Imputing Responses for Manufacturing Establishments Using a Mixed Model under a Matrix Sub-sample Design

Yeng Xiong

US Census Bureau

Federal Statistics Committee Meetings 2024

Team

- Stephen Kaputa (US Census Bureau)
- Scott H. Holan (University of Missouri, US Census Bureau)

Sample Design

AIES

Subsample

Matrix Sample

- Economy wide
- Stratified by NAICS3 and geography
- Sequential probability proportional to size (payroll)
- 4 core items

- Manufacturing-focused
- Same stratum definitions
- Equal probability
- 50+ items

AIES = Annual Integrated Economic Survey
NAICS = North American Industry Classification System

United States® Census Bureau

		Frame covariate	AIES core items		Matrix item
AIES Sampled Unit	Matrix Sample Indicator	MOS	Item 1	Item 2	Item 3
1	1	x_1	y_{11}	y_{21}	z_{11}
2	0	x_2	y_{12}	y_{22}	?
3	0	x_3	y_{13}	y_{23}	?
4	1	x_4	y_{14}	y_{24}	z_{14}
5	1	x_5	y_{15}	y_{25}	z_{15}
6	0	x_6	y_{16}	y_{26}	?
7	0	x_7	y_{17}	y_{27}	?
8	0	x_8	y_{18}	y_{28}	?
9	1	x_9	y_{19}	y_{29}	z_{19}
…	…	…	…	…	…

Imputation approach

1. Fit a Bayesian linear mixed model using frame covariates and responses from the matrix sample

2. Impute responses for AIES units not selected in matrix sample

3. Compute estimates of domain totals

- Goal: have a lower root mean squared prediction error than design-based estimates

		Matrix item
AIES Sampled Unit	Matrix Sample Indicator	Item 3
1	1	z_{11}
2	0	z_{12}
3	0	z_{13}
4	1	z_{14}
5	1	z_{15}
6	0	z_{16}
7	0	z_{17}
8	0	z_{18}
9	1	z_{19}
…	…	…

United States® Census Bureau

Challenges in model building

- Multiple outcome variables with complex relationships
- Frequent zero-valued observations in a some of the variables
- Highly skewed data
- Varying industry and geography estimation levels
- Need a model that is generalizable
    - Fit all (or most) outcomes
    - Handle zeros
    - Applicable across estimation levels

United States® Census Bureau

Items and outcome variables

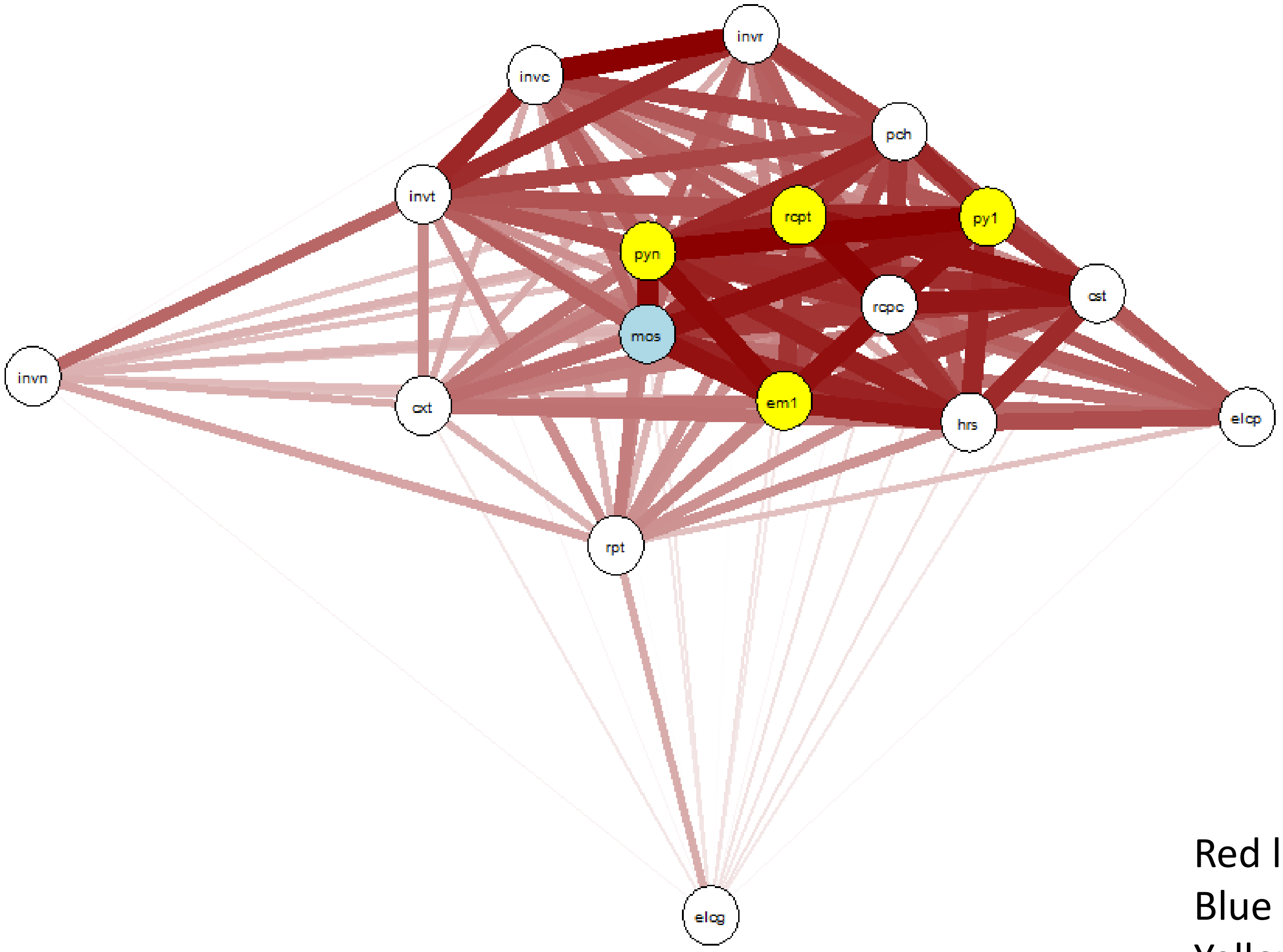
Outcome	Description	# of detailed items	Additional variables
hrstotm	Total production worker hours	3	
rcpecomt	Receipts for e-commerce ($)		rcpecmtp = % of e-commerce receipts
pchtt	Total purchased services	12	
rptot	Total rent payments	2	
cstmtot	Total cost of materials	5	
elecgen	Electricity generated		
elecpch	Electricity purchased		
elecsld	Electricity sold		
cextot	Total capital expenditures	7	cexbld = expenditures on buildings cexmch = expenditures on machinery cexmcha = expenditures on automobiles cexmchc = expenditures on robotics
invtotb/invtote	BOY/EOY value of inventory	6	
invcb/invce	BOY/EOY LIFO inventory valuation	2	
invnctb/invncte	BOY/EOY non-LIFO inventory valuation	8	
invvtob/invvtoe	BOY/EOY total inventory valuation	2	
invrsvb/invrsve	BOY/EOY LIFO reserve	2	
valaddm	Value added		

BOY = Beginning of year
EOY = End of year

Selected outcomes

Outcome	Description	# of detailed items	Additional variables
hrstotm	Total production worker hours		
rcpecomt	Receipts for e-commerce ($)		
pchtt	Total purchased services		
rptot	Total rent payments		
cstmtot	Total cost of materials		
elecgen	Electricity generated		
elecpch	Electricity purchased		
cextot	Total capital expenditures		
invtote	EOY value of inventory		
invce	EOY LIFO inventory valuation		
invncte	EOY non-LIFO inventory valuation		
invrsve	EOY LIFO reserve		

BOY = Beginning of year
EOY = End of year



Red lines = positive correlation
Blue circle = frame covariate
Yellow circle = AIES core items
White circle = outcomes

Source: U.S. Census Bureau, 2023; DRB approval number CBDRB-FY24-ESMD005-003

Linear mixed model – full model

Within NAICS4, each matrix item is independently model as

$$\log z_{vsji} = \beta_0 + \beta_1 \log x_i + \beta_2 \log y_{3i} + \beta_3 \log y_{4i} + \gamma_s + \delta_j + \epsilon_i$$

- z_{vsji} = response of v^{th} matrix item for establishment i in state s operating in NAICS6 industry j

Linear mixed model – full model

Within NAICS4

$$\log z_{vsji} = \beta_0 + \beta_1 \log x_i + \beta_2 \log y_{3i} + \beta_3 \log y_{4i} + \gamma_s + \delta_j + \epsilon_i$$

- Linear regression modeling a national relationship between response and frame covariate MOS

Linear mixed model – full model

Within NAICS4

$$\log z_{vsji} = \beta_0 + \beta_1 \log x_i + \beta_2 \log y_{3i} + \beta_3 \log y_{4i} + \gamma_s + \delta_j + \epsilon_i$$

- Linear regression modeling a national relationship between response and AIES core items of receipts and employment

Linear mixed model – full model

Within NAICS4

$$\log z_{vsji} = \beta_0 + \beta_1 \log x_i + \beta_2 \log y_{3i} + \beta_3 \log y_{4i} + \gamma_s + \delta_j + \epsilon_i$$

$$\gamma_s \sim N(0, \sigma_s^2)$$

- Random effect for state allowing deviation from the national trend

United States® Census Bureau

Linear mixed model – full model

Within NAICS4

$$\log z_{vsji} = \beta_0 + \beta_1 \log x_i + \beta_2 \log y_{3i} + \beta_3 \log y_{4i} + \gamma_s + \delta_j + \epsilon_i$$

$$\delta_j \sim N\left(0, \sigma_j^2\right)$$

- Random effect for NAICS6 industry allowing deviation from national NAICS4 industry trend

Linear mixed model – full model

Within NAICS4

$$\log z_{vsji} = \beta_0 + \beta_1 \log x_i + \beta_2 \log y_{3i} + \beta_3 \log y_{4i} + \gamma_s + \delta_j + \epsilon_i$$

$$\epsilon_i \sim N\left(0, \sigma_i^2\right)$$

- Residual error

Simulation

- Generate 1000 samples from research frame following the AIES and matrix sample designs
- Fit full linear mixed model to impute missing responses
  - Use maximum likelihood approximation
- Calculate domain estimates $\hat{\theta}^{d}$ with a ratio estimator
  - Domain = NAICS4 x state
- Showing results for single NAICS4 industry

Evaluation Criteria

Relative absolute bias (RAB)

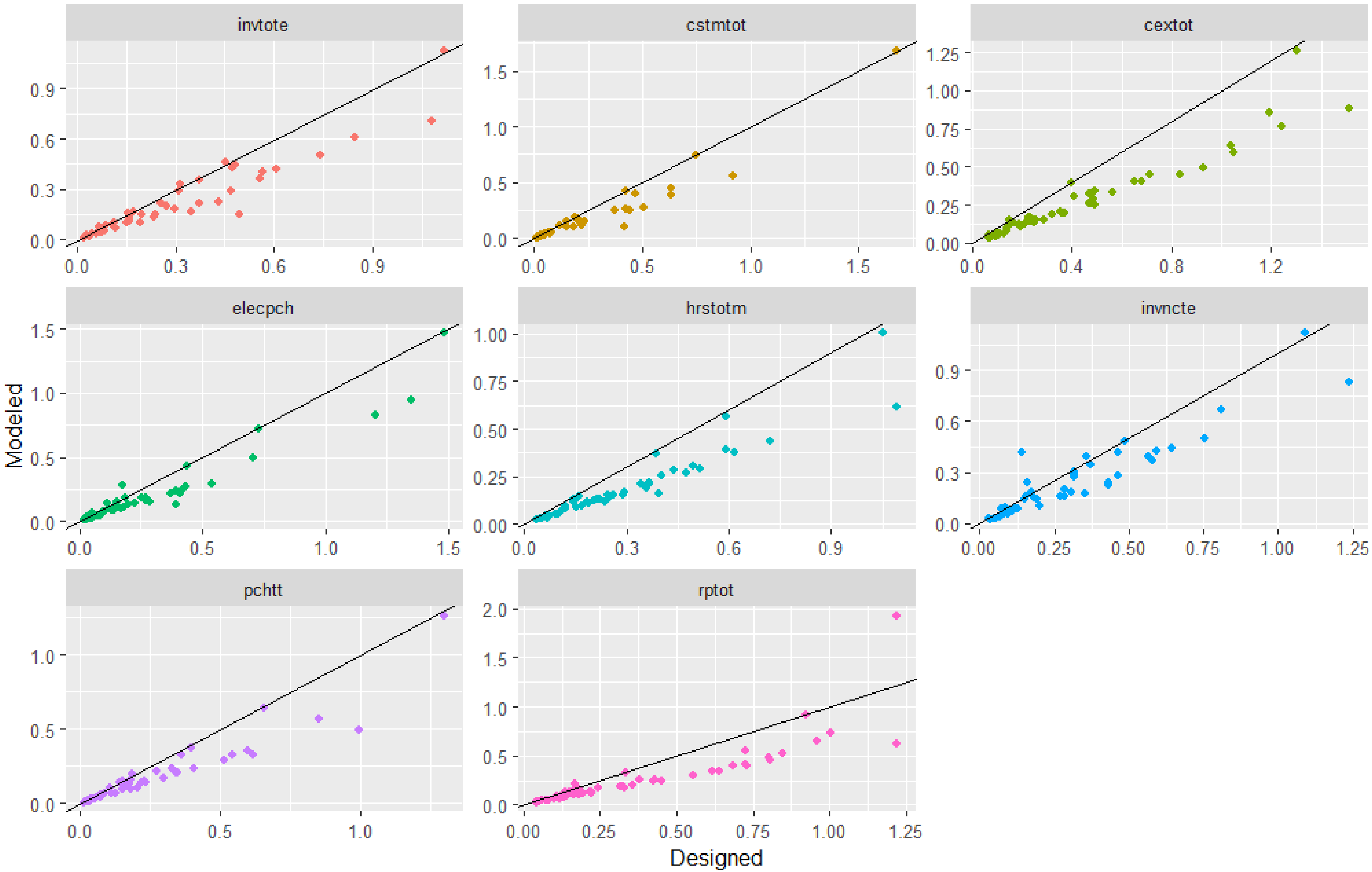
$$\text{RAB} = \frac{1}{1000}\sum_{r}\frac{\left|\hat{\theta}_r^d - \theta_r^d\right|}{\theta_r^d}$$

Reduction in root mean-squared prediction-error (RMSPE)

$$\text{RMSPE} = \sqrt{\frac{1}{1000}\sum_{r}\left(\hat{\theta}_r^d - \theta_r^d\right)^2}$$

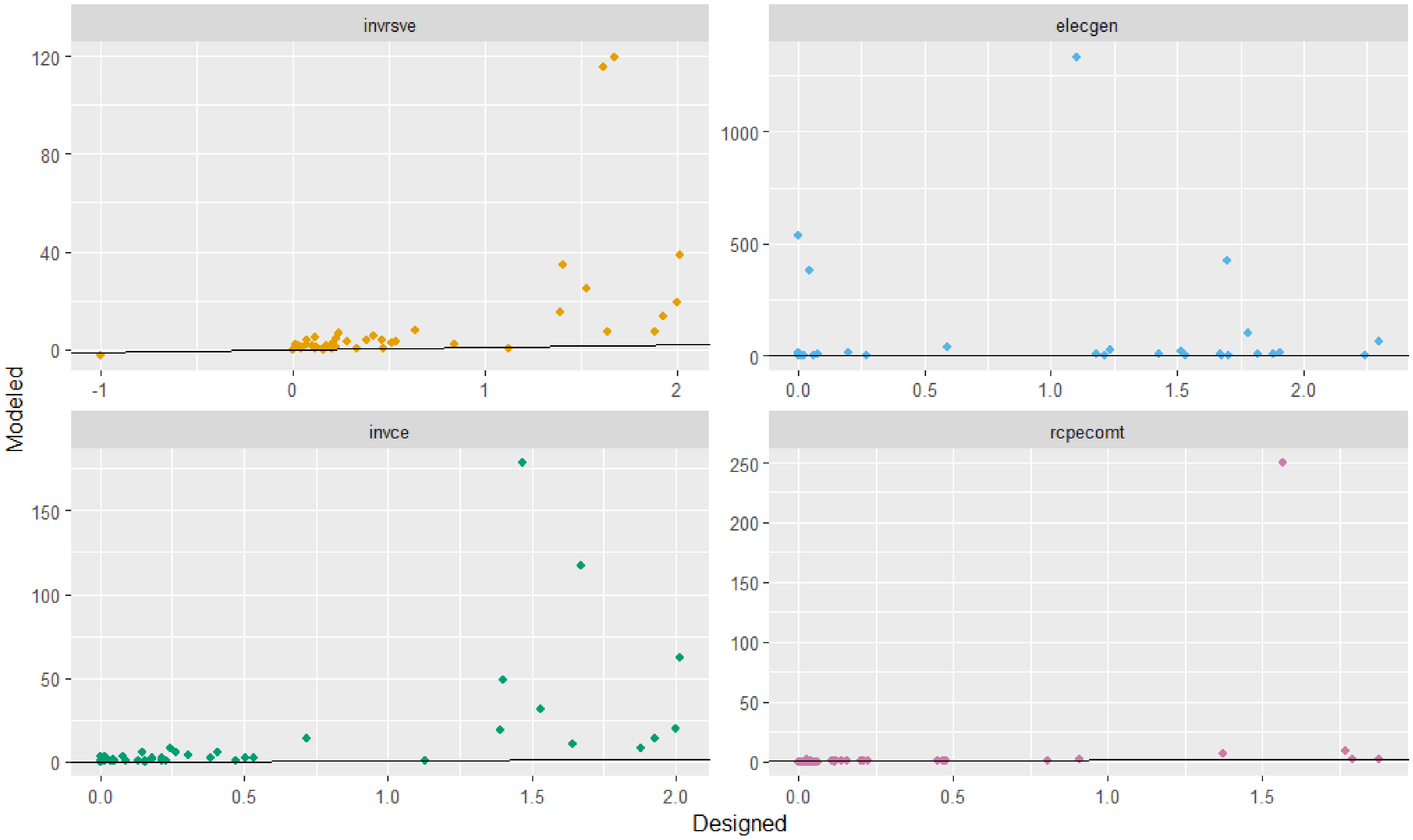
$$\text{Reduction} = \frac{\text{RMSPE}_{\text{model}}}{\text{RMSPE}_{\text{designed}}}$$

United States® Census Bureau

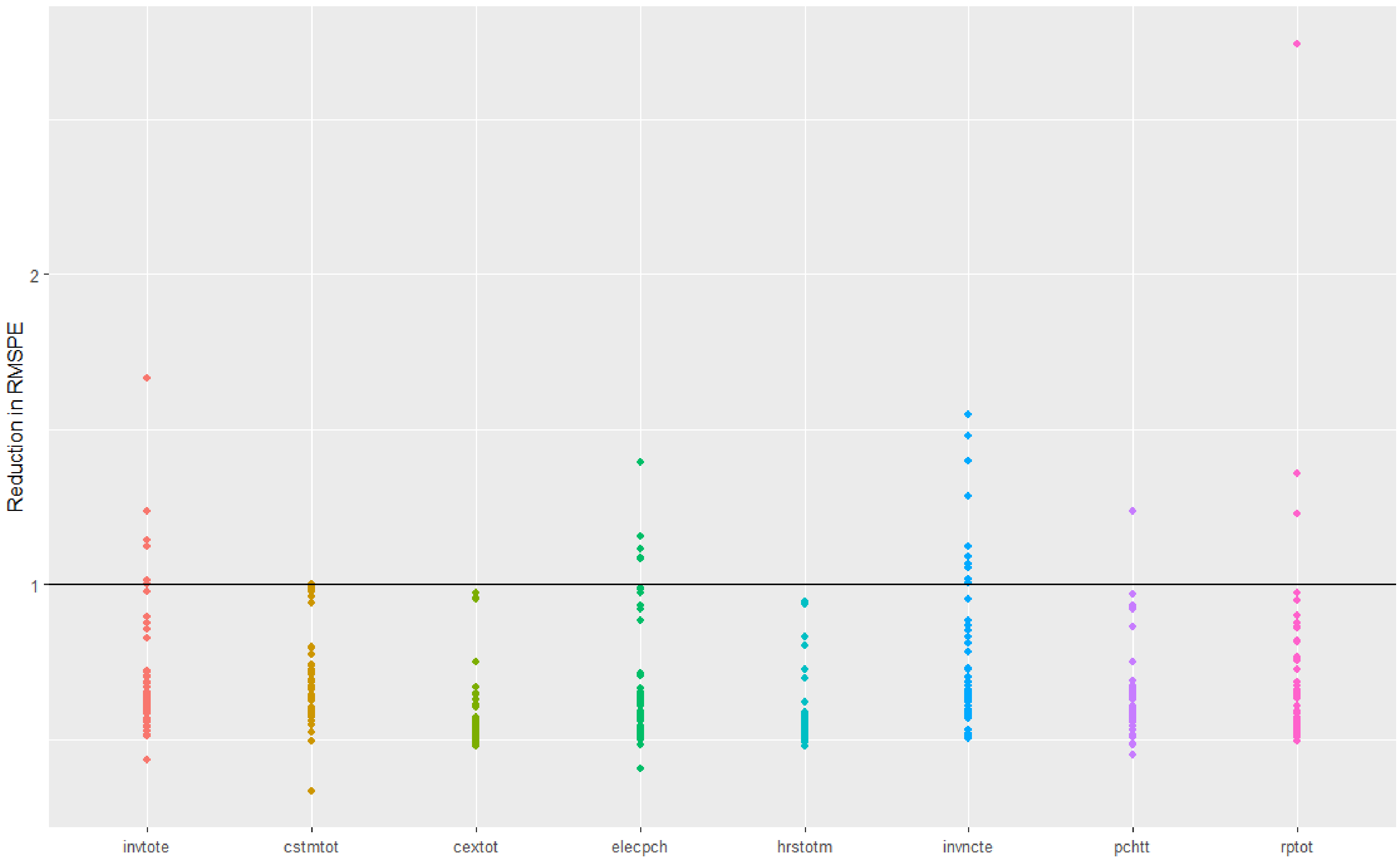


Source: U.S. Census Bureau, 2023; DRB approval number CBDRB-FY24-ESMD005-003

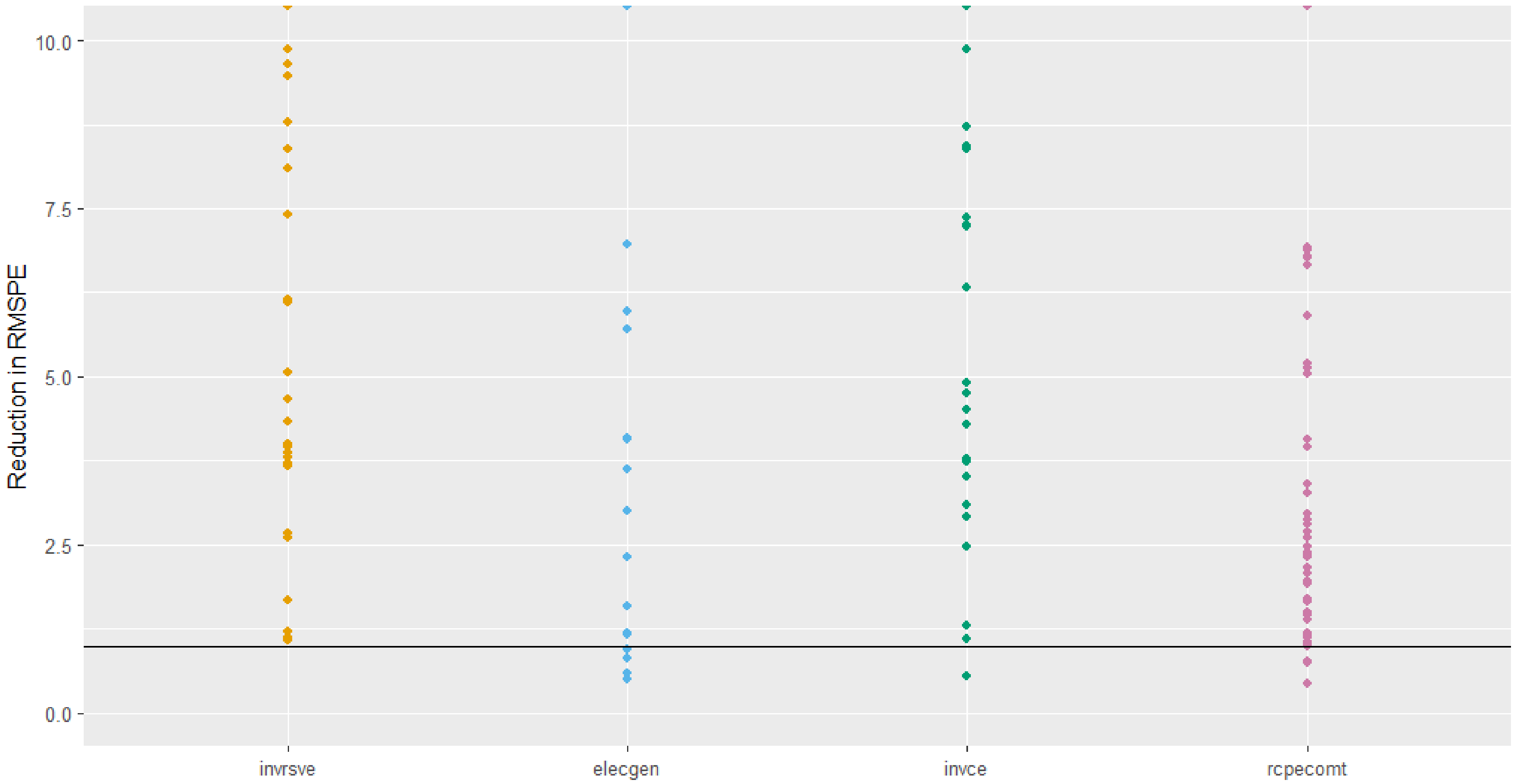
United States Census Bureau



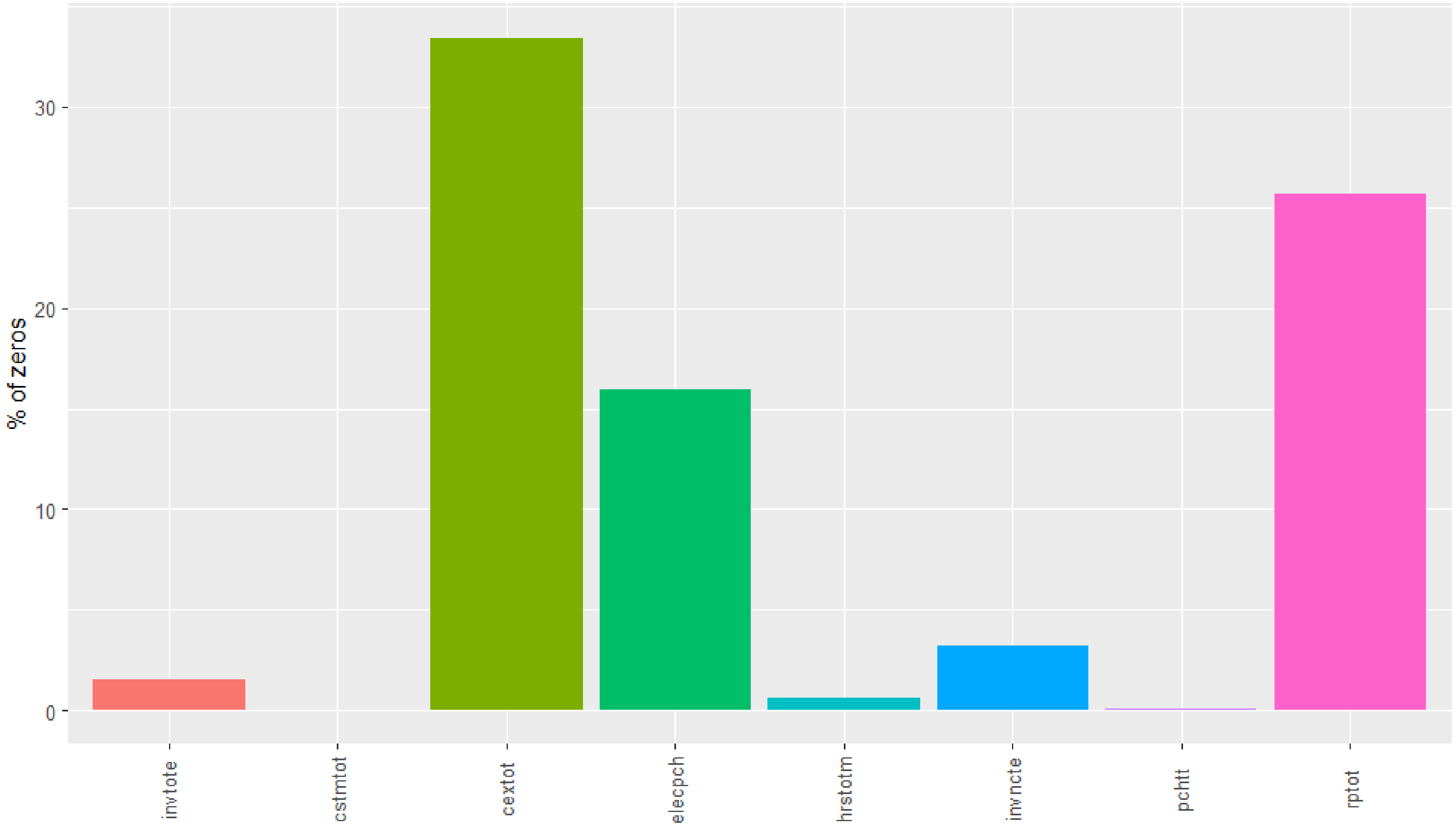
Source: U.S. Census Bureau, 2023; DRB approval number CBDRB-FY24-ESMD005-003



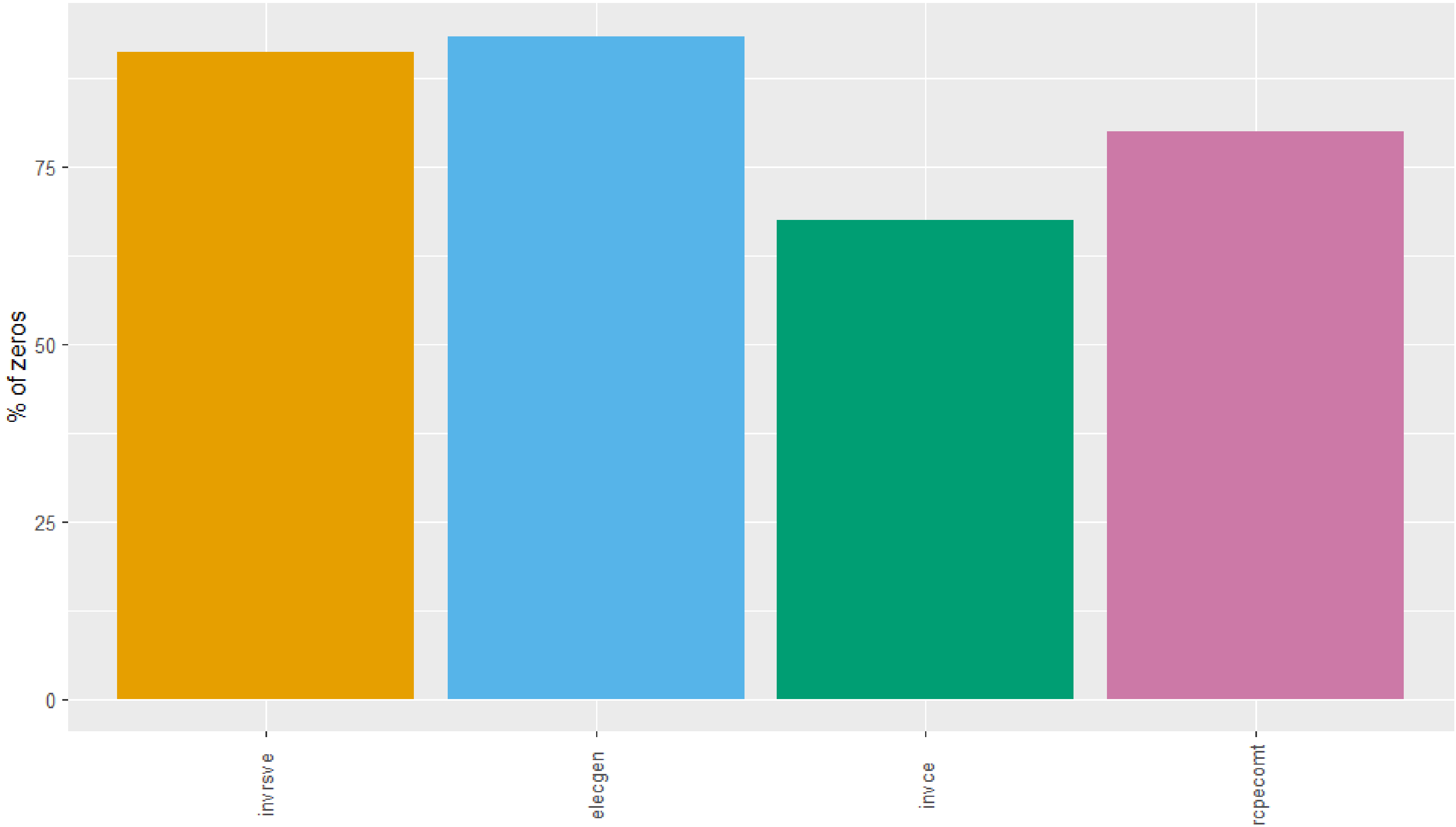
Source: U.S. Census Bureau, 2023 ; DRB approval number CBDRB-FY24-ESMD005-003



Source: U.S. Census Bureau, 2023 ; DRB approval number CBDRB-FY24-ESMD005-003
Note: y-axis is truncated



Source: U.S. Census Bureau, 2023 ; DRB approval number CBDRB-FY24-ESMD005-003



Source: U.S. Census Bureau, 2023 ; DRB approval number CBDRB-FY24-ESMD005-003

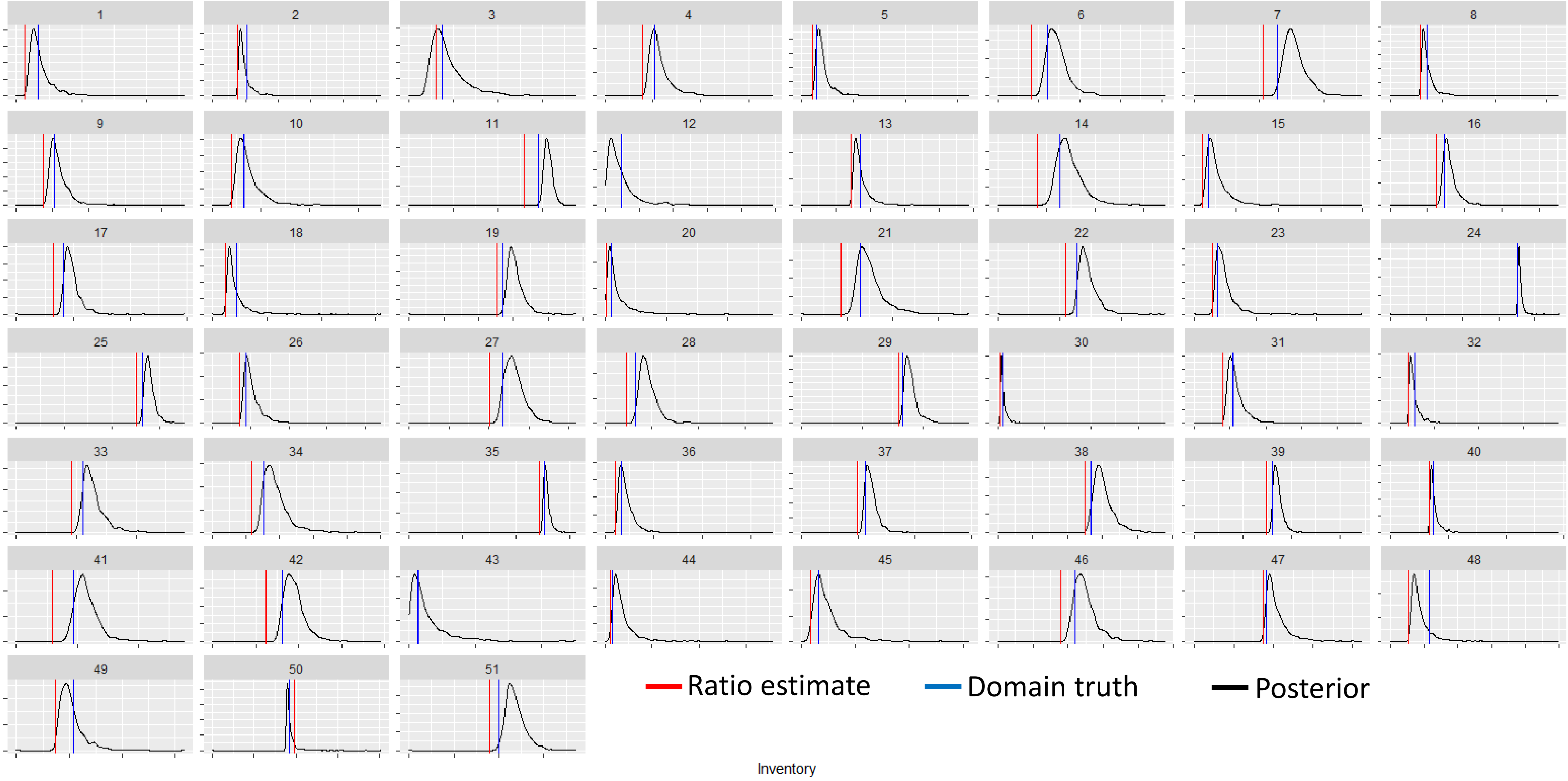
Empirical Application

- Take a sample from production frame following AIES and matrix sample designs

- Fit full linear mixed model
  - Use Bayesian imputation model
  - Implemented with “Stan” in R

- Obtain posterior distribution of estimated domain totals
  - Back-transform variable and ratio adjust totals
  - Domain = naics4 x state

- Compare with design-based ratio estimate and true domain total



Source: U.S. Census Bureau, 2023 ; DRB approval number CBDRB-FY24-ESMD005-003

Discussion

- Work in progress but promising results
  - Generally comparable or lower bias for most variables
  - Generally lower MSE for most variables
  - Can produce estimates for small domains with no observable data
- Future research
  - Improve prediction for variables with high percentage of zeros
    - Predicting zeros first and then positive values
  - Produce estimates for detailed items
  - Combining sampling variability and imputation variability
  - Evaluate whether variable should be included in short-form survey

United States® Census Bureau

Thank you!

- Email: yeng.xiong@census.gov