Weight-aware imputation for policy microsimulation

Max Ghenis* María Juaristi*

Abstract

Policy microsimulation models rely on imputation to attach variables one survey observes to records from another, and the imputed file is then reweighted — calibrated to administrative totals, filtered to subpopulations, stressed under reform scenarios. Common imputation practice fits models to survey records while ignoring their design weights, which biases imputed conditionals toward the sample rather than the population and plants extreme low-weight donor values that a later reweighting can amplify. We present the imputation operator of POPULACE, POLICYENGINE's open-source microdata stack: a regime-gated, sequentially-chained, weighted-bootstrap quantile-regression-forest estimator whose interface makes an unweighted fit impossible to request by accident. To evaluate it we formalize a *population-view harness*: one latent population observed through survey views, with any candidate weighted file scored by projecting it through each view against that survey's holdout — a strictly proper joint score, a coverage axis invariant to reweighting of the candidate, a classifier two-sample test, and an uncapped tail block on the imputed columns, which we add after showing that subsampled joint geometry certifies files whose top percentile is wrong by a factor of two. Under this harness and paired repeated holdouts we benchmark the estimator against standard survey-imputation methods — unweighted and weighted quantile regression forests, ordinary least squares, linear quantile regression, and nearest-neighbour hot-deck statistical matching — on within-SCF wealth, SCF-onto-CPS population synthesis, zero-inflated CPS income components, and six cross-domain datasets. Ablations attribute performance to each design choice, and the experiment runs at two scopes: a minimal six-predictor instantiation and a populace-scale one (ten shared predictors, a four-component chained balance sheet, three pooled receiver vintages). Where the survey design is informative, weighting dominates: the identical estimator fit unweighted is six

*PolicyEngine. Corresponding author: max@policyengine.org

times worse on net-worth marginal fit within the donor survey, and on the population view its imputed net-worth 99th percentile is 1.9 to 2.3 times the holdout's — an inflation the geometry metrics tie on and only the tail block detects. Regime gating carries zero-inflated components (0.6 versus 4.4 percentage points of dividend zero-share error against the same forest ungated, and 25–84 points against linear and Gaussian baselines). Chaining is a joint property that binds at scale: with two targets it is indistinguishable, while on the four-component balance sheet fitting targets independently costs 5.8 points of two-sample AUC — the classifier detects balance sheets that do not add up. Hot-deck matching is a genuine near-peer on the population view; marginal metrics alone rank an unconditional weighted draw at or near the top throughout, and at populace scale the classifier separates it almost perfectly (AUC 0.97 versus the candidate's 0.76). A closed-form reweighting stress test shows every tail-capable method matching the held-out truth's own concentration while Gaussian-residual imputation understates it by factors of 7 to 11 — missing tails, not robustness. Wealth imputation is what makes asset-tested programs such as Supplemental Security Income simulable on the CPS at all, and the population-view results identify which methods preserve the demographics–wealth joint that such simulation depends on. The estimator is available as `populace-fit`, installable independently of the rest of the stack.

Keywords: microsimulation, imputation, survey weights, quantile regression forests, statistical matching

1 Introduction

Microsimulation models of tax and transfer policy require individual records that span domains no single survey covers (Bourguignon and Spadaro, 2006; Sutherland and Figari, 2013). The Current Population Survey observes employment and income but not wealth; the Survey of Consumer Finances observes wealth in detail but lacks the CPS's coverage and program detail; administrative tax records observe income components precisely but only for filers, and with narrow demographics. Model builders bridge these gaps with imputation: fit a conditional model of the missing variables on a donor survey, then draw values for the receiver's records (Little and Rubin, 2002; D'Orazio et al., 2006).

Two properties of this setting distinguish it from generic missing-data problems. First, both files are *weighted samples*: each record stands in for a different number of population units, and the quantity of interest is a population distribution, not a sample one. An imputation

model fit to unweighted records recovers the sample conditional, which differs from the population conditional exactly when the design (or a prior calibration) correlates weights with outcomes. Second, the imputed file is not an endpoint. It is *reweighted downstream* — calibrated to administrative margins (Deville and Särndal, 1992; Hainmueller, 2012), filtered to subpopulations, stressed under reforms — so an imputation that looks acceptable under the weights it shipped with can fail under weights it never saw. A rare donor record carrying an extreme value at near-zero weight is invisible in the shipped aggregates; a later reweighting that assigns it mass turns it into a distortion of the population total. We refer to such records as landmines, and we measure exposure to them directly with a reweighting stress test (Section 5.2, Table 3).

This paper presents and evaluates the imputation operator of POPULACE, POLICYENGINE's open-source microdata stack (PolicyEngine, 2026). The estimator combines three design choices, each motivated by one of the failure modes above: a *weighted bootstrap* that materializes survey weights into the training data of each forest, so the learned conditionals are population conditionals; *regime gates* that classify each record's sign class (negative, zero, positive) before magnitudes are modeled, so zero-inflated and sign-mixed variables — ubiquitous in economic microdata (Mullahy, 1986; Lambert, 1992) — retain their mass structure; and *sequential chaining* across imputed variables, so joint structure among targets survives the imputation (Raghunathan et al., 2003; Van Buuren and Groothuis-Oudshoorn, 2011). The interface enforces the first choice: fits are weighted by construction, and an unweighted fit must be requested explicitly.

We make three contributions:

1. **A weight-aware imputation estimator, available standalone.** We describe the estimator and its interface contract, and release it as `populace-fit`, installable and usable on plain data frames independently of the rest of the POPULACE stack.

2. **The population-view harness.** We formalize the evaluation question as one latent population observed through survey *views* — variable subsets, designs, measurement idioms — and score any candidate weighted file by projecting it through each view against that survey's holdout, using a strictly proper joint score (Gneiting and Raftery, 2007; Székely and Rizzo, 2013), a support-based coverage axis that is invariant to reweighting of the candidate (Naeem et al., 2020), a weighted classifier two-sample test, and an uncapped tail block on imputed columns — added after demonstrating that capped sample-geometry metrics are blind to twofold errors in the extreme tail,

where economic variables carry their policy weight. The harness is generator-agnostic and non-self-referential (holdouts appear nowhere upstream), and it cleanly separates what record generation must get right from what calibration can later change.

3. **A benchmark against standard practice with attribution, plus a reweighting stress test.** We compare the estimator against unweighted and weighted quantile regression forests (Meinshausen, 2006), ordinary least squares imputation (Von Hippel, 2007), linear quantile regression (Koenker and Bassett, 1978), and nearest-neighbour hot-deck statistical matching (Andridge and Little, 2010; D'Orazio et al., 2021), under a paired, repeated holdout protocol. Ablations knock out one design choice at a time, attributing performance to weighting, gating, and chaining separately, and a closed-form fragility diagnostic measures each method's worst-case single-record exposure under bounded reweightings of the imputed file — the landmine failure mode.

The policy stakes are concrete. Supplemental Security Income conditions eligibility on countable resources (Social Security Administration, 2024), so simulating its baseline caseload on the CPS — or reforms such as the SSI Savings Penalty Elimination Act, which would raise the resource limits (119th United States Congress, 2025) — is impossible without imputed wealth, and the composition of the simulated caseload is exactly a functional of the imputed demographics–wealth joint. The population-view experiment of Section 4.2 scores that joint directly: it asks whether households assembled from CPS demographics and imputed wealth are distributionally exchangeable with households the SCF actually observed.

The rest of the paper proceeds as follows. Section 2 situates the estimator in the survey-imputation literature. Section 3 specifies the estimator, the ablation design, the metrics, and the protocol. Section 4 describes the four benchmark tasks and their data. Section 5 reports results, Section 6 discusses limitations, and Section 7 concludes.

2 Background

2.1 Imputation methods in survey practice

Survey agencies and microsimulation teams draw on a small set of imputation families. *Hot-deck and statistical matching* methods donate observed values from a matched donor record; they preserve marginal distributions by construction because every imputed value is a real donor value (Andridge and Little, 2010; D'Orazio et al., 2006), and remain the

standard data-fusion tool in European microsimulation practice (Sutherland and Figari, 2013). *Regression-based* methods impute from a fitted conditional model — ordinary least squares with residual draws (Von Hippel, 2007), linear quantile regression (Koenker and Bassett, 1978; Koenker, 2005) — and extend to multiple imputation for uncertainty propagation (Rubin, 1987; Kennickell, 1998). *Tree-ensemble* methods, in particular quantile regression forests (Meinshausen, 2006; Breiman, 2001), estimate full conditional distributions non-parametrically and have become the default in POLICYENGINE's data pipelines (Ghenis, 2018; Woodruff and Ghenis, 2024) for their ability to capture nonlinear structure without per-variable specification. The `microimpute` package (PolicyEngine, 2025) implements these families under one interface and selects among them by cross-validation; its central finding — no family dominates across datasets — motivates empirical comparison on each new task rather than a prescribed method.

2.2 Chained imputation and joint structure

Imputing several variables one at a time, each conditional only on shared predictors, destroys the dependence among imputed variables; sequential (chained-equations) imputation conditions each variable on those already imputed, retaining joint structure (Raghunathan et al., 2003; Van Buuren and Groothuis-Oudshoorn, 2011). In data fusion the same issue appears as the conditional independence assumption of statistical matching: matched files preserve marginals but attenuate cross-source correlations (D'Orazio et al., 2006; Meinfelder, 2011). Synthetic-data generation faces the identical trade-off at whole-file scale (Rubin, 1993).

2.3 Weights in imputation

Design weights enter imputation practice unevenly. The model-based literature often treats weights as a nuisance — if the model is correctly specified and the design variables are conditioned on, weighting is unnecessary — while the design-based tradition weights every estimator (Little and Rubin, 2002). In production microdata pipelines the model is never correctly specified and the design variables are never fully available, so the choice is consequential: an unweighted fit targets the sample conditional, and any correlation between weights and outcomes within predictor cells propagates into the imputed population distribution. Tree ensembles add a mechanical subtlety: a fully-grown forest's *predictive distribution* is a leaf-membership distribution, which per-record `sample_weight` arguments do not reweight, so honoring weights requires materializing them into the training data — the weighted bootstrap of Section 3. Zero-inflated targets add a second subtlety: weighting

the zero/nonzero gate by resampling can delete a rare sign class outright, so the gate must be weighted directly. We are not aware of prior survey-imputation work that combines weight-aware forest training, sign-regime gating, and chaining in one estimator; documenting and testing that combination is the gap this paper fills.

2.4 Evaluating imputed files

Because an imputation is a draw from an estimated conditional distribution, pointwise accuracy metrics understate quality; the relevant question is distributional. We follow the quantile-loss tradition for conditional calibration (Koenker, 2005; Meinshausen, 2006) and use Wasserstein distance for marginal fit. For the joint, the evaluation perspective this paper formalizes — one latent population, surveys as partial views of it, candidates scored through each view against held-out survey samples — has ancestry in several literatures that have not, to our knowledge, been assembled into an operational benchmark: the survey-statistics treatment of multiple sources observing one population (Lohr and Raghunathan, 2017), the missing-data view of fusion as block missingness on a single file (Little and Rubin, 2002; Rässler, 2002), strictly proper scoring of distributional claims against realized samples (Gneiting and Raftery, 2007; Székely and Rizzo, 2013), and the precision–recall–density–coverage geometry of generative-model evaluation (Naeem et al., 2020). Population-synthesis practice validates against input marginals rather than held-out views, and reviews of the field note the absence of a standard benchmark. Section 3.3 operationalizes the combination; the reweighting fragility diagnostic of Section 3.4 is, to our knowledge, new.

3 Methodology

3.1 The estimator

`populace-fit` models $P(\text{targets} \mid \text{predictors})$ with a quantile-regression-forest estimator (Meinshausen, 2006; Zillow Group, 2024) wrapped in three mechanisms.

Weighted bootstrap. Random forests cannot honor a per-record `sample_weight` in their predictive distribution: a fully grown leaf holds individual training rows, and weighting the impurity criterion does not change which values a quantile query reads out. `populace-fit` therefore materializes weights into the data: before each forest is grown, training rows are resampled with replacement with probability proportional to weight, so leaf distributions reflect the weighted population. Draws from the fitted forest then sample the *weighted*

conditional.

Regime gates. A numeric target's sign support — which of $\{\text{negative}, \text{zero}, \text{positive}\}$ occur — defines its regime. A single regressor over a zero-inflated or sign-mixed target either drops a tail or interpolates across the empty gap at zero; hurdle-style two-part models are the classical remedy (Mullahy, 1986; Lambert, 1992). `populace-fit` fits a classifier that gates each row into a sign class and a separate magnitude forest within each nonzero sign. Regime detection is structural (it reads the unweighted support, since which signs *exist* is a fact about the variable, not the weighting), while the gate itself is weighted directly through its `sample_weight` — not by resampling, which would delete a rare sign class whose total weight share is negligible. Draws sample the gate's predicted class probabilities rather than taking the modal class, preserving the zero mass.

Sequential chaining. Targets are imputed in order, each conditioning on the predictors plus the targets already drawn (Raghunathan et al., 2003; Van Buuren and Groothuis-Oudshoorn, 2011), so dependence among imputed variables survives.

Interface contract. Fitting over a POPULACE frame reads the frame's typed survey weights by default; fitting over a plain data frame — the standalone path — requires the weights explicitly (a weight column name, a weight vector, or the literal `"none"`), and omitting them is an error rather than a silent unweighted fit. The contract makes the weight-blindness failure mode of Section 2.3 unrepresentable by accident.

3.2 Ablations and baselines

Each design choice is evaluated by knocking it out with everything else held fixed: the full estimator versus itself with `weights="none"` (the weighted bootstrap's contribution), versus fitting each target independently (chaining's contribution), and versus a single ungated forest at matched hyperparameters (gating's contribution). Baselines are standard practice: quantile-regression-forest, ordinary-least-squares, and linear quantile-regression imputers from `microimpute` (PolicyEngine, 2025), and nearest-neighbour-distance hot-deck statistical matching from `py-statmatch` (D'Orazio et al., 2021). The candidate and the plain-forest baselines share the same underlying forest implementation (Zillow Group, 2024), so their differences isolate the design choices rather than implementation quality.

3.3 The population-view harness

Per-variable metrics cannot adjudicate the question this paper cares about, because the methods differ most in how they treat the *joint*. We therefore evaluate within a framework we call the population-view harness. One latent population distribution produces individuals; each survey s is a *view* V_s of it — a variable subset, a sampling design, a measurement idiom — and the survey's holdout is a weighted sample from $V_s(P)$ (Lohr and Raghunathan, 2017). A candidate weighted file Q (any generator's) is scored by projecting it through each view and comparing $V_s(Q)$ to that view's holdout *in the view's own variable space*. Cross-survey consistency is never required — the candidate is only asked to explain each view in its own idiom — which keeps the harness usable even though surveys disagree with one another about the population they share.

Each view carries four complementary blocks. The *weighted energy distance* (Székely and Rizzo, 2013) is the sample form of a strictly proper scoring rule (Gneiting and Raftery, 2007): the true distribution uniquely minimizes its expectation, so a candidate hedged toward modal households scores strictly worse than one matching the full distribution. PRDC *coverage* (Naeem et al., 2020) — the weighted fraction of holdout points with a candidate point inside their k-nearest-neighbour radius — is the explicit anti-collapse axis; because it depends on the candidate only through its support, it is invariant to any reweighting of the candidate, making it the calibration-blind block of the scorecard. A *weighted classifier two-sample test* (cross-validated AUC of a gradient-boosted classifier at equal class mass; 0.5 is indistinguishable) is the omnibus check. The fourth block exists because we caught the first three missing a real failure: pairwise sample-geometry metrics run on capped subsamples, and a candidate whose imputed wealth 99th percentile is double the holdout's can tie on energy distance because the discrepant mass is a sliver of standardized pairwise space. Economic variables live in heavy right tails, so each *imputed* column also carries a *tail block*, computed on the full weighted samples with no cap: its weighted Wasserstein-1 distance to the holdout (scaled by the holdout's weighted standard deviation) and its weighted q90 and q99 ratios (candidate over holdout; 1 is a perfect tail). Holdouts are used nowhere upstream, so the harness is a non-self-referential test surface; a sampling-noise floor — the donor split itself scored as a candidate — anchors what "as good as another sample of the survey" means on each axis, including how much a tail ratio wobbles under sampling noise alone.

3.4 Marginal and stress metrics

All metrics are computed under survey weights, on held-out receiver rows. *Weighted pinball loss*, averaged over a decile grid, scores conditional distributional calibration (Koenker, 2005). *Weighted Wasserstein-1 distance* to the weighted donor distribution scores marginal fit. *Zero-share error* scores preservation of the exact-zero mass of zero-inflated targets. Finally, *reweighting fragility* formalizes the landmine diagnostic: over the family of bounded multiplicative reweightings $w_i \mapsto m_i w_i$ with $m_i \in [1/\kappa, \kappa]$ (the hard weight-ratio bounds production calibration guards use), the worst-case single-record share of a target aggregate has the closed form $\kappa^2 c^* / (\kappa^2 c^* + S - c^*)$, where c^* is the largest baseline contribution $w_i|a_i|$ and S their sum. We report it at $\kappa = 5$ for every method next to the held-out truth's own fragility — the exposure a method should not exceed.

3.5 Protocol

Every task uses an 80/20 donor/receiver holdout repeated over ten seeds, with splits paired across methods: the split is a pure function of the seed, so every method fits and is scored on identical partitions. Task inputs are frozen once and pinned by hash; only the method varies within a sweep. All results tables are generated from the sweep artifacts by the repository's command line; no result is hand-entered. Where a method is unavailable or a metric undefined for a task, the cell is reported as absent rather than omitted.

4 Data and tasks

Four task families span the settings the estimator is used in. Where a task imputes within one survey, the protocol of Section 3.5 applies: 80/20 donor/receiver holdouts repeated over ten paired seeds. The population-view experiment additionally projects candidates onto a second survey. Every table in Section 5 regenerates from the repository's committed run configurations; row caps and skipped cells are recorded in the run manifests.

4.1 Task 1: wealth within the SCF

The Survey of Consumer Finances (Board of Governors of the Federal Reserve System, 2023) observes household wealth in detail. From the 2022 summary extract (first implicate, survey weights) we impute `debt` — zero-inflated and nonnegative — and `networth` — sign-mixed, with a heavy right tail — from age, sex, education class, marital status, children,

total income, and wage income. The two targets are chosen deliberately: their sign structures exercise the regime gates, and their strong mutual dependence exercises chaining.

4.2 Task 2: the SCF→CPS population view

Each method fits wealth conditionals on an SCF donor split and imputes them onto real CPS ASEC households sharing the predictor set — households built from the Census Bureau's ASEC public-use files (U.S. Census Bureau, 2025), read directly from `census.gov` so that no task input flows through any PolicyEngine processing. The experiment runs in two profiles. The *minimal* profile — six shared predictors, the two wealth targets of Section 4.1, a single receiver vintage (ASEC 2025) — is the controlled instantiation in which every mechanism is legible. The *populace-scale* profile mirrors the production stack's shape: the shared predictor set widens to everything the two surveys genuinely share (ten variables, adding education class, race, homeownership, and labor-force status, each mapping verified against both files' codebooks), the receiver pools three ASEC vintages (2023–2025, weights scaled to one household population, 168,852 households) as the production support spine does, and the targets widen to a four-component chained balance sheet — financial assets, nonfinancial assets, debt, then net worth — whose accounting identity networth $=$ fin $+$ nfin $-$ debt holds exactly in the extract, so chaining's contribution is directly interpretable. The candidate population — CPS demographics plus imputed wealth, under CPS household weights — is then scored through the harness of Section 3.3 against the *held-out* SCF: the joint of shared predictors and wealth, under SCF weights. The `scf_sample_reference` rows score the donor split itself against the holdout, anchoring the sampling-noise floor. This is the paper's direct test of whether conditional-distribution methods preserve joint structure that matching's conditional-independence gluing loses (Rässler, 2002; D'Orazio et al., 2006), measured where it matters: on a receiver file the donor survey never saw. Because the demographic block of every candidate is the same real CPS sample, differences between methods isolate the wealth block and its coupling to demographics; the shared demographic offset between the two surveys is common to all methods and bounded below by the reference floor.

4.3 Task 3: zero-inflated income components in the CPS

From the Census Bureau's ASEC 2025 public-use person file (income reference year 2024), read directly from `census.gov` — never a processed or enhanced artifact, which would embed the production imputations this paper is about — we impute interest and dividend

income for adults from age, sex, and employment income, under the ASEC person supplement weights. (Survey files publish per-person design weights; POPULACE's production convention deliberately carries calibrated *household* weights onto persons, since its calibration operates at household grain. The task fixes the survey's own published weights, and because every method shares whatever weights the task fixes, the method comparison does not turn on this choice.) Both components are zero for most adults (41.7 and 85.2 percent, respectively) with heavy positive tails — the regime surface where gating and weighting matter most. Public-use ASEC amounts are topcoded, which truncates the extreme tail relative to an administrative or SCF-style source; the fragility comparisons of Section 5.2 are therefore conservative. This task also carries the reweighting stress test of Section 3.4: prior POLICYENGINE production experience motivates it, where an unweighted fit broadcast rare high-value donor records across a receiver file as near-zero-weight point masses that later reweighting amplified. All fragility numbers reported here are measured under this paper's protocol.

4.4 Task 4: cross-dataset transfer

To test behavior beyond economic microdata, we run the method surface over the six OpenML AutoML Benchmark regression datasets used in the `microimpute` manuscript's appendix (Vanschoren et al., 2014) — `space_ga`, `elevators`, `brazilian_houses`, `onlinenewspopularity`, `abalone`, and `house_sales` — treating each dataset's target as the imputed column under uniform weights.

Full multi-view synthesis — one candidate scored simultaneously against CPS, SIPP (U.S. Census Bureau, 2023), and PSID (Institute for Social Research, University of Michigan, 2023) views — is the production release-gate use of the harness and is left to future work; PSID access requires registration, and the two-view experiment of Section 4.2 already identifies the joint-structure differences the paper is about.

5 Results

Throughout, `plain_qrf` — the ungated, unchained forest — is the same estimator and configuration as the `microimpute` QRF baseline, so it appears once, as an ablation row; OLS, quantile regression, and hot deck are the remaining standard-practice baselines, and the weighted marginal draw is the unconditional lower bound.

Method	Debt pinball	Debt W_1	Debt zero-share err.	Net worth pinball	Net worth W_1
populace-fit	49,201 (2,637)	**13,001** (6,285)	0.020 (0.012)	551,317 (36,012)	**135,871** (51,106)
– unweighted	50,218 (2,452)	32,247 (16,930)	**0.012** (0.014)	703,457 (77,173)	829,158 (445,108)
– unchained	49,201 (2,637)	**13,001** (6,285)	0.020 (0.012)	556,622 (23,349)	158,606 (44,310)
– ungated/unchained forest	50,433 (3,120)	16,619 (5,358)	0.018 (0.018)	560,819 (36,903)	234,367 (142,057)
OLS	70,053 (17,637)	156,295 (74,822)	0.226 (0.018)	1,116,572 (383,007)	3,565,923 (2,062,063)
Quantile regression	49,797 (3,391)	29,460 (20,456)	0.151 (0.031)	534,420 (37,801)	295,470 (110,986)
NND hot deck (py-statmatch)	49,537 (2,627)	13,770 (5,141)	0.017 (0.013)	644,921 (51,255)	509,013 (399,757)
Weighted marginal draw	**48,277** (2,586)	14,617 (4,732)	0.020 (0.017)	**443,554** (20,651)	215,398 (87,264)

Table 1: Wealth imputation within the SCF (debt and net worth from demographics and income): weighted pinball loss, Wasserstein-1 to the weighted donor distribution, and zero-share error, mean (sd) over ten paired seeds. Bold marks the best non-reference value per column. Lower is better.

5.1 Wealth within the SCF

Table 1 reports the within-SCF task. The headline is the weighting ablation: switching the candidate to `weights="none"` degrades net-worth Wasserstein-1 by a factor of six (135,871 to 829,158) and debt by a factor of 2.5 (13,001 to 32,247). The SCF's list-sample design oversamples wealthy households at low weight, so an unweighted fit learns the sample's wealth distribution rather than the population's — this is the weight-blindness failure mode of Section 1 measured directly, and no other design choice comes close to its effect size. Gating and chaining contribute more modestly: the ungated, unchained forest sits at 234,367 on net-worth W_1 (1.7 times the candidate), and removing chaining raises net-worth W_1 by 17 percent (135,871 to 158,606).

Against the baselines, the candidate's margins are largest where tails and joints matter: hot-deck matching nearly ties it on debt marginal fit (13,770 vs. 13,001) but is 3.7 times worse on net worth (509,013), OLS is catastrophic on both (156,295 and 3.6 million — normal residuals cannot represent a wealth distribution), and linear quantile regression leaves 15 percent of the debt zero mass unaccounted for, against at most 2 percent for the forest-based methods.

One result reads wrong until the metric is examined: the *unconditional* weighted marginal draw posts the best pinball loss on both targets. With predictors as weak as demographics and income are for wealth, receiver-side marginal metrics cannot distinguish a method that draws from the correct weighted marginal while ignoring predictors entirely from one that also gets the conditionals right. This is not an artifact to explain away; it is the motivating observation for the population-view harness of Section 5.5, which scores the joint.

Method	Interest pinball	Interest zero-share	Dividend pinball	Dividend zero-share
populace-fit	1,491 (63.58)	0.013 (0.011)	320.68 (45.26)	0.006 (0.004)
– unweighted	**1,491** (63.62)	0.011 (0.007)	320.68 (45.29)	0.008 (0.005)
– unchained	1,491 (63.58)	0.013 (0.011)	320.69 (45.26)	**0.006** (0.005)
– ungated/unchained forest	1,493 (63.34)	0.014 (0.011)	322.17 (45.24)	0.044 (0.010)
OLS	2,527 (63.57)	0.422 (0.012)	961.52 (112.29)	0.844 (0.005)
Quantile regression	1,493 (63.23)	0.270 (0.016)	320.91 (45.19)	0.251 (0.011)
NND hot deck (py-statmatch)	1,493 (63.89)	0.021 (0.019)	320.94 (45.22)	0.009 (0.009)
Weighted marginal draw	1,491 (63.61)	**0.011** (0.013)	**320.68** (45.25)	0.006 (0.004)

Table 2: Zero-inflated income components in the raw CPS (interest and dividend income for adults): weighted pinball loss and zero-share error, mean (sd) over ten paired seeds.

Method	Interest fragility	(truth)	Dividend fragility	(truth)
populace-fit	0.384 (0.051)	0.372 (0.043)	0.675 (0.100)	0.614 (0.104)
– unweighted	0.395 (0.043)	0.372 (0.043)	0.699 (0.057)	0.614 (0.104)
– unchained	0.384 (0.051)	0.372 (0.043)	0.661 (0.100)	0.614 (0.104)
– ungated/unchained forest	0.364 (0.064)	0.372 (0.043)	**0.633** (0.087)	0.614 (0.104)
OLS	0.055 (0.010)	0.372 (0.043)	0.055 (0.009)	0.614 (0.104)
Quantile regression	**0.368** (0.105)	0.372 (0.043)	0.444 (0.073)	0.614 (0.104)
NND hot deck (py-statmatch)	0.391 (0.027)	0.372 (0.043)	0.647 (0.140)	0.614 (0.104)
Weighted marginal draw	0.356 (0.042)	0.372 (0.043)	0.746 (0.067)	0.614 (0.104)

Table 3: Reweighting fragility at $\kappa = 5$: worst-case single-record share of each component's aggregate over bounded multiplicative reweightings, next to the held-out truth's own fragility. The target is the truth's exposure — bold marks the closest method; substantially lower means the method fails to generate realistic tails, substantially higher means it plants landmines.

5.2 Zero-inflated components and reweighting fragility

Table 2 reports the CPS components task. Pinball loss barely separates the non-OLS methods — age, sex, and earnings carry little conditional signal for capital income — but the zero mass separates them sharply. The candidate holds zero-share error to 0.6 and 1.3 percentage points (dividends, interest); the ungated forest drifts to 4.4 points on dividends; linear quantile regression misses by 25–27 points because a 99-level linear quantile grid cannot hold a five-sixths zero mass; and OLS misses the dividend zero share by 84 points — with normal residual draws, nearly every adult becomes a dividend recipient. Hot deck, which donates observed values, holds the zero mass within 0.9–2.1 points, as expected.

Table 3 reports the landmine diagnostic. Two readings matter. First, the held-out truth is itself concentrated: under $\kappa = 5$ reweighting a single record can carry 37 percent of the interest aggregate and 61 percent of the dividend aggregate at this sample size — heavy-tailed capital income is genuinely fragile, and a faithful method should *match* that exposure, not minimize it. Second, every forest- and donation-based method tracks the truth's fragility closely, while

Task	Target	Metric	Δ unweighted	Δ unchained	Δ plain forest
cps components	dividend_income	pinball_loss	0.000	0.005	1.48
cps components	dividend_income	wasserstein1	36.11	11.29	308.65
cps components	dividend_income	zero_share_error	0.002	-0.000	0.038
cps components	interest_income	pinball_loss	-0.016	0.000	1.57
cps components	interest_income	wasserstein1	-26.28	0.000	116.69
cps components	interest_income	zero_share_error	-0.002	0.000	0.001
scf wealth	debt	pinball_loss	1,017	0.000	1,232
scf wealth	debt	wasserstein1	19,246	0.000	3,618
scf wealth	debt	zero_share_error	-0.008	0.000	-0.001
scf wealth	networth	pinball_loss	152,140	5,305	9,502
scf wealth	networth	wasserstein1	693,287	22,735	98,496
scf wealth	networth	zero_share_error	-0.000	0.000	0.001

Table 4: Paired-seed deltas, ablation minus candidate (positive = ablation worse on these lower-is-better metrics), mean over ten seeds.

OLS sits far below it (0.06 against 0.37 and 0.61 — understating the true exposure by factors of roughly 7 and 11): its aggregate is spread across many moderate records because the method never generates realistic tails — the same failure that its zero-share and Wasserstein numbers show from other angles. No method exceeds the truth's exposure materially on this within-survey task; the landmine pathology of Section 1 arises when unweighted cross-file fits meet downstream reweighting, and the direct evidence for it here is the tail block of the population-view experiment (Section 5.5).

5.3 Ablations: attributing the gains

Table 4 attributes the candidate's behavior to its design choices with paired-seed deltas. The ordering is task-dependent in an informative way. On the SCF — informative weights, heavy tails — weighting dominates ($+693{,}287$ net-worth W_1 when removed), the structural forest machinery matters next ($+98{,}496$), and chaining contributes a consistent but smaller $+22{,}735$. On the CPS components — weakly informative weights, extreme zero inflation — weighting is approximately free ($\Delta \approx 0$), while gating carries the load ($+534$ dividend W_1 for the plain forest, and the zero-share drift of Table 2). The design choices are complements across regimes, not redundant safeguards: which one binds depends on the survey's design and the target's structure, and none of them hurts where it does not help.

5.4 Cross-dataset transfer

Table 5 runs the surface over the six OpenML regression datasets under uniform weights. Two observations. First, with weights uninformative by construction, the candidate and its

Method	abalone	brazilian_houses	elevators	house_sales	onlinenewspopularity	space_ga	Mean rank
populace-fit	0.254	166.64	0.000	7,080	274.67	0.008	3.83
– unweighted	0.257	184.92	0.000	6,744	233.10	0.008	**3.33**
– unchained	0.254	166.64	0.000	7,080	274.67	0.008	3.83
– ungated/unchained forest	0.174	191.36	0.000	8,461	546.16	0.012	5.00
OLS	0.532	212.09	0.002	90,518	6,117	0.015	8.83
Quantile regression	0.321	212.04	0.001	24,115	382.98	0.085	7.83
NND hot deck (py-statmatch)	0.176	199.12	0.000	10,212	161.43	0.010	4.00
Weighted marginal draw	0.148	210.50	0.000	7,135	143.52	0.011	**3.33**

Table 5: Wasserstein-1 to the donor distribution across the six OpenML regression datasets (uniform weights), mean over ten paired seeds, with mean rank across datasets.

Method	Energy distance	Coverage	C2ST AUC	NW W_1/sd	NW q99 ratio
SCF sample (floor)	0.009 (0.006)	0.987 (0.009)	0.496 (0.023)	0.013 (0.003)	1.10 (0.202)
populace-fit	**0.139** (0.026)	0.960 (0.014)	0.732 (0.013)	0.031 (0.007)	0.696 (0.112)
– unweighted	0.140 (0.026)	0.958 (0.014)	**0.731** (0.016)	0.159 (0.041)	1.93 (0.255)
– unchained	0.139 (0.026)	0.958 (0.016)	0.742 (0.013)	0.029 (0.009)	0.765 (0.177)
– ungated/unchained forest	0.140 (0.027)	0.957 (0.012)	0.775 (0.008)	0.029 (0.007)	0.805 (0.156)
OLS	0.166 (0.045)	0.732 (0.087)	0.988 (0.005)	0.434 (0.282)	**0.967** (0.505)
Quantile regression	0.141 (0.027)	0.938 (0.028)	0.927 (0.018)	0.052 (0.013)	0.530 (0.118)
NND hot deck (py-statmatch)	0.141 (0.027)	**0.962** (0.013)	0.753 (0.014)	0.051 (0.016)	0.762 (0.148)
Weighted marginal draw	0.139 (0.026)	0.937 (0.005)	0.848 (0.008)	**0.014** (0.004)	1.09 (0.222)

Table 6: The population-view experiment, *minimal* profile (six shared predictors, two targets, ASEC 2025 receiver): candidates assembled from real CPS demographics plus each method's imputed wealth, under CPS household weights, scored against the held-out SCF through the SCF view. The first row is the sampling-noise floor. Energy distance lower is better; coverage higher; C2ST AUC of 0.5 is indistinguishable; net-worth W_1/sd lower is better; a q99 ratio of 1 is a perfect extreme tail.

unweighted ablation are statistically the same method, and they rank near the top (mean ranks 3.8 and 3.3) alongside the unconditional marginal draw (3.3); the linear methods rank last (7.8 and 8.8). Second — and more important for this paper's argument — the best mean rank on a *marginal* distance is achieved by a method with no conditional structure at all. Marginal metrics saturate: they reward drawing from the right marginal and cannot rank what imputation is actually for. The population-view harness exists because of exactly this ceiling.

5.5 The SCF→CPS population view

Tables 6 and 7 report the population-view experiment in its two profiles. The floor behaves as designed in both: another sample of the same survey is statistically indistinguishable (energy 0.009 and 0.012, coverage 0.987 and 0.986, AUC 0.496 and 0.499), and its q99 ratio of 1.10 calibrates how much a 99th-percentile ratio moves under sampling noise alone. Four findings follow.

Method	Energy distance	Coverage	C2ST AUC	NW W_1/sd	NW q99 ratio
SCF sample (floor)	0.012 (0.005)	0.986 (0.003)	0.499 (0.021)	0.013 (0.003)	1.10 (0.202)
populace-fit	**0.118** (0.024)	0.969 (0.009)	0.761 (0.026)	0.027 (0.008)	0.767 (0.190)
– unweighted	0.120 (0.024)	**0.971** (0.008)	**0.754** (0.011)	0.133 (0.038)	2.25 (0.443)
– unchained	0.118 (0.024)	0.967 (0.011)	0.819 (0.015)	0.028 (0.006)	0.699 (0.112)
– ungated/unchained forest	0.126 (0.025)	0.965 (0.011)	0.817 (0.009)	0.033 (0.022)	1.15 (0.445)
OLS	0.152 (0.059)	0.877 (0.055)	0.995 (0.002)	0.435 (0.282)	**0.963** (0.503)
Quantile regression	0.120 (0.026)	0.951 (0.011)	0.970 (0.007)	0.051 (0.013)	0.485 (0.098)
NND hot deck (py-statmatch)	0.119 (0.025)	0.971 (0.010)	0.768 (0.009)	0.024 (0.008)	0.747 (0.107)
Weighted marginal draw	0.119 (0.024)	0.931 (0.015)	0.968 (0.003)	**0.015** (0.003)	1.12 (0.241)

Table 7: The population-view experiment, *populace-scale* profile (ten shared predictors, four chain-ordered balance-sheet targets, pooled ASEC 2023–2025 receiver). Columns as in Table 6.

First, *the weighting failure is a tail failure, and only the tail block sees it*. The unweighted ablation ties the candidate on energy distance, coverage, and the C2ST in both profiles — and its imputed net-worth q99 is 1.93 times the holdout's in the minimal profile and 2.25 times at populace scale, with tail W_1/sd five times the candidate's (0.159 vs. 0.031 minimal; 0.133 vs. 0.027 at scale). Fitting the SCF unweighted learns the wealthy oversample's conditionals, and the resulting inflation concentrates beyond the resolution of capped pairwise geometry: we report this as a finding about evaluation practice. Any harness whose joint metrics subsample — which at these sample sizes they must — needs an uncapped tail block for heavy-tailed economic variables, or it will certify files whose top percentile is wrong by a factor of two.

Second, *chaining is a joint property, and it binds at scale*. In the minimal profile (two targets) the unchained ablation is indistinguishable from the candidate. At populace scale, where the four targets satisfy the accounting identity networth = fin + nfin − debt exactly in the donor, fitting them independently costs 5.8 points of C2ST AUC (0.819 vs. 0.761) while leaving every marginal and tail statistic essentially unchanged — the classifier detects balance sheets that do not add up, which no marginal metric can. The same pattern holds for gating: the ungated forest loses 5.6 AUC points at scale (0.817).

Third, *conditioning richness makes the omnibus test decisive*. The unconditional weighted marginal draw — which ties or beats every method on marginal metrics throughout this paper — moves from AUC 0.848 in the minimal profile to 0.968 at populace scale, while the candidate stays closest to the floor (0.732 and 0.761). With ten shared predictors and a four-component wealth block, a classifier finds the missing demographics–wealth coupling almost perfectly. This is the empirical case for carrying strong auxiliary predictors in a production imputation stack: richer conditioning is what turns joint-structure failures from

undetectable into obvious.

Fourth, *hot-deck matching is a genuine near-peer on this task*. It sits within noise of the candidate on energy and coverage in both profiles, within 2.1 and 0.7 AUC points, and its donated values slightly beat the candidate's on tail W_1 at scale (0.024 vs. 0.027) — donation preserves marginals and tails by construction. The candidate's advantages over matching live where this task does not press hard: informative-weight regimes (Table 1), extreme zero inflation (Table 2), and multi-target coherence (the chaining result above, which matching handles only through whole-record donation at the cost of reusing donors). One more tail observation cuts against every conditional method, the candidate included: forest draws do not extrapolate beyond training support, so their imputed q99 ratios sit at 0.70–0.81 — they understate the extreme tail by 20–30 percent where donation-based draws sit near 1. No method in this surface gets the extreme tail right for free; the tail block is what keeps that visible.

6 Discussion

The results attribute the estimator's behavior to its design choices with a regularity worth stating plainly. *Weighting is the dominant choice wherever the design is informative*: on the SCF, whose list sample oversamples wealthy households, removing the weighted bootstrap costs a factor of six on net-worth marginal fit within the donor survey and inflates the imputed 99th percentile on the population view by a factor of 1.9 to 2.3 — larger than the gap between the candidate and any competing method. On the CPS components task, whose person weights carry little outcome information, weighting is approximately free. An estimator cannot know in advance which regime it is in, which is the argument for weighting by construction rather than by option. *Gating carries the zero-inflation surface* (0.6 versus 4.4 percentage points of dividend zero-share error against the same forest ungated, and 5.6 AUC points on the populace-scale population view). *Chaining is a joint property that binds at scale*: invisible with two targets, it is worth 5.8 AUC points when the four imputed components must satisfy a balance-sheet identity — and only the omnibus classifier sees the violation, because a balance sheet that does not add up has perfectly reasonable marginals. None of the three choices hurts where it does not help.

The evaluation findings are as load-bearing as the method findings, and one of them corrected this paper's own harness. Marginal metrics saturate: with weak predictors, an unconditional weighted draw from the donor marginal ties or beats every conditional method on pinball

loss and Wasserstein distance — on the OpenML suite it attains the best mean rank — while carrying no conditional structure at all; at populace scale the classifier separates it almost perfectly (AUC 0.968 against 0.761). But the joint metrics have their own blind spot: capped pairwise geometry tied the unweighted ablation with the candidate while its imputed net-worth 99th percentile was double the holdout's, because the discrepant mass is a sliver of standardized pairwise space. We caught this by comparing weighted quantiles directly, added the uncapped tail block to the harness, and report it as a general lesson: sample-geometry evaluations of heavy-tailed economic microdata need an explicit tail axis, with a sampling floor to calibrate its noise (the floor's own q99 ratio wobbles to 1.10). The fragility diagnostic adds the complementary caution — the target is the truth's own tail exposure, not minimal exposure, and the one method that minimizes it (OLS, at factors of 7 to 11 below truth) does so by failing to generate realistic tails at all.

Several limitations bound the claims. First, the population-view experiment runs one view pair: the common CPS–SCF offset (differing income concepts, ASEC topcoding, and income reference years spanning 2021 to 2024 across the pooled vintages) is shared by all methods but bounds how close any candidate can come to the floor; a multi-view instantiation (adding SIPP and PSID) would tighten the identification and is the production release-gate use of the harness. Second, the joint-geometry blocks still run on weight-proportional subsamples (2,048 points per side); the tail block and the C2ST, which use all rows, are the axes least affected. Third, no conditional method gets the extreme tail right: forest draws do not extrapolate beyond training support, so imputed q99 ratios sit at 0.70–0.81 for the candidate and its forest relatives, where donation-based draws sit near 1 — a genuine advantage of hot-deck matching that the tail block keeps visible, and a caution against reading the candidate's wins as uniform. Baseline quantile models are additionally converted to samplers through a 99-level grid, truncating draws beyond the 1st and 99th conditional percentiles. Fourth, the `microimpute` forest's internal randomness is not seedable through its public constructor, so its paired-seed dispersion includes forest noise the candidate's does not. Fifth, comparisons hold shared forest backends and default hyperparameters fixed rather than tuning each method per task; the ablations are exact controls, the cross-family comparisons are defaults-versus-defaults. Finally, we evaluate single imputations distributionally and do not propagate imputation uncertainty in the multiple-imputation sense (Rubin, 1987); for the population statistics microsimulation consumes, the distributional view is the relevant one, but inference from imputed files would require the fuller treatment.

7 Conclusion

Imputation sits upstream of every population statistic a microsimulation model produces, and the files it produces are reweighted by consumers the imputer never sees. We presented an estimator that treats survey weights as a structural obligation rather than an option — weighted bootstrap into the forests, directly weighted regime gates, sequential chaining, and an interface that refuses a silent unweighted fit — and a population-view harness that evaluates any candidate population against held-out survey views with a strictly proper joint score, a reweighting-invariant coverage axis, a classifier two-sample test, and an uncapped tail block. Empirically, the interface rule is the paper's largest single effect: on a survey whose design is informative, the identical estimator fit without weights is six times worse on net-worth marginal fit, and on the population view it inflates the imputed 99th percentile by a factor of 1.9 to 2.3 — a failure that capped joint geometry certifies and only the tail block catches, which is itself the paper's central lesson about evaluating imputed files. Chaining binds where targets are jointly constrained (5.8 AUC points on a balance-sheet identity at populace scale), gating carries zero-inflated components that plain forests and linear quantile models miss by 4 to 27 percentage points of zero mass, and marginal metrics — the field's default — rank an unconditional marginal draw at or near the top throughout while the harness separates it decisively. The estimator is available standalone as `populace-fit`; the harness, task configurations, and every number in this paper regenerate from the accompanying repository.

Conflict of interest

All authors are affiliated with POLICYENGINE, the nonprofit organization that develops POPULACE, `populace-fit`, and `microimpute`, the software evaluated in this paper. The authors have no other competing interests.

Funding

[TODO: Funding statement.]

Data and code availability

The paper and experiment code are at https://github.com/PolicyEngine/imputation-paper; every results table regenerates from committed run configurations via the repository's command line. The estimator is implemented in `populace-fit`, part of POPULACE, open source at https://github.com/PolicyEngine/populace. Baseline implementations are `microimpute` (https://github.com/PolicyEngine/microimpute) and `py-statmatch` (https://github.com/CosilicoAI/py-statmatch). The reported sweeps ran against `populace-fit` at populace commit `71b5a83`, `microimpute` 1.1.2, and `py-statmatch` at commit `094a242`; the SCF 2022 summary extract from the Federal Reserve (SHA-256 beginning `3bb4d890`) and the Census ASEC 2025 public-use bundle (`asecpub25csv.zip`, SHA-256 beginning `318845a2`) are the pinned survey inputs — both read directly from their agencies of origin — and each run directory's manifest records rows, caps, seeds, and methods.

References

119th United States Congress. SSI savings penalty elimination act, 2025. URL https://www.congress.gov/bill/119th-congress/senate-bill/1234. S. 1234 / H.R. 2540, introduced April 1, 2025.

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.

Board of Governors of the Federal Reserve System. Changes in U.S. family finances from 2019 to 2022: Evidence from the Survey of Consumer Finances. Technical Report 4, Federal Reserve Bulletin, October 2023.

François Bourguignon and Amedeo Spadaro. Microsimulation as a tool for evaluating redistribution policies. *Journal of Economic Inequality*, 4(1):77–106, 2006.

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

Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

Marcello D'Orazio, Marco Di Zio, and Mauro Scanu. *Statistical Matching: Theory and Practice*. John Wiley & Sons, 2006.

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.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Jens Hainmueller. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Institute for Social Research, University of Michigan. Panel Study of Income Dynamics: Public use dataset. Ann Arbor, MI: Survey Research Center, 2023. URL https://psidonline.isr.umich.edu/.

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.

Roger Koenker. *Quantile regression*. Cambridge University Press, 2005.

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

Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2nd edition, 2002.

Sharon L. Lohr and Trivellore E. Raghunathan. Combining survey data with other data sources. *Statistical Science*, 32(2):293–312, 2017.

Florian Meinfelder. On the simulation of contingency tables with given margins and statistical matching when observed information is incomplete. *Computational Statistics & Data Analysis*, 55(12):3257–3267, 2011.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

John Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020.

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

PolicyEngine. Populace: A microsimulation dataset construction pipeline, 2026. URL https://github.com/PolicyEngine/populace. Software.

Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16, 2003.

Susanne Rässler. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer, New York, 2002.

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

Donald B Rubin. Statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

Social Security Administration. SSI annual statistical report, 2024. Technical report, Social Security Administration, Office of Retirement and Disability Policy, 2024. URL https://www.ssa.gov/policy/docs/statcomps/ssi_asr/.

Holly Sutherland and Francesco Figari. EUROMOD: The European Union tax-benefit microsimulation model. *International Journal of Microsimulation*, 6(1):4–26, 2013.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

U.S. Census Bureau. Survey of Income and Program Participation (SIPP): 2022 panel. Washington, DC: U.S. Census Bureau, 2023. URL https://www.census.gov/programs-surveys/sipp.html.

U.S. Census Bureau. Current population survey, 2025 annual social and economic (ASEC) supplement. Technical report, U.S. Census Bureau, 2025. URL https://www2.census.gov/programs-surveys/cps/datasets/2025/march/asecpub25csv.zip.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luís Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

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.

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

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