Model-Based or Hot Deck? Imputing Item Non-Response Data in the Survey of Income and Program Participation (SIPP)

Besufekad Alemu, Angelica Phillips,
Ruth E. Sarafin, Joey Marshall, and Sandy Dietrich
U.S. Census Bureau

Social, Economic, and Housing Statistics Division

American Association for Public Opinion Research Conference
May 14, 2026

United States® Census Bureau

Disclaimers

This presentation is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed are those of the authors and not those of the U.S. Census Bureau.

Data Management System (DMS) number: P-7512692, Disclosure Review Board (DRB) approval numbers: CBDRB-FY25-SEHSD003-005, CBDRB-FY25-SEHSD003-108, CBDRB-FY26-SEHSD003-017.

The estimates are unweighted and therefore cannot be generalized to the population. Comparative statements in this presentation are not supported by statistical testing, and the results presented are only applicable to the described cohort of SIPP survey respondents.

United States® Census Bureau

Background

- Hot deck is the default imputation methods for most variables in the Survey of Income and Program Participation (SIPP).

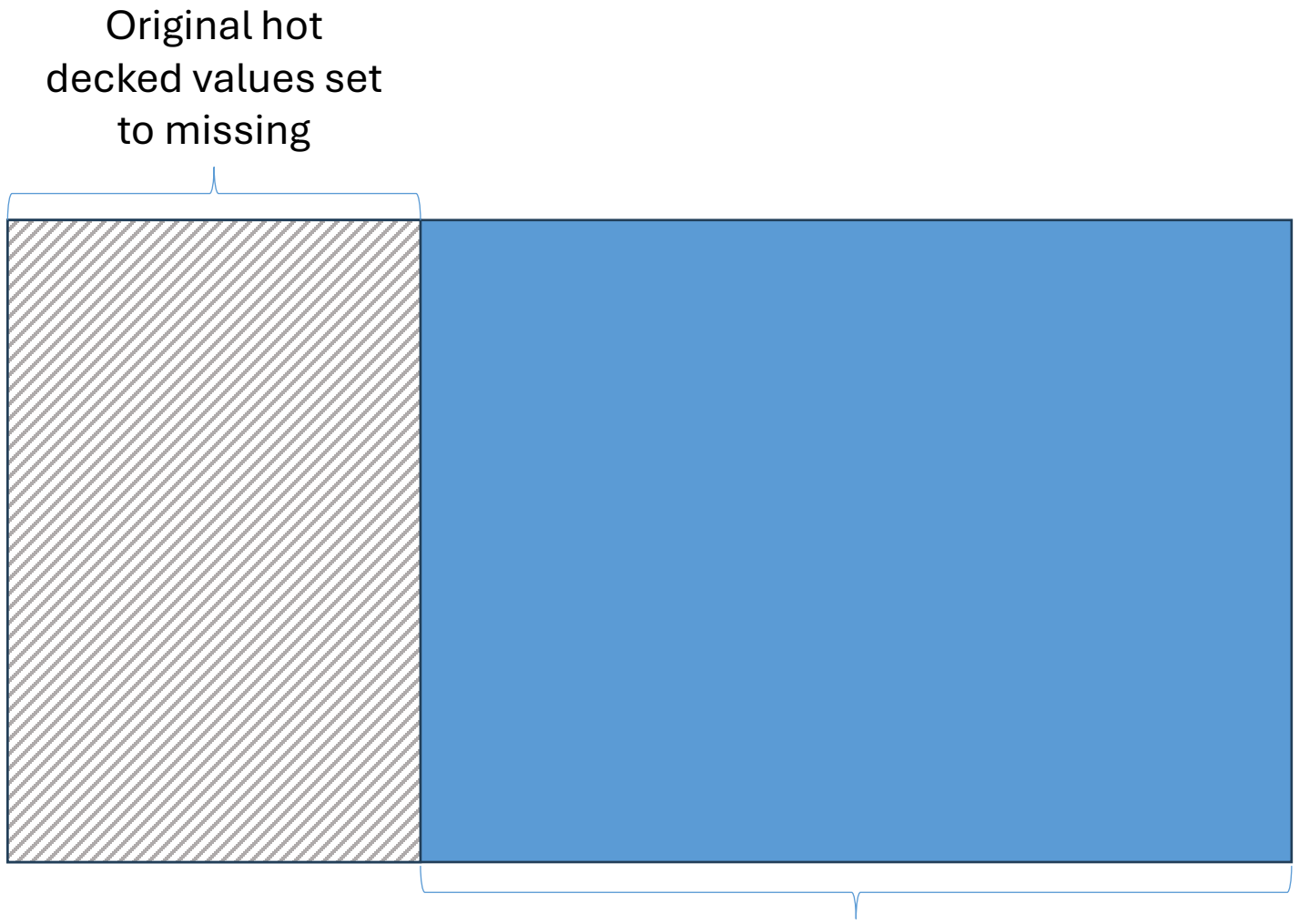
- Limitations of hot deck.
  - Relies on categorical or discretized matching variables.
  - Quality of imputes depend on how donor cells are defined.
  - Limited for multivariate relationships.

- Need to modernize imputation method.

- Goal: Using SIPP 2022, compare modern imputation methods against hot deck.

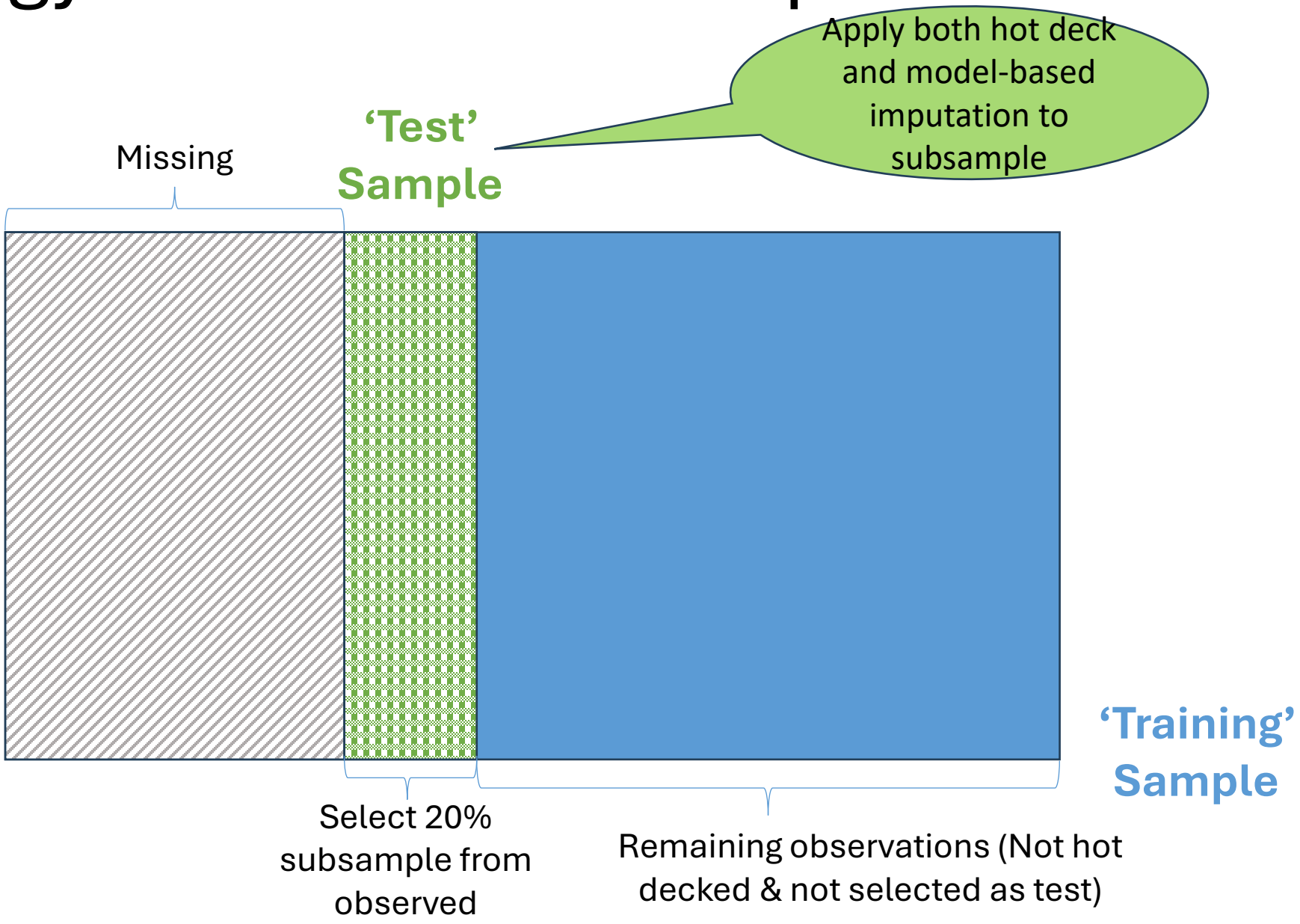
United States® Census Bureau

Methodology: Starting Scenario



United States® Census Bureau

Methodology: Random Subsample from Observed



Methodology: Hot Deck vs. Modern Methods

- Models requiring fully observed predictors.
  - Linear regression.
  - Logistic regression.

- Models that account for predictors with missing values.
  - Multiple Imputation by Chained Equations (MICE).
    - Uses linear, logit, and mlogit models to generate imputed values.
  - eXtreme Gradient Boosting (XGBoost).
    - A tree-based ensemble method.

United States® Census Bureau

Focal Variables

Variable	Type	In-Universe Observations	Missing Rate
Multiple partner fertility	Binary	21,600	**5.5%**
Class of worker	Multi-categorical	20,470	**25.9%**
SNAP benefit amount	Continuous	2,700	**12.4%**
Primary residence	Mixed (2 binary, 2 continuous, 1 multi-categorical)	9,300	**7.3%**

United States® Census Bureau

Focal Variables and Models

	Hot Deck	Linear Regression	Logistic Regression	MICE	XGBoost
Multiple partner fertility	✓		✓	✓	✓
Class of worker	✓		✓	✓	✓
SNAP benefit amount	✓	✓		✓	✓
Primary residence	✓			✓	

United States® Census Bureau

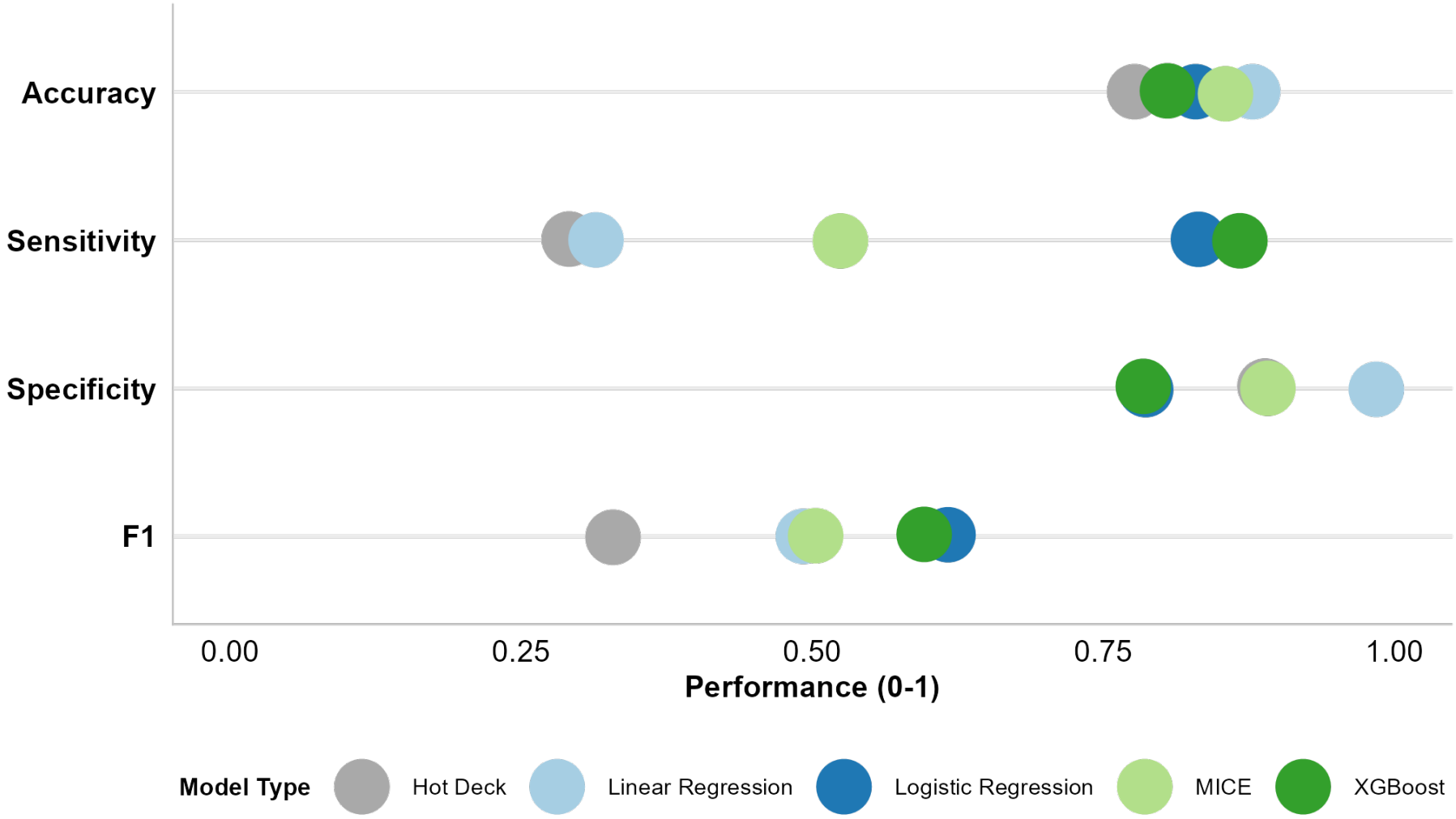
Results: Multiple Partner Fertility

$$\frac{Correct\ predictions}{All\ predictions}$$

$$\frac{Correct\ positive\ predictions}{All\ positive\ cases}$$

$$\frac{Correct\ negative\ predictions}{All\ negative\ cases}$$

$$2^* \frac{Precision * sensitivity}{Precision + sensitivity}$$



United States® Census Bureau

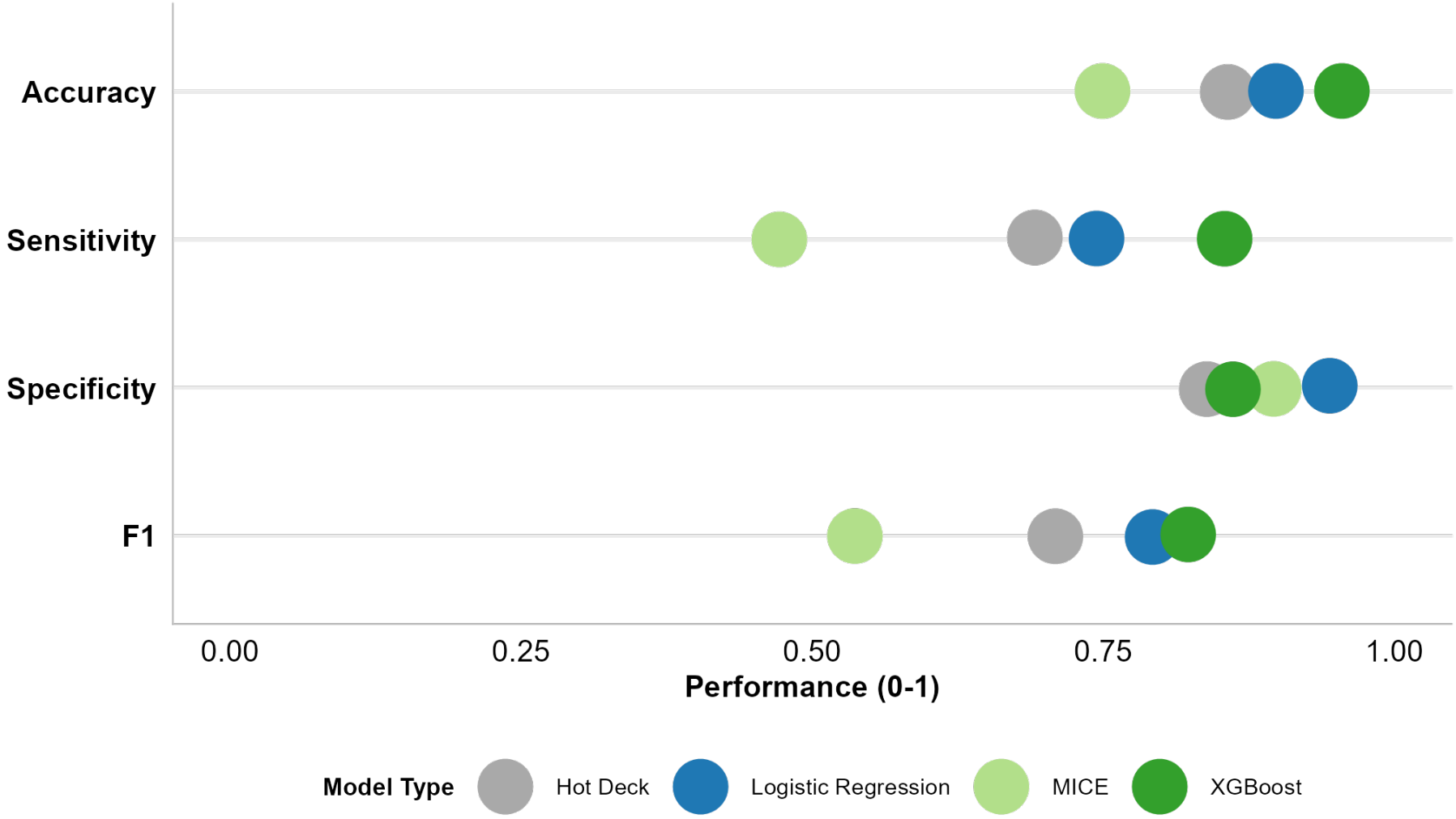
Results: Class of Worker

$$\frac{Correct\ predictions}{All\ predictions}$$

$$\frac{Correct\ positive\ predictions}{All\ positive\ cases}$$

$$\frac{Correct\ negative\ predictions}{All\ negative\ cases}$$

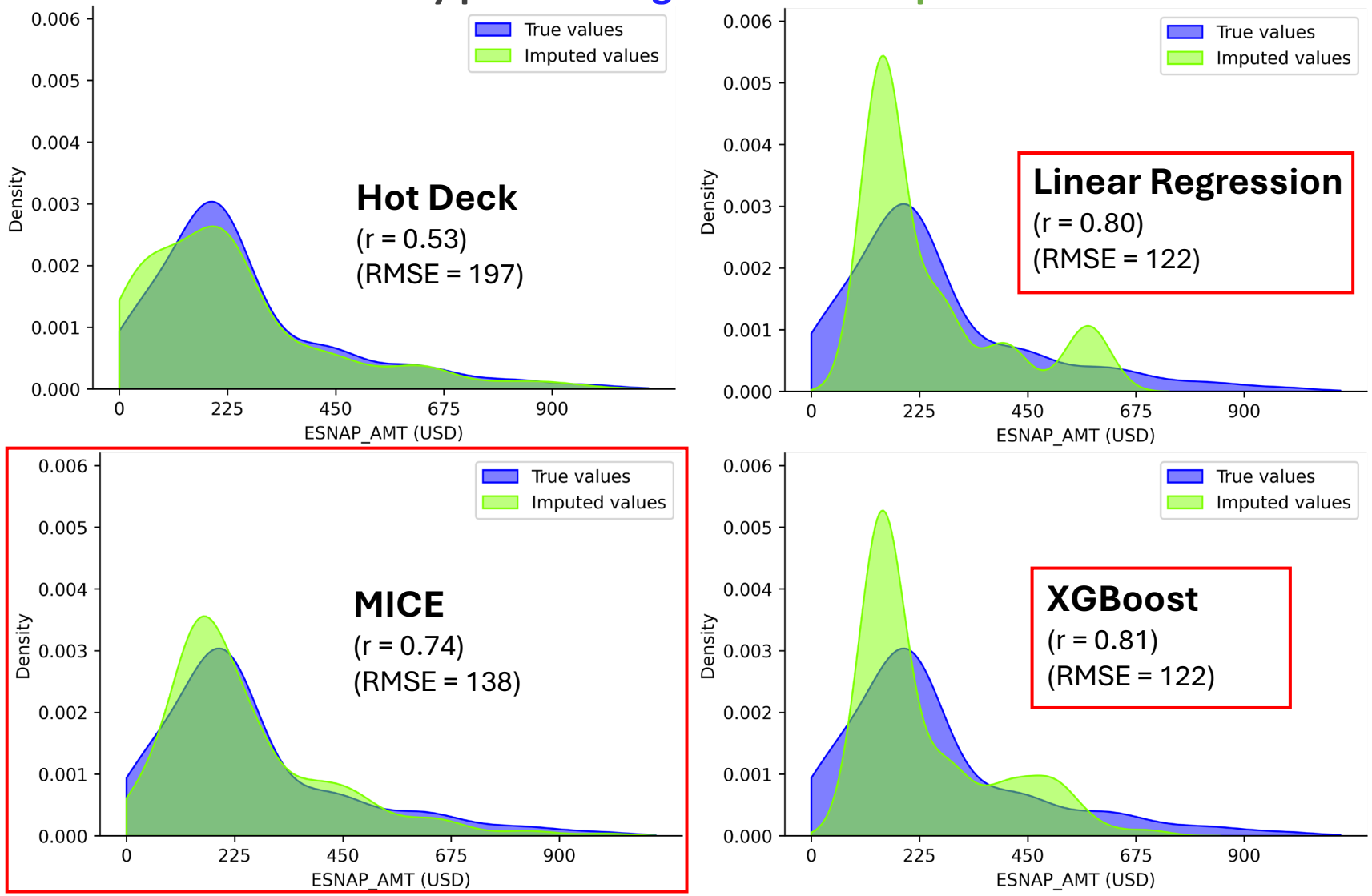
$$2*\frac{Precision * sensitivity}{Precision + sensitivity}$$



United States® Census Bureau

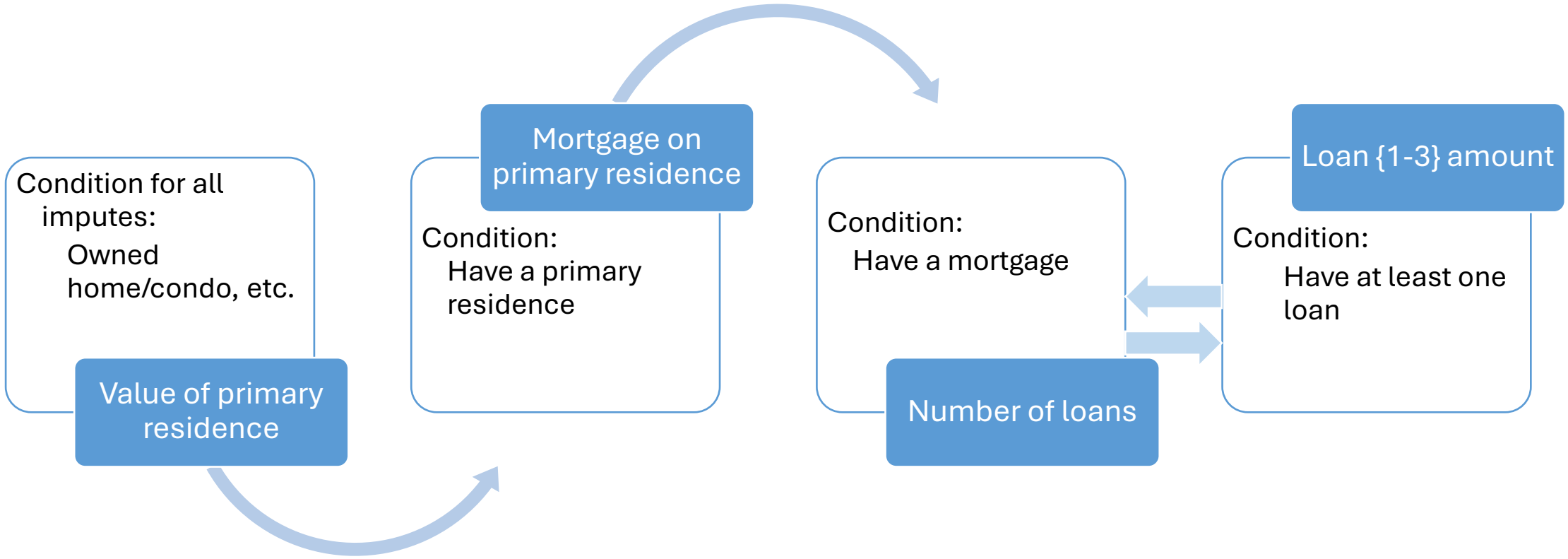
Results: SNAP Benefit Amount

Kernel density plots of original versus imputed values



United States® Census Bureau

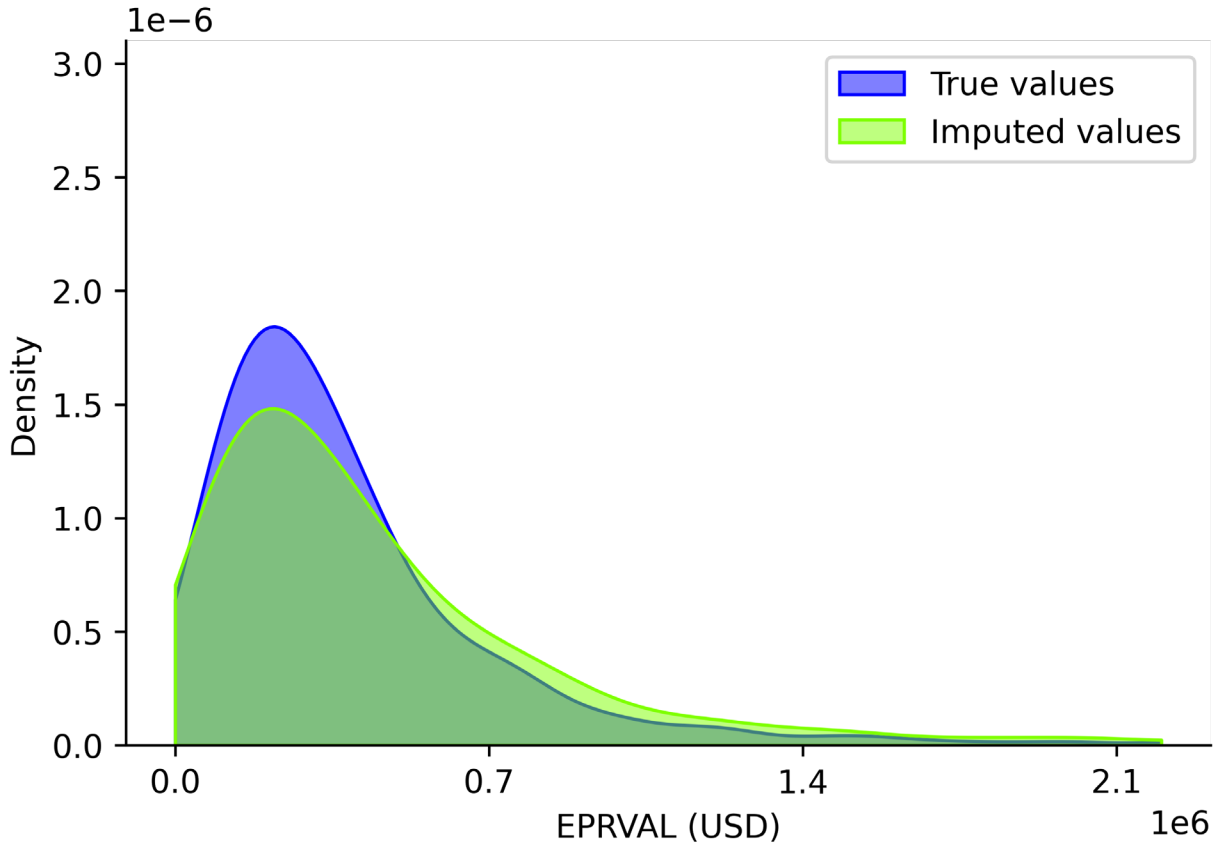
Primary Residence Hot Deck



United States® Census Bureau

Results: Value of Primary Residence

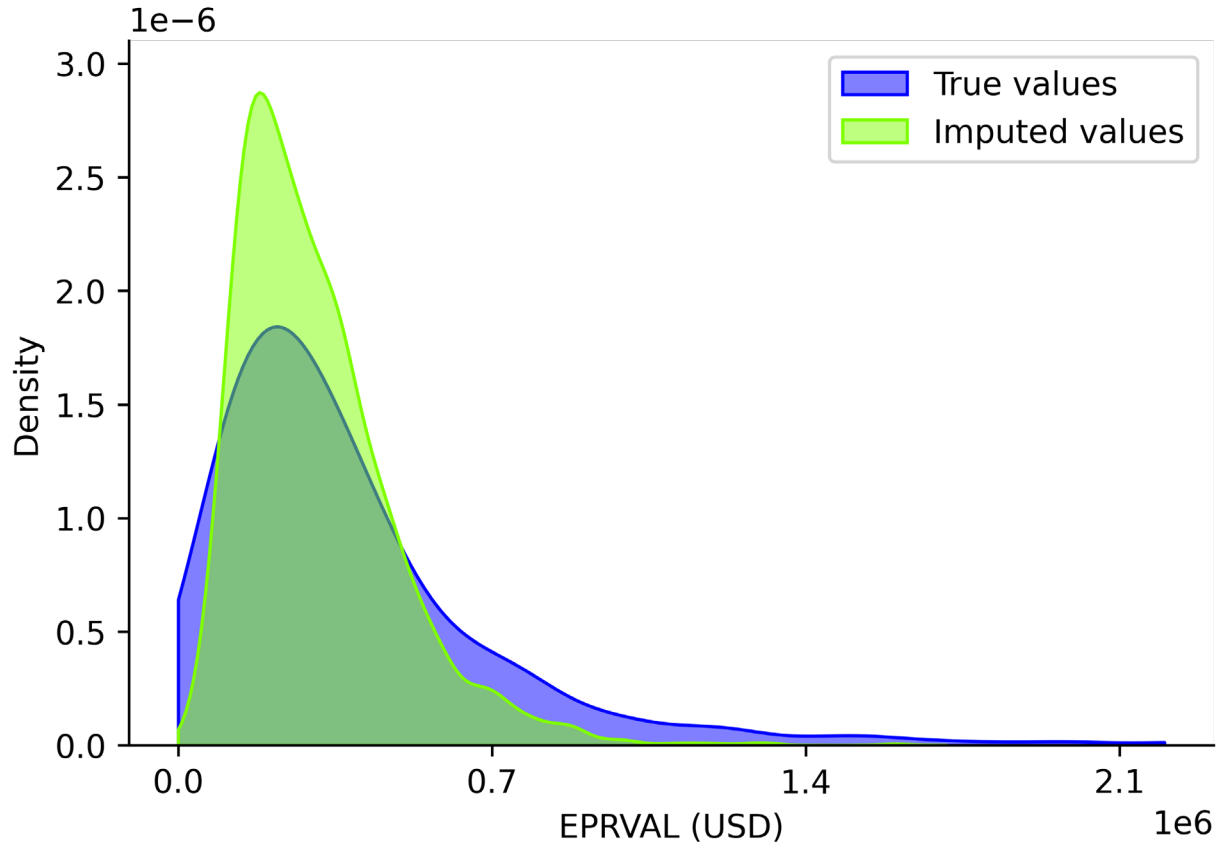
Kernel density plots of original versus imputed values



Hot Deck

(r = 0.46)

(RMSE = 510,100)



MICE

(r = 0.56)

(RMSE = 358,100)

United States® Census Bureau

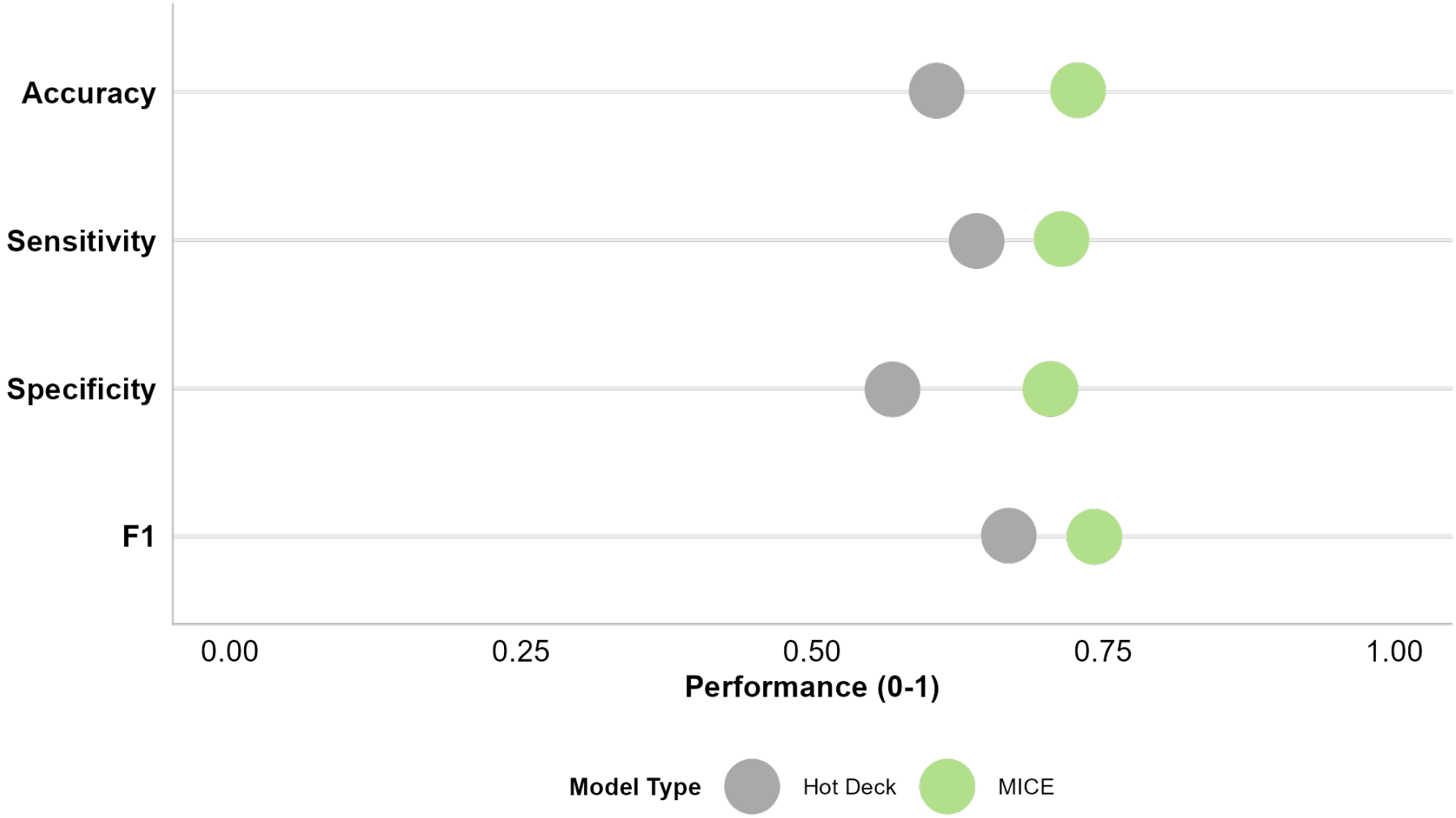
Results: Have Loan(s) on Primary Residence

$$\frac{Correct\ predictions}{All\ predictions}$$

$$\frac{Correct\ positive\ predictions}{All\ positive\ cases}$$

$$\frac{Correct\ negative\ predictions}{All\ negative\ cases}$$

$$2 * \frac{Precision * sensitivity}{Precision + sensitivity}$$

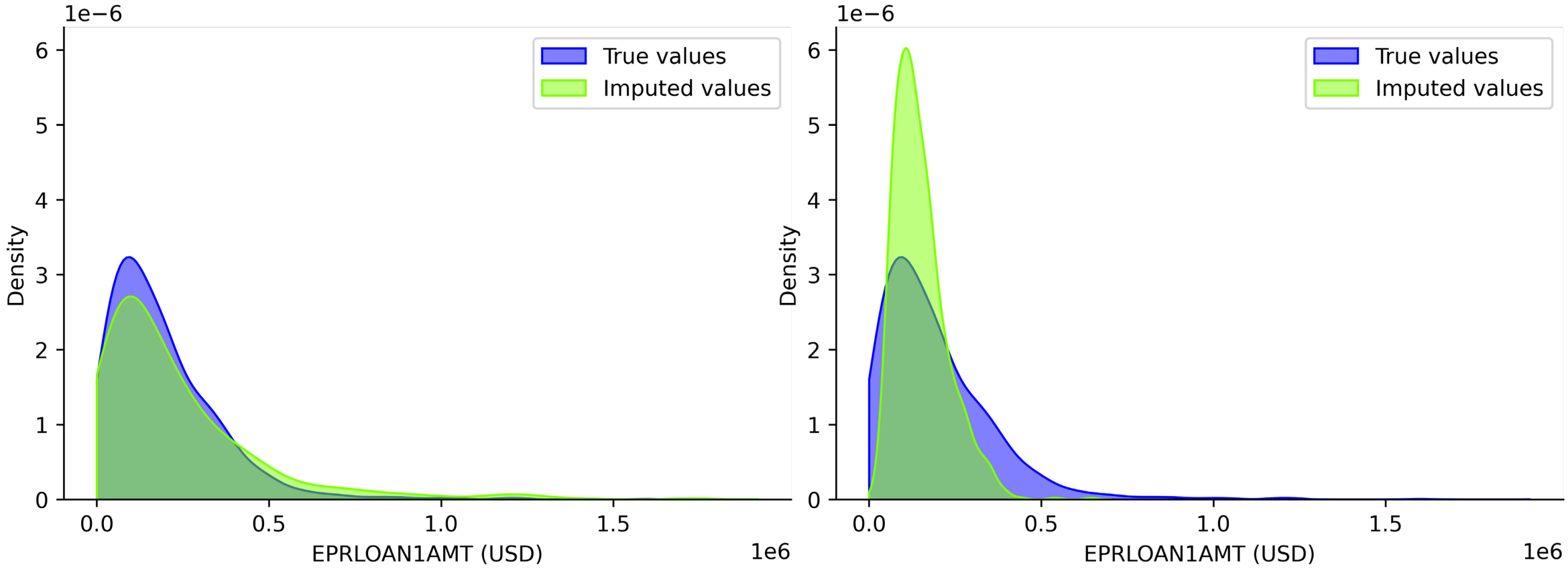


United States® Census Bureau

DMS P-7512692; DRB CBDRB-FY26-SEHSD003-017

Results: First Loan Amount

Kernel density plots of original versus imputed values



Hot Deck

(r = 0.73)

(RMSE = 173,411)

MICE

(r = 0.56)

(RMSE = 186,192)

United States® Census Bureau

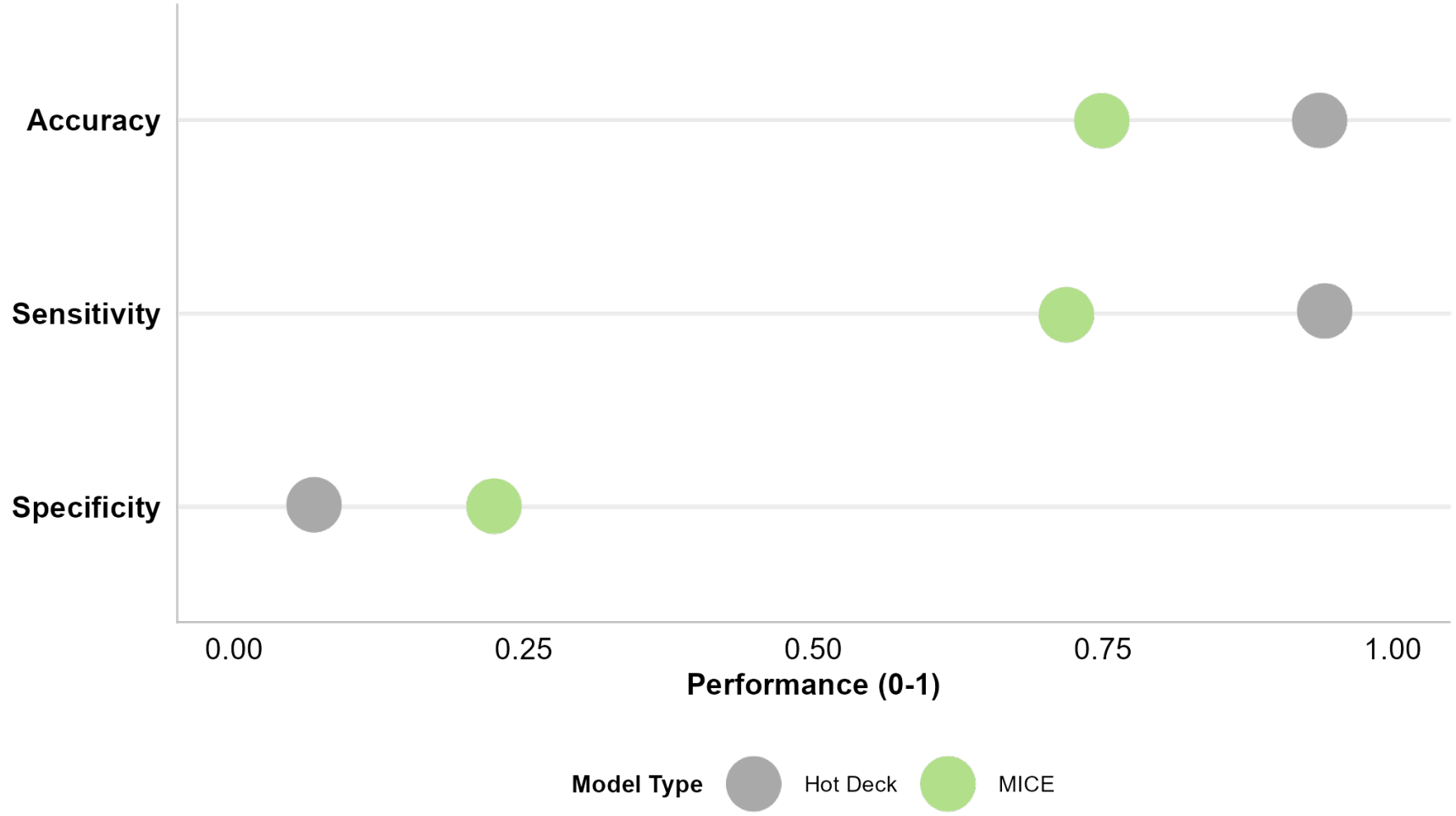
DMS P-7512692; DRB CBDRB-FY26-SEHSD003-017

Results: Number of Loans

$$\frac{\textit{Correct predictions}}{\textit{All predictions}}$$

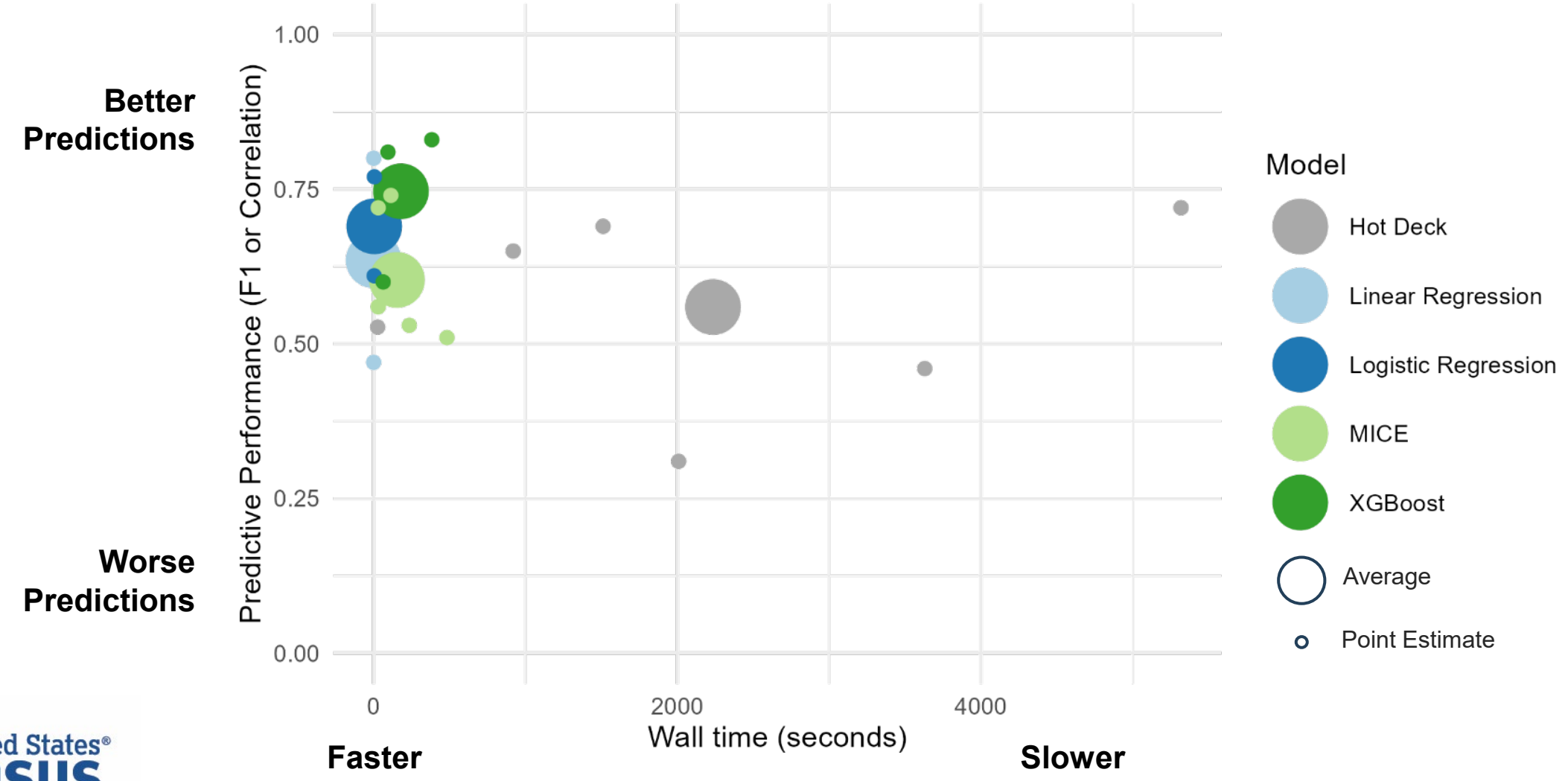
$$\frac{\textit{Correct positive predictions}}{\textit{All positive cases}}$$

$$\frac{\textit{Correct negative predictions}}{\textit{All negative cases}}$$



United States® Census Bureau

Accuracy vs. Processing Time



United States® Census Bureau

Discussion

- Hot deck generally underperforms compared to model-based imputation in both accuracy and processing time.

- Imputation recommendations.
  - XGBoost for categorical variables.
  - MICE framework with linear regression for continuous variables.
  - Hot deck may still be best for low-incidence variables.

- Caveats.
  - Test sample selected at random (assumes MCAR) -- no k-fold validations.
  - True missing rates may differ from the 20% benchmark we used.

United States® Census Bureau

Broader Implications for Survey Processing

- Things to consider.
  - When selecting focal variable:
    - How much missingness in variable?
    - How complicated are edit specs? More complexity → More processing time
    - Little dependency with highly imputed variables?
    - Which groups of variables could be imputed in the same framework?
  - When selecting predictors:
    - Are predictors highly predictive of outcomes?
    - Are edited values available at point of imputation relative to focal variable?
    - How consistently are they collected across waves?
    - Do they have high rates of missingness?

United States® Census Bureau

DMS P-7512692; DRB CBDRB-FY26-SEHSD003-017

Thank you!

Besufekad.Alemu@census.gov

SEHSD.SIRB.list@census.gov

United States® Census Bureau

Supplementary Slides

MICEnsus Creation

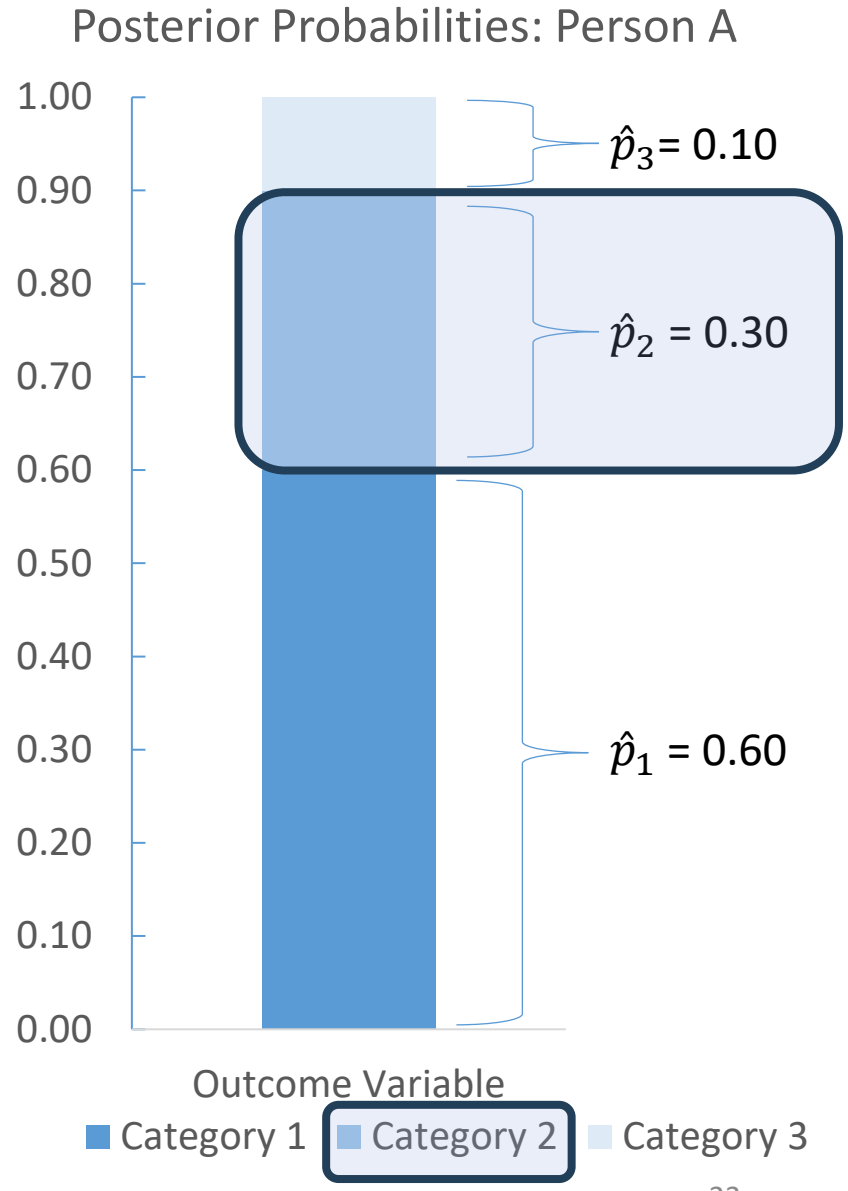
- Significant problems with existing Python MICE packages.
  - Treats all predictors as continuous.
  - Cannot handle multinomial models.
  - Cannot access intermediate models (black box).
- Our “custom-built” package:
  - Appropriately imputes all variable types, including handling of predicted probabilities in nonlinear models.
  - Allows for model specification for each focal variable.
  - Allows for conditional imputation based on imputed up-stream variables.
  - Produces “noisy” output in a log file.
  - Modeled after R’s MICE package.
- Stress testing needed before production!

United States® Census Bureau

DMS P-7512692; DRB CBDRB-FY26-SEHSD003-017

Custom MICE Process

- Initialize missing predictors with the mode (categorical) or mean (continuous).
- To simulate posterior predictive distribution:
  - **Continuous outcomes**: Random draw from the five real (observed) values closest to the predicted value (Predictive Mean Matching).
  - **Categorical outcomes**: Random draw weighted by predicted probabilities.

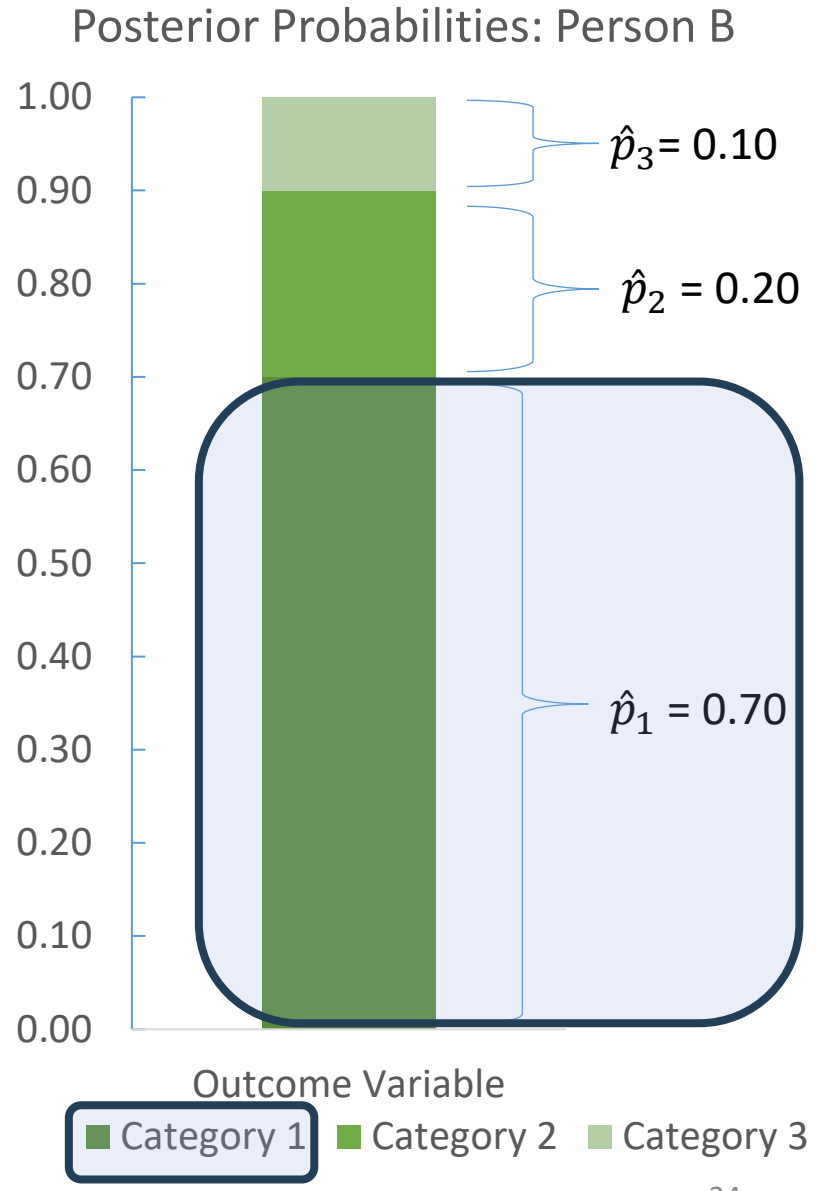


United States Census Bureau

DMS P-7512692; DRB CBDRB-FY26-SEHSD003-017

Custom MICE Process

- Initialize missing predictors with the mode (categorical) or mean (continuous).
- To simulate posterior predictive distribution:
  - **Continuous outcomes**: Random draw from the five real (observed) values closest to the predicted value (Predictive Mean Matching).
  - **Categorical outcomes**: Random draw weighted by predicted probabilities.



DMS P-7512692; DRB CBDRB-FY26-SEHSD003-017

MICE Discussion

- MICE program has more flexibility, needs more stress testing.
  - Inclusion of strata, scaling of imputed value by RMSE rather than predictive mean matching, incorporation of non-parametric models, etc.
- MICE under performing:
  - Predictive mean matching vs. deterministic imputed values of other methods.
  - MICE incorporates missing values in predictors compared to linear, logistic regression, which used edited values (cannot accommodate missing values in predictors).
  - Possible improvement with other pulled levers.
- Can uniquely impute multiple variables at once.
  - In our test, each focal variable was imputed separately. In production, multiple focal variables can be imputed within the same MICE framework.
  - Processing time for MICE will be lower overall than suggested here due to joint imputation of multiple focal variables.

United States® Census Bureau

Primary Residence

Primary Residence Predictors

1. Replicating current hot deck

- Government program participation.
- Age.
- Household income .

2. Hot deck on test dataset

- Demographics (education, marital status, race, Hispanic).
- House condition (cracks in floor, walls).
- Metropolitan status of residence location.
- Family structure.
- # of people in household.

3. Model-based imputation on test dataset

- Value of retirement account, asset income.
- State of residence location.
- Married or divorced within reference period.

United States® Census Bureau

Primary Residence MICE Specifications

- Revise MICE to allow conditional imputation:
  - Impute EPRDEBT only if EPRVAL > 0.
  - Impute EPRLOAN1AMT only if EPRDEBT=1.
- Omit variables that do not need to be imputed:
  - EPRLOAN_NUM.
- Imputed variables can serve as inputs to downstream variables:
  - EPRVAL serves as predictor when imputing EPRDEBT.
  - EPRVAL, EPRDEBT predictors for imputing EPRLOAN1AMT.
  - EPRVAL, EPRDEBT, EPRLOAN1AMT predictors for imputing EPRLOAN2AMT.
  - EPRVAL, EPRDEBT, EPRLOAN1AMT , EPRLOAN2AMT predictors for imputing EPRLOAN3AMT.
  - EPRLOAN_NUM is imputed outside of MICE based on EPRLOAN#AMT.

United States® Census Bureau

Variable Descriptives

EMPF Distribution

EMPF – SIPP 2022	Total Edited Distribution	Hot-Decked Distribution
Yes	3,600 (16.7% of total)	200 (16.7% of hot decked)
No	18,000	1,000
Column Totals	21,600	1,200 (5.6% of all obs)

United States® Census Bureau

EJB1_CLWRK Distribution

EJB1_CLWRK – SIPP 2022	Total Edited Distribution	Hot-Decked Distribution
1 – Federal government	600 (2.9% of total)	70 (2.8% of hot decked)
2 – Active duty military	70 (0.3%)	15 (0.6%)
3 – State government	1,100 (5.4%)	150 (5.9%)
4 – Local government	1,100 (5.4%)	150 (5.9%)
5 – Private, for-profit company	13,500 (66%)	1,900 (75%)
6 – Private, not-for-profit company	1,700 (8.3%)	250 (9.9%)
7 – Self-employed, incorporated business	800 (3.9%)	0
8 – Self-employed, not incorporated business	1,600 (7.8%)	0
Column Totals	20,470	2,535 (12.5% of total)

United States® Census Bureau

ESNAP Distribution

ESNAP_AMT – Quintile	Range ($)	Total Edited Distribution	DSD Imputed Distribution
1	[9,108]	550 (20.4% of total)	150 (21.4% of hot decked)
2	(108,193]	550 (20.4%)	150 (21.4%)
3	(193,250]	650 (24.1%)	150 (21.4%)
4	(250,387]	400 (14.8%)	100 (14.3%)
5	(387,1278]	550 (20.4%)	150 (21.4%)
Column Totals	[9,1278]	2,700	700 (25.9% of total)

Mean: $259
Standard deviation: $198

Min: $9
Max: $1,278

United States® Census Bureau

PR Variable Distributions: Continuous

Variable	Mean	SD	Median	n
EPRVAL	406,200	453,700	299,600	9,300
EPRLOAN1AMT	194,400	184,500	149,600	5,200
EPRLOAN2AMT	50,400	59,610	29,190	300

United States® Census Bureau

PR Variable Distributions: Categorical

EPRDEBT	Count	Percentage
0	4,100	44.1%
1	5,200	55.9%
N	9,300	100.0%

EPRLOAN_NUM	Count	Percentage
1	4,900	94%
2	300	5.8%
3+	15	0.3%
N	5,215	100.0%

United States® Census Bureau

Primary Residence Variables

Variable	Definition	N (observed and imputed)	Hot-Decked	In-Universe and Observed After Listwise Deletion
EPRVAL	primary residence value	11,000	1,100	9,300
EPRDEBT	mortgages or loans against residence	11,000	450	9,200
EPRLOAN_NUM	# of loans	6,200	350	5,200
EPRLOAN1AMT	amount of loan 1	6,200	900	5,200
EPRLOAN2AMT	amount of loan 2	350	40	308

United States® Census Bureau

Stratifiers and Predictors

EMPF Predictors

1. Replicating current hot deck

2. Hot deck on test dataset

3. Model-based imputation on test dataset

- Demographics: age, sex, marital status.
- Number of times married.
- Education.
- Race/ethnicity.
- Fertility: children ever born, age at first birth, age at first birth squared.
- 2021 variables: marital status, EMPF.
- Age squared.
- Interactions.

United States® Census Bureau

EJB1_CLWRK Predictors

1. Replicating current hot deck

2. Hot deck on test dataset

3. Model-based imputation on test dataset

- Sex
- 13-category occupation code.
- Demographics – Education.
- Residence – Region of residence, metro status, state of residence.
- Job information - # of months working, union status, working from home, 14-category industry.
- Previous year CLWRK.
- Residence – Located in state capital zip code.

United States® Census Bureau

ESNAP_AMT Predictors

1. Replicating current hot deck

2. Hot deck on test dataset

3. Model-based imputation on test dataset

- Household size.
- Contiguous U.S. vs. AK/HI.
- Number of months worked in year.
- Demographics - sex, age, race/ethnicity, education, marital status.
- Program receipt - WIC, TANF.
- Work-limiting disability.
- Previous years – SNAP receipt, receipt of SNAP 2+ years prior.
- SNAP information – # of months received, reasons for starting/ending receipt.
- # of recipients – Within spell, within household.

United States® Census Bureau