore adaptive tree construction methods that ensure minimum node occupancy or hybrid approaches that combine QRF with parametric methods at distribution extremes.

These enhancements would position `microimpute` further as a comprehensive solution for survey data imputation challenges while maintaining its current strengths in ease of use and methodological rigor.

7 Conclusion

This paper has reviewed methodological advances in microdata imputation, focusing on techniques suitable for complex survey data like wealth. We highlighted the limitations of traditional methods and the advantages of novel approaches, particularly Quantile Regression Forests, for handling skewed distributions and non-linearities.

The key contributions of this work are the synthesis of current knowledge on imputation for challenging microdata, a detailed examination of QRF's suitability, and an introduction to the `microimpute` package. Our review suggests that QRF represents a step forward in preserving the statistical integrity of imputed microdata, increasing the robustness of economic analysis. The implementation of QRF in the `microimpute` package provides a practical tool for researchers seeking to combine detailed microdata across datasets.

Future research should continue to refine QRF for imputation, particularly in response to challenges like limited data at extreme quantiles. Comparative studies against other emerging techniques, like deep learning models (Alaa et al., 2024), are also vital. Continued innovation in imputation methodology supports the reliability of evidence-based research and policymaking.

References

Arun Advani and Andy Summers. Capital gains and UK inequality. Technical Report 1260, University of Warwick, Department of Economics, May 2020.

Arun Advani, Andy Summers, and Hannah Tarrant. Measuring top income shares in the

UK. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):241–265, 2023. doi: 10.1093/jrsssa/qnac008.

Ahmed M. Alaa, Boris Van Breugel, David Sontag, and Mihaela van der Schaar. Deep learning for missing data imputation: A comprehensive review and new perspectives. *Foundations and Trends in Machine Learning*, 17(2):139–301, 2024.

Rebecca R. Andridge and Roderick J. A. Little. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64, 2010.

Anil. Regression imputation - deterministic & stochastic. URL https://www.rpubs.com/anilpmm/regression_det_stoc. Retrieved May 23, 2025.

Cristina Barceló. Imputation of household wealth components in the spanish survey of household finances. Technical Report 0629, Banco de España, 2006.

Cristina Barceló. The impact of alternative imputation methods on the measurement of income and wealth: Evidence from the spanish survey of household finances. Technical Report 0829, Banco de España, 2008.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Christian Bruch. Imputation of missing values in survey data: An overview. Technical report, GESIS - Leibniz Institute for the Social Sciences, 2023.

Victoria Bryant. General description booklet for the 2015 public use tax file demographic file. Technical documentation, Statistics of Income Division, Internal Revenue Service, 2023. URL https://drive.google.com/file/d/1WoTU7OGEjYMO0KHsHvTTH0NwCc-kN5cE/view.

Jie Chen and Jun Shao. Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16(2):113–131, 2000.

Qizhi Chen and Keming Yu. Confidentiality protection by data synthesis using quantile regression and hot deck imputation. In *Proceedings of the Survey Research Methods Section*. American Statistical Association, 2007.

Shu Chen, David Haziza, and Christian Léger. Imputation for skewed data using a cube-root transformation. *Journal of Survey Statistics and Methodology*, 8(3):545–567, 2020.

Arthur P. Dempster and Donald B. Rubin. Introduction. In William G. Madow, Ingram Olkin, and Donald B. Rubin, editors, *Incomplete data in sample surveys: Vol. 2. Theory and bibliographies*, pages 3–10. Academic Press, 1983.

Department for Work and Pensions. Family resources survey: Financial year 2021 to 2022, July 2023. URL https://www.gov.uk/government/statistics/family-resources-survey-financial-year-2021-to-2022/.

Marcello D'Orazio, Marco Di Zio, and Mauro Scanu. Statistical matching and imputation of survey data with r. *Journal of Statistical Software*, 98(1):1–35, 2021.

Max Ghenis. Quantile regression: From linear models to trees to deep learning. Towards Data Science, 2018. URL https://medium.com/data-science/quantile-regression-from-linear-models-to-trees-to-deep-learning-af3738b527c3. Medium blog post.

John W. Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

Lingxin Hao and Daniel Q. Naiman. *Quantile regression*. Sage Publications, 2007.

David Haziza. Imputation and inference in the presence of missing data. In Danny Pfeffermann and C. R. Rao, editors, *Handbook of statistics, Vol. 29A: Sample surveys: Design, methods and applications*, pages 215–246. Elsevier, 2009.

Arthur B. Kennickell. Multiple imputation in the survey of consumer finances. Technical report, Board of Governors of the Federal Reserve System, September 1998. Prepared for the August 1998 Joint Statistical Meetings, Dallas, TX.

Kristian Kleinke and Markus Fritsch. Robust multiple imputation based on quantile forests. In *GESIS Workshop on Longitudinal Data Analysis*, 2023.

Kristian Kleinke and Jost Reinecke. Multiple imputation of incomplete zero-inflated count data. *Statistica Neerlandica*, 74(4):515–539, 2020.

Roger Koenker and Gilbert J. Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Zhixin Lun and Ravindra Khattree. Multiple imputation for skewed multivariate data: A marriage of the mi and copula procedures. In *Proceedings of the SAS Global Forum 2019 Conference*, number Paper 3605-2019. SAS Institute Inc., 2019. URL https://support.sas.com/resources/papers/proceedings19/3605-2019.pdf.

Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Office of Tax Analysis. Revenue estimating models at the u.s. treasury department. Technical Report Technical Paper 12, U.S. Department of the Treasury, 2012.

D. Parker. Missing data and quantile regression [lecture notes]. URL https://www.reed.edu/economics/parker/312/online_slides/312_4-29.pdf. Retrieved May 23, 2025.

PolicyEngine. Microimpute documentation, 2025. URL https://policyengine.github.io/microimpute/. Retrieved from.

Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 1987.

Juned Siddique and Thomas R. Belin. Multiple imputation using chained equations: What is it and how does it work? *American Journal of Public Health*, 98(4):644–659, 2008.

Fei Tang and Hemant Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

Paul T. Von Hippel. Should a normal imputation model be modified to impute skewed variables? Technical report, LBJ School of Public Affairs, University of Texas at Austin, 2007.

Ying Wei, Yang Ma, and Raymond J. Carroll. Multiple imputation in quantile regression. *Biometrics*, 70(1):196–205, 2014.

Nikhil Woodruff and Max Ghenis. Enhancing survey microdata with administrative records: A novel approach to microsimulation dataset construction. Technical report, 2024.

Qian Y. Zhao, Shuangge Ma, and Lan Wang. Quantile regression for electronic medical records analysis with nonignorable missing responses. *Biometrics*, 79(3):2036–2048, 2023.

Zillow Group. quantile-forest: Scikit-learn compatible quantile regression forests, 2024. URL https://zillow.github.io/quantile-forest/.