Variance Modeling for Differentially Private Counts

Kyle M. Irimata

Center for Statistical Research and Methodology
U.S. Census Bureau

August 8, 2024
JSM 2024

Disclaimer

This presentation is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the presenter and not those of the U.S. Census Bureau.

Outline

1. Census Redistricting Data
2. Variance Estimation and Modeling
3. 2010 PL94-171 Modeling
4. Conclusions

United States Census Bureau

Redistricting Data

PL94-171 Congressional Redistricting data

- Primary use of this data is for creating redistricting plans
- Provides estimates for the counts by OMB race and Hispanicity groups
- For the 2010 Decennial Census, this data was released in Summary File 1 (SF1), which was privacy-protected using data swapping
- Beginning with the 2020 Decennial Census, these counts were protected using differential privacy through the TopDown Algorithm (TDA)

United States Census Bureau

Differential Privacy

- Differential Privacy (DP) involves the addition of distributional noise to the true measurements
  - $\tilde{\theta}_i = \theta_i + \xi_i$, where $\xi_i \sim (0, \sigma^2_{DP})$
- Two commonly used distributions in the application of DP to count data are the Laplace and Discrete Gaussian distributions
- The variance of the random noise is defined by a privacy loss budget (PLB), commonly denoted by ϵ or ρ, which is set in advance
  - σ^2_{DP} is inversely proportional to the PLB
- The PLB is publicly available/published

United States Census Bureau

TopDown Algorithm (TDA)

The TDA was used to protect respondent confidentiality in the 2020 Decennial Census

- Incorporates both the additive differentially private noise, as well as post-processing
- Post-processing was used to enforce constraints such as:
  - Invariants (state level population counts are fixed)
  - Hierarchical consistency (lower level counts sum to higher level counts)
  - Non-negativity
  - Housing and group quarters constraints
- Produced counts at every level of geography
  - For this work, the focus will be on block groups

United States Census Bureau

TopDown Algorithm (TDA)

- In preparation for the 2020 data release, the Census Bureau released a series of "demonstration products" using the 2010 PL94-171 Census counts
- The PLB and post-processing were adjusted iteratively
- The "production settings" were used for the 2021-06-08 demonstration data release with 2010 data
- Second application of the "production settings" were used for the 2023-04-03 demonstration data release with 2010 data
- **How can we quantify the error in the TDA (DP and post-processing) using only publicly available data?**

United States Census Bureau

Esri Extended Error Metrics

- Esri (2023) produced the "Esri Extended Error Metrics" and presented an arcgis tool to visualize TDA noise in the 2020 Census PL94-171 data at various geographies
- Grouped geographies by population decile
- Mean absolute percent error (MAPE) between the 2010 PL94-171 demonstration product and the 2010 SF1 counts
  - Used the MAPE from 2010 as a measure of variability on 2020 Census counts
  - Assigned the same MAPE to each geographic unit within the decile

United States Census Bureau

Small Area Estimation: The Fay-Herriot Model

- θ_i is the population characteristic of interest for area i.
- Y_i is the direct survey estimate of θ_i:

$$Y_i = \theta_i + e_i \qquad \theta_i = x_i'\beta + v_i \qquad i = 1, \ldots, m$$

- First part is sampling model; the second part is linking model.
- e_i is the sampling error in Y_i, generally assumed to be $\overset{ind}{\sim} N(0, \sigma_i^2)$, with σ_i^2 known.
- v_i is the area i random effect, commonly assumed to be $\overset{iid}{\sim} N(0, \phi^2)$ and independent of the e_i.
- x_i is the vector of covariates and β the vector of regression coefficients and intercept.

United States Census Bureau

Small Area Estimation: Variance Models

$$Y_i = \theta_i + e_i \qquad \theta_i = x_i'\beta + v_i \qquad i = 1, \ldots, m$$

- Generally, small area estimation (SAE) models assume some known sampling variances σ_i^2
- Misspecification of the σ_i^2 can lead to issues in the estimation, especially of the MSE
- Typically, sampling variances are estimated (s_i^2) and common ways to deal with this are:
  - Assume the variance is known
  - Generalized variance functions (GVF)
  - Joint modeling (modeling both the point estimates and variance)

United States Census Bureau

Initial Variance Estimation

- In SAE, we usually assume the availability s_i^2; however, these are not immediately available in this application
- Modification of Esri estimators: decile grouping by state
- For group $g(s) = 1, ..., 10$, within state s

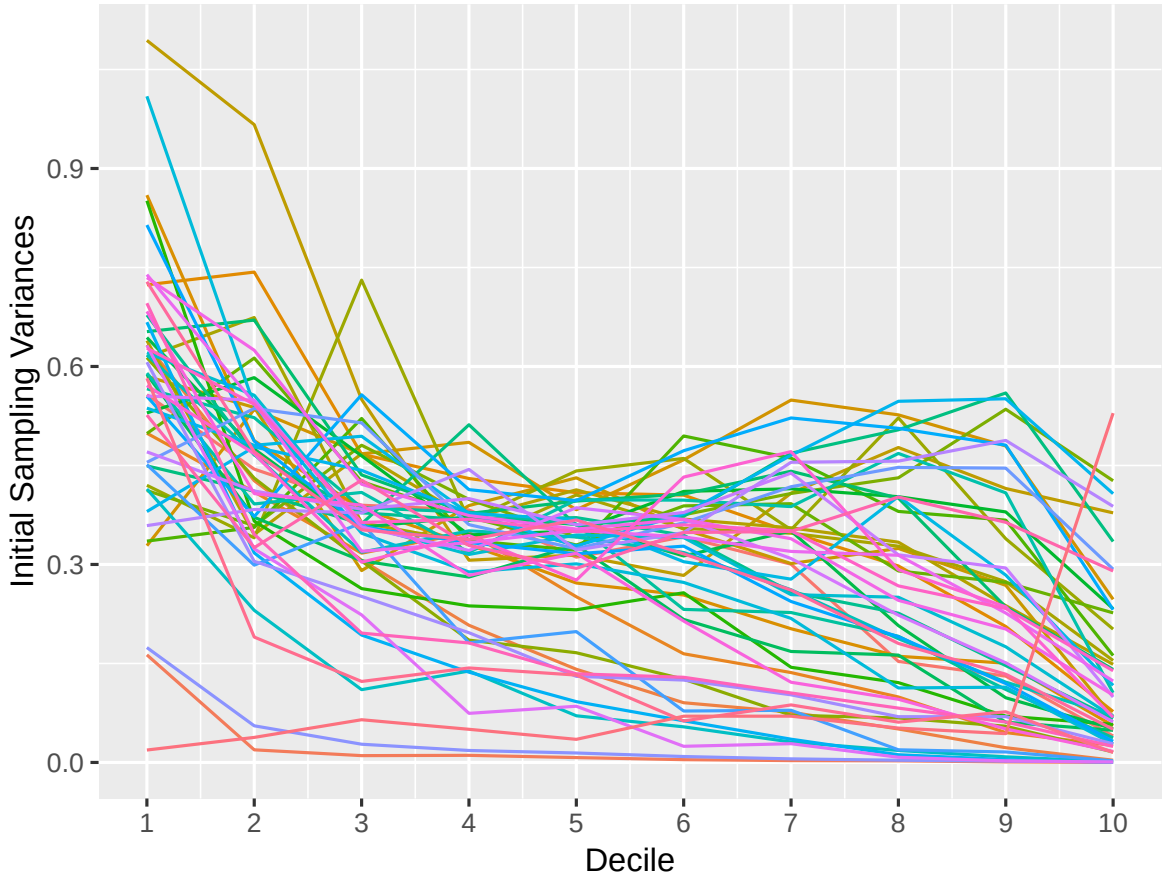
$$s^2_{g(s)} = \frac{1}{n_{g(s)}} \sum_{i=1}^{n_{g(s)}} (y_{ig(s)} - \theta^{SF}_{ig(s)})^2$$

- $y_{ig(s)}$ denotes the count from the 2010 demonstration product
- $\theta^{SF}_{ig(s)}$ denotes the count from the 2010 SF1 (swapping counts)
- For each block group $ig(s)$, set $s_i^2 = s^2_{g(s)}$

United States Census Bureau

Initial Variance Estimates by state: AIAN Alone

American Indian and Alaska Native (AIAN) alone



United States Census Bureau

Joint Hierarchical Bayesian FH Model

Extension of You (2021):

- $y_i|\theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$
- $d_i s_i^2|\sigma_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$
  - n_i is the sample size in area i
- $\theta_i|\beta, \phi^2 \sim N(x_i'\beta, \phi^2)$
- $log(\sigma_i^2) \sim N(z_i'\delta, \tau^2)$

for $i = 1, ..., m$ denoting the areas and x_i, z_i denoting suitable covariates

- Flat priors for $\beta, \delta, \phi^2, \tau^2$
- Gibbs sampling for computing

United States Census Bureau

PL94-171 Modeling

- 2023-04-03 demonstration data (2010 PL94-171)
  - Obtain initial decile within state variance estimates (s_i^2)
  - Select covariates (x_i, z_i), develop model form
- 2021-06-08 demonstration data (2010 PL94-171)
  - Fit the model and obtain estimates
  - Use the s_i^2 from the 2023-04-03 data as the initial variance estimates
- Population counts for:
  - Total population
  - American Indian and Alaska Native (AIAN) alone

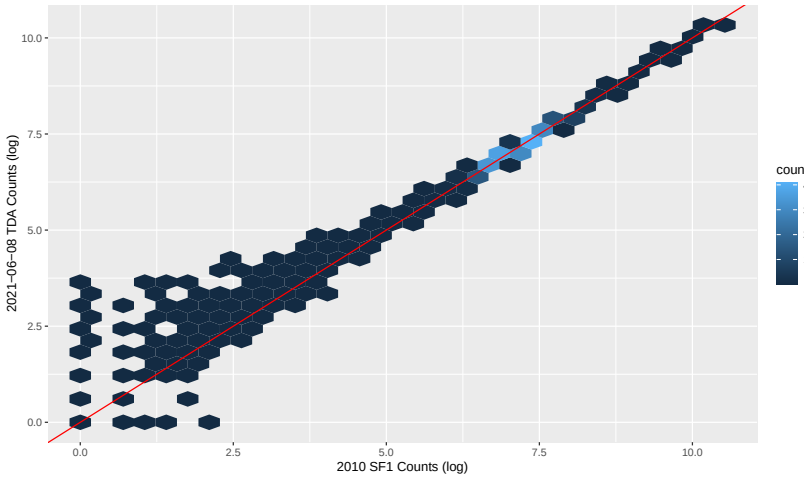
United States Census Bureau

Covariate Data

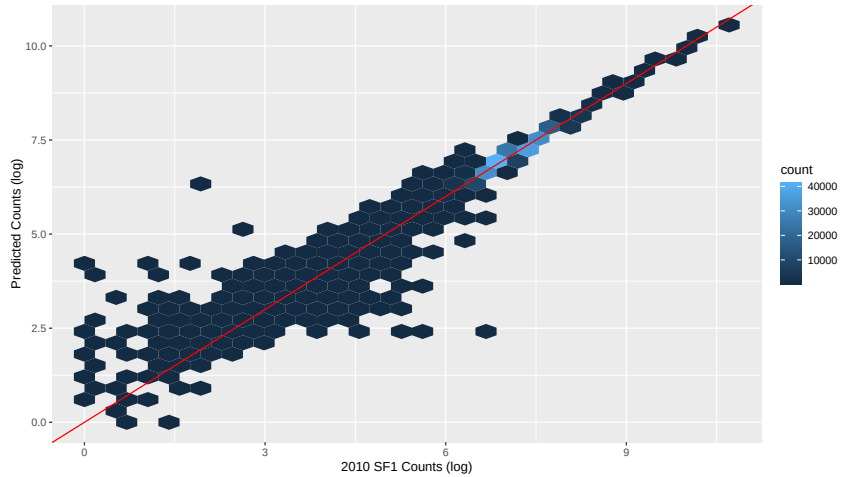
- Covariates for the Mean:
  - 2010 ACS Table B03002: Hispanic or Latino Origin by Race
  - Block group total population (when modeling racial subgroups)
- Covariates for the Variance:
  - Number of block groups in the grouping
  - Number of blocks in each block group
  - Count of housing units (HU)
  - Count of group quarters (GQ)
  - Census 2000 PL94-171 counts
    - Block equivalency file to produce estimates using 2010 geographies

United States Census Bureau

Point Estimates for Total Population

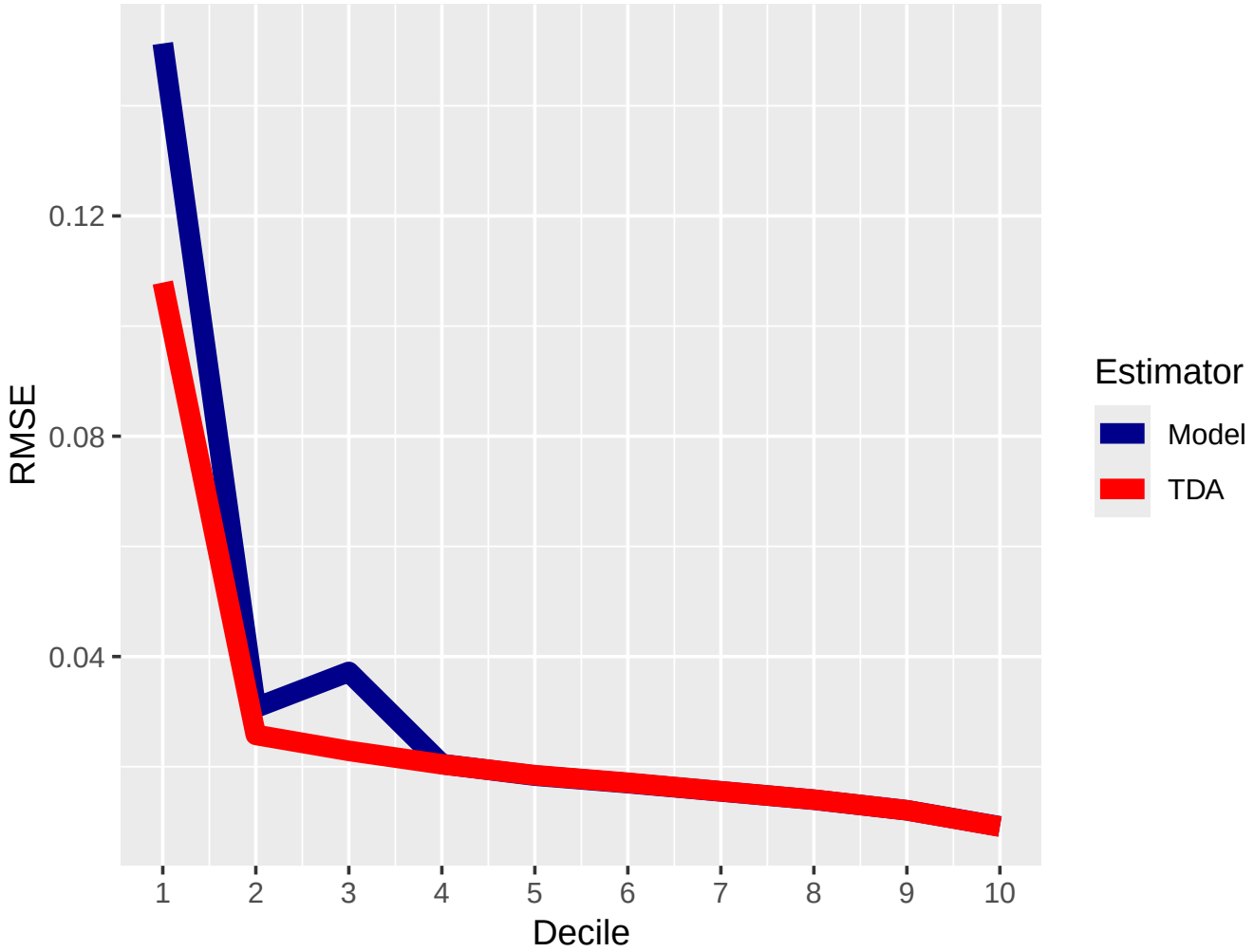


(a) TDA

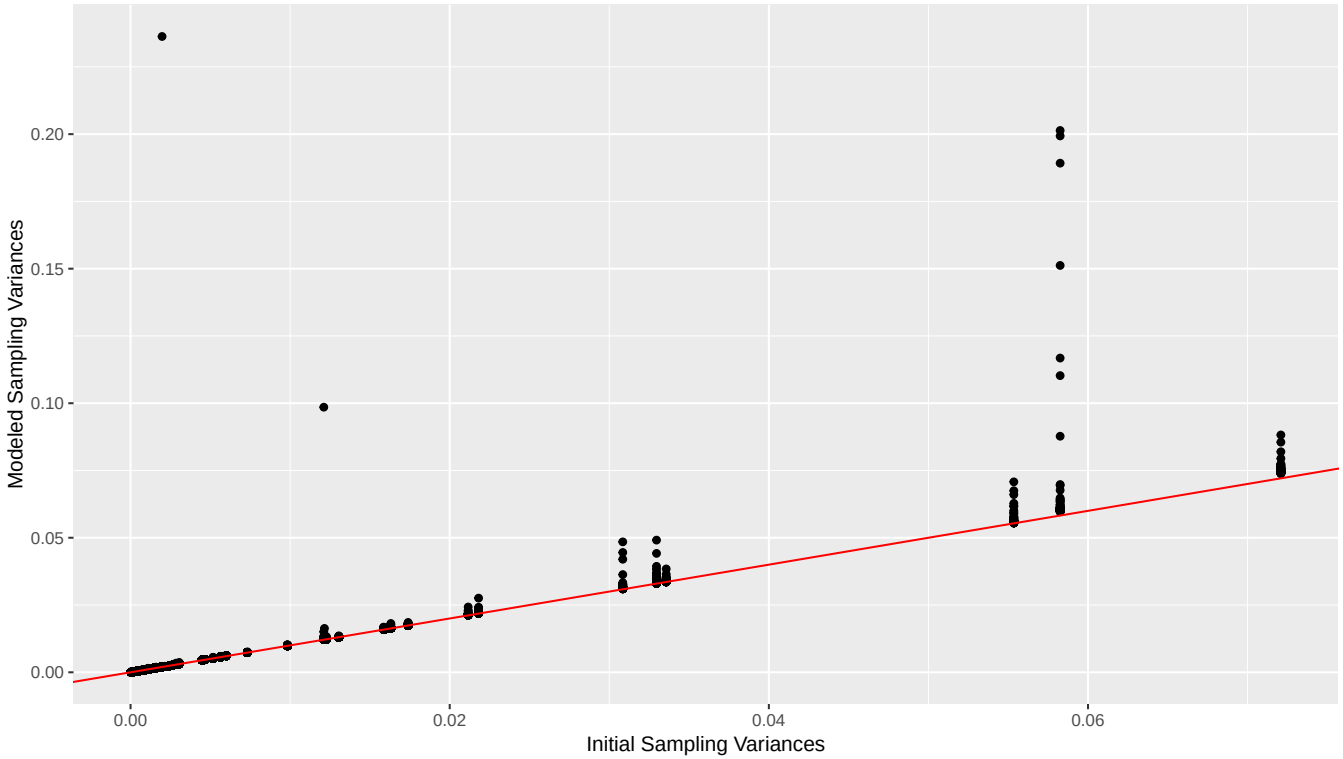


(b) Predicted

RMSE by Decile for Total Population

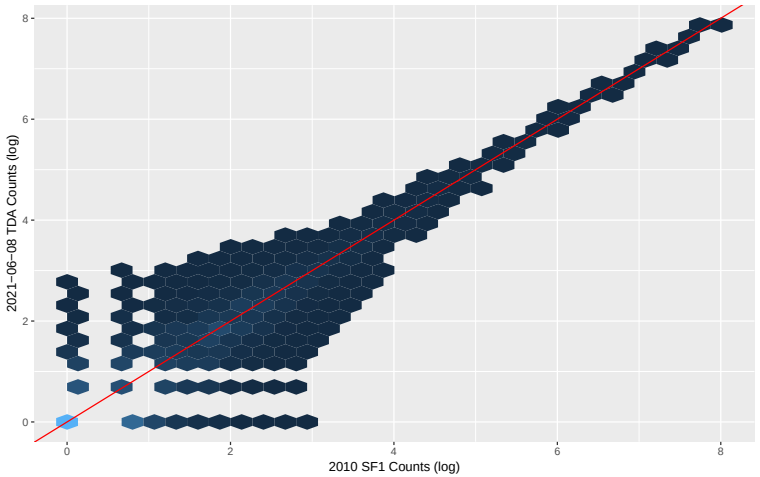


Variance Estimation for Total Population

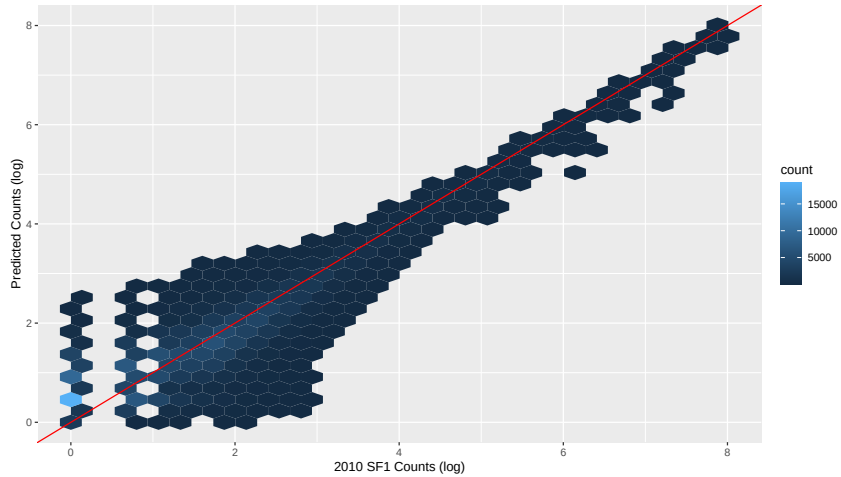


United States Census Bureau

Point Estimates for AIAN Alone Population

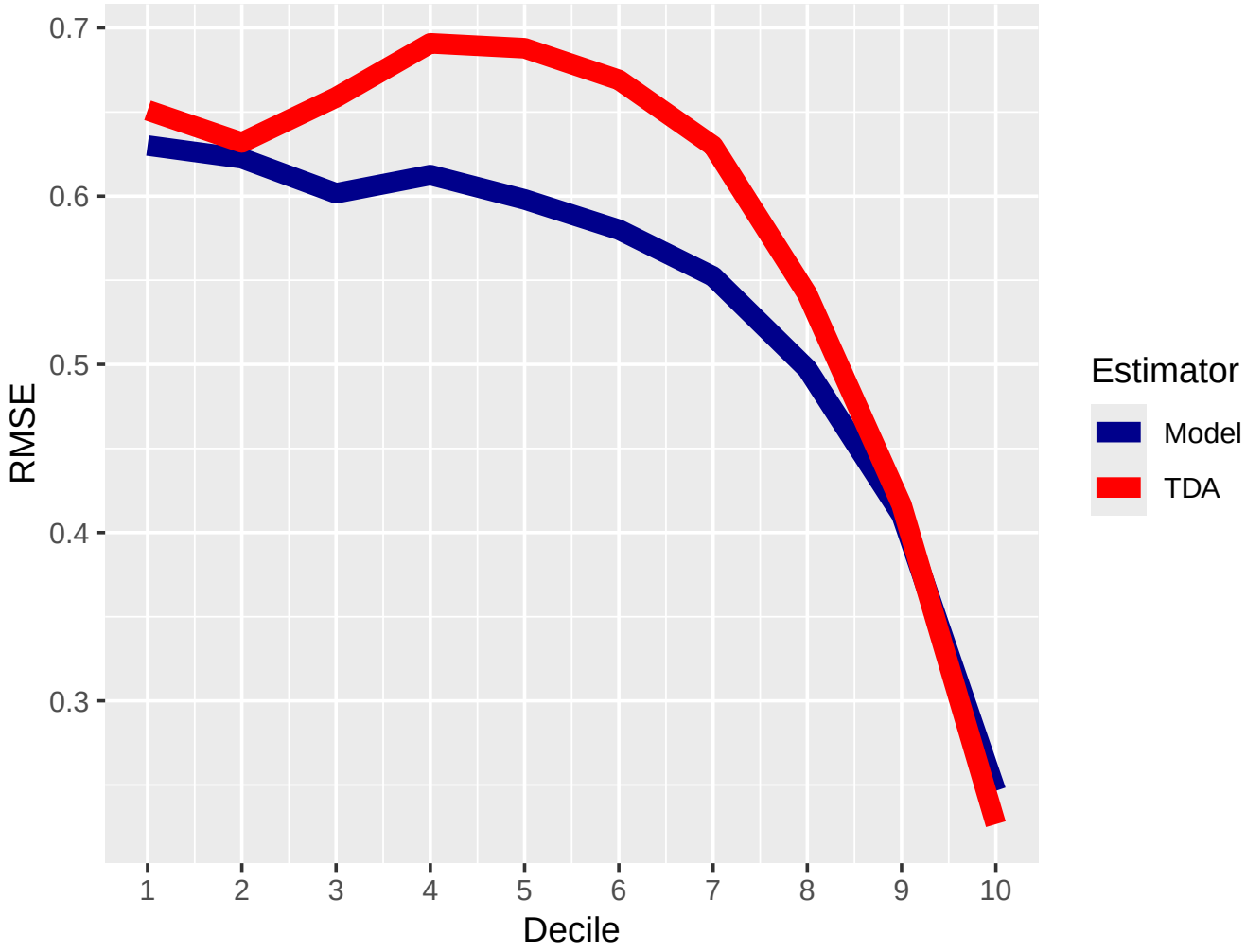


(a) TDA



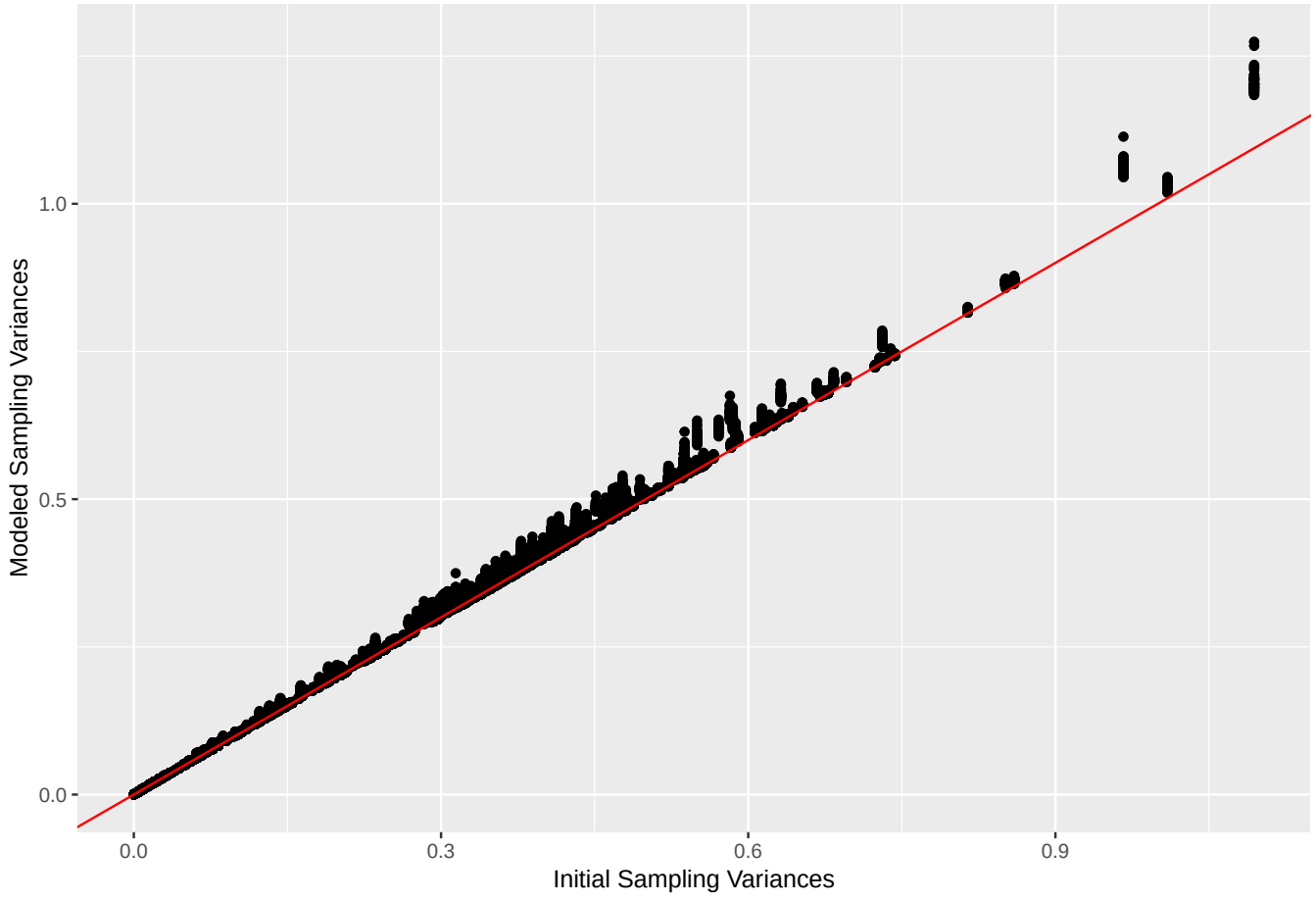
(b) Predicted

RMSE by Decile for AIAN Alone Population



United States Census Bureau

Variance Estimation for AIAN Alone



United States Census Bureau

Main message

- Though the random noise mechanism used in the implementation of differential privacy in the TDA is known, the effect of post-processing is more difficult to quantify
- Variance estimators similar to the form suggested by Esri are easy to understand, but may be over-generalized
- The use of a joint hierarchical Bayesian model can improve the estimation of the variance, while also producing model-estimated population counts
- TDA counts for the total are largely accurate, there is more room for improvement in the race subgroup estimation

United States Census Bureau

Next Steps and Limitations

- Additional estimators for the mean
- More use of the TDA constraints and published specifications
- Goodness of fit for the variance model
- Better initial estimates of the variance
  - Other estimators
  - Different methods of grouping for replication
- Apply the model to 2020 using estimated variances from the 2010 data

United States Census Bureau

Thank you!

Kyle Irimata
kyle.m.irimata@census.gov

United States Census Bureau