Estimation for Dependent Multi-type Survey Data

Zewei Kong [1] , Paul A. Parker [2] and Scott H. Holan [1,3]

[1]University of Missouri

[2]University of California, Santa Cruz

[3]US Census Bureau

August 04, 2024

This research was partially supported by the U.S. National Science Foundation (NSF). This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the NSF or U.S. Census Bureau.

Introduction

- When performing small area estimation, our focus is on efficiently utilizing multivariate spatial correlations
- Utilizing information from strongly correlated responses combined with multivariate spatial correlations can enhance the accuracy of predictions
- Selected poverty rate and median income as spatially correlated responses for each area in our study

Methodology: Univariate Spatial Model

- The Univariate Spatial Model (UNIS model), accommodating various response types and incorporating both spatial and random effects
- Univariate spatial model [Fay III and Herriot (1979)] for Gaussian Response

$$Y_i = \theta_i + \epsilon_i$$
$$\theta_i = X_i\beta + S_i\eta + \zeta_i$$

- The variance for each area is predetermined
- S represents the basis function, η is the spatial random effect, and ζ is the random effect

Methodology: Univariate Spatial Model

- the UNIS model for binomial response

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta + S_i\eta + \zeta_i$$

- Effective sample size [Chen et al. (2014)]: m_i^E for a specific proportion $\hat{P}_i$ using the formula

$$m_i^E = \frac{\hat{P}_i(1-\hat{P}_i)}{\widehat{\text{var}}(\hat{P}_i)}; y_i^E = m_i^E * \hat{P}_i$$

NOTE: to simplify notation, we let $\phi_i = X_i\beta + S_i\eta + \zeta_i$

Methodology: Polya-Gamma Data Augmentation

- Polya-Gamma Data Augmentation [Polson et al. (2013)] for logistic regression

- **Lemma 1.** Let $p(\omega)$ denote the density of the random variable $\omega \sim \text{PG}(b, 0), b > 0$. Then the following integral identity holds for all $a \in \mathbb{R}$.

$$\frac{\left(e^{\psi}\right)^{a}}{\left(1+e^{\psi}\right)^{b}} = 2^{-b} e^{k\psi} E_{\omega|\cdot}\left[\exp \frac{-\omega\psi^{2}}{2}\right],$$

  where $\kappa = a - b/2$

- **Lemma 2.** If $\omega \sim PG(b, 0)$ then $\omega|\cdot \sim PG(b, \psi)$.

Methodology: Polya-Gamma Data Augmentation

- Apply Data Augmentation for Parameter Updating (Using β as an example),

$$L(\beta) \propto \frac{[\exp(\phi_i)]^{\sum y_i}}{[1+\exp(\phi_i)]^N}$$

- By Lemma 1 & 2.

$$\omega_i|\beta \sim PG(m_i^E, \phi_i)$$
$$\beta|\Omega, y \propto \mathcal{N}(\boldsymbol{\mu_\beta}, \Sigma_\beta)$$

Methodology: Multi-type Spatial Model

- We proposed the Multi-type Spatial Model (MUTS Model):

$$
\begin{aligned}
\boldsymbol{Z_1}|\cdot &\sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma_\epsilon}) \\
\boldsymbol{Z_1} &= \boldsymbol{\theta} + \boldsymbol{\epsilon} \\
\boldsymbol{\theta} &= \boldsymbol{X}_1\beta_1 + \tau_1\boldsymbol{S}\boldsymbol{\eta} + \zeta_1 \\
\boldsymbol{Z_2}|\cdot &\sim \mathcal{B}(\boldsymbol{m}^*, \boldsymbol{\pi}) \\
logit(\boldsymbol{\pi}) &= \boldsymbol{X}_2\beta_2 + \tau_2\boldsymbol{S}\boldsymbol{\eta} + \tau_3\boldsymbol{S}\boldsymbol{\kappa} + \zeta_2
\end{aligned}
$$

- Incorporate $\boldsymbol{\eta}$ as a spatial random variable shared between two responses
- $\boldsymbol{S}$ represents spatial basis function and τ is a scale parameter
- κ is a spatial random effect, ζ_1 and ζ_2 is random effect

Simulation Study

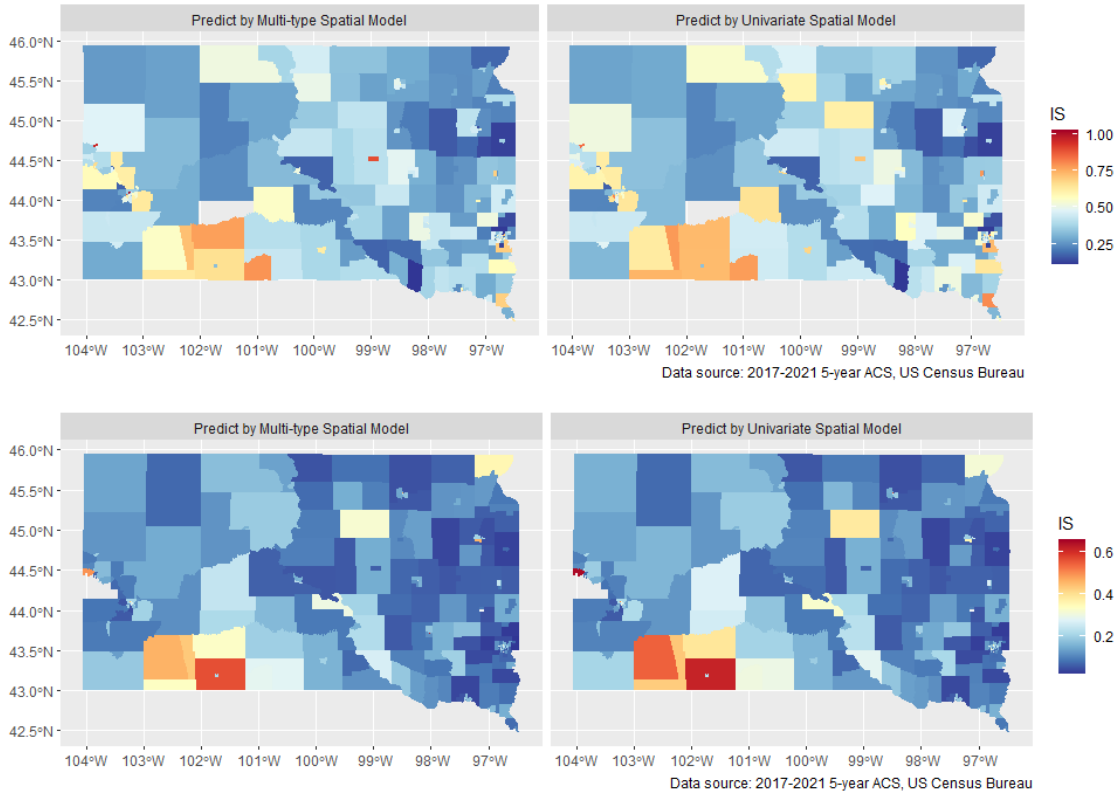
- South Dakota was selected for the empirical simulation study, 241 tracts used
- Covariates choice: *the proportion of white people* and *the proportion of bachelor degrees*
- Used the Census Bureau's direct estimates as true values, generated 100 datasets, and performed estimations
- For binomial responses, we compared the MSE of logit(p)

Simulation Study: Interval Score

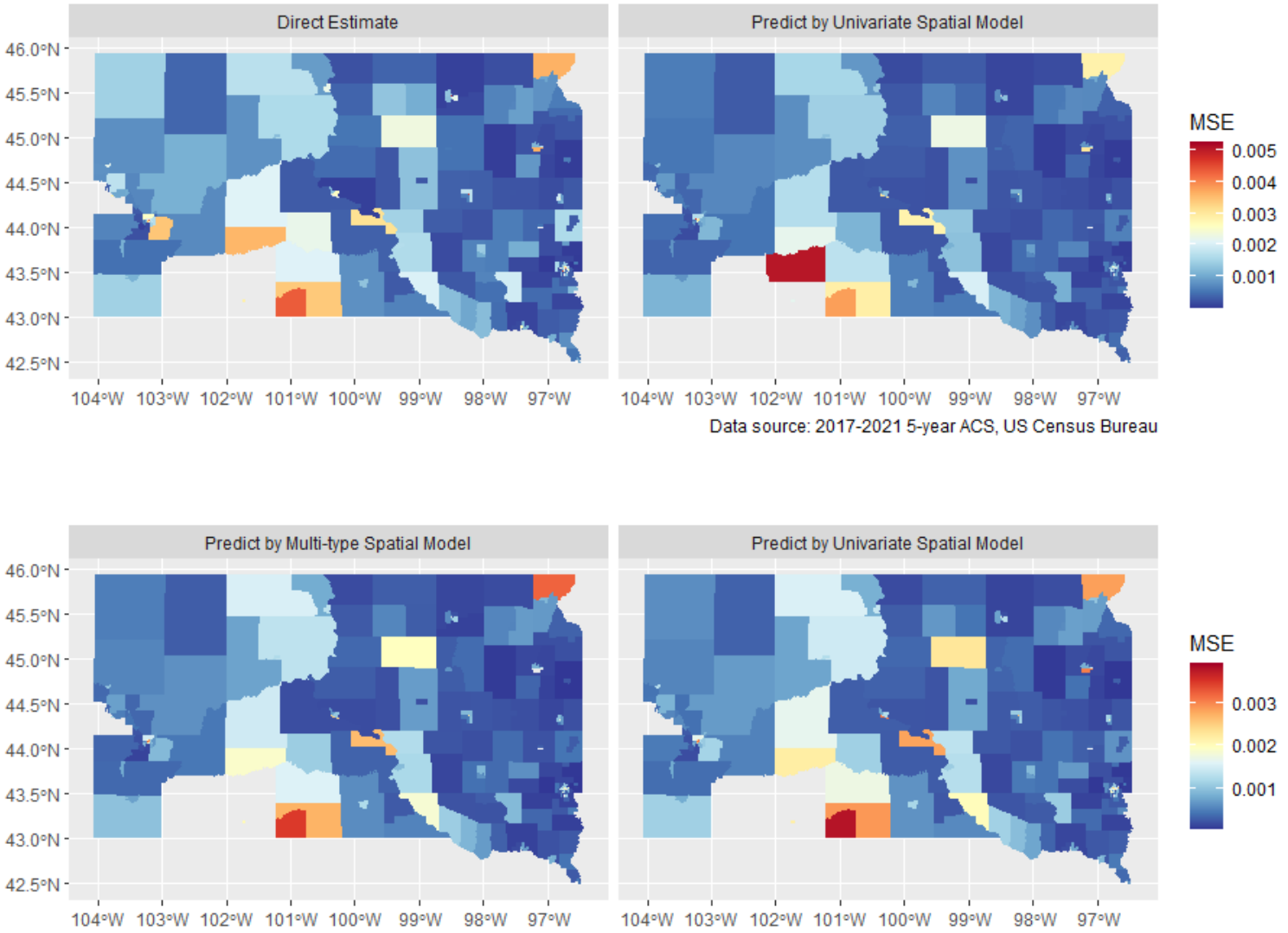
- Interval score [Gneiting and Raftery (2007)]:

$$S_{\text{int}}^{\alpha}(l, u; x) = (u - l) + \frac{2}{\alpha}(l - x)1_{\{x<l\}} + \frac{2}{\alpha}(x - u)1_{\{x>u\}}$$

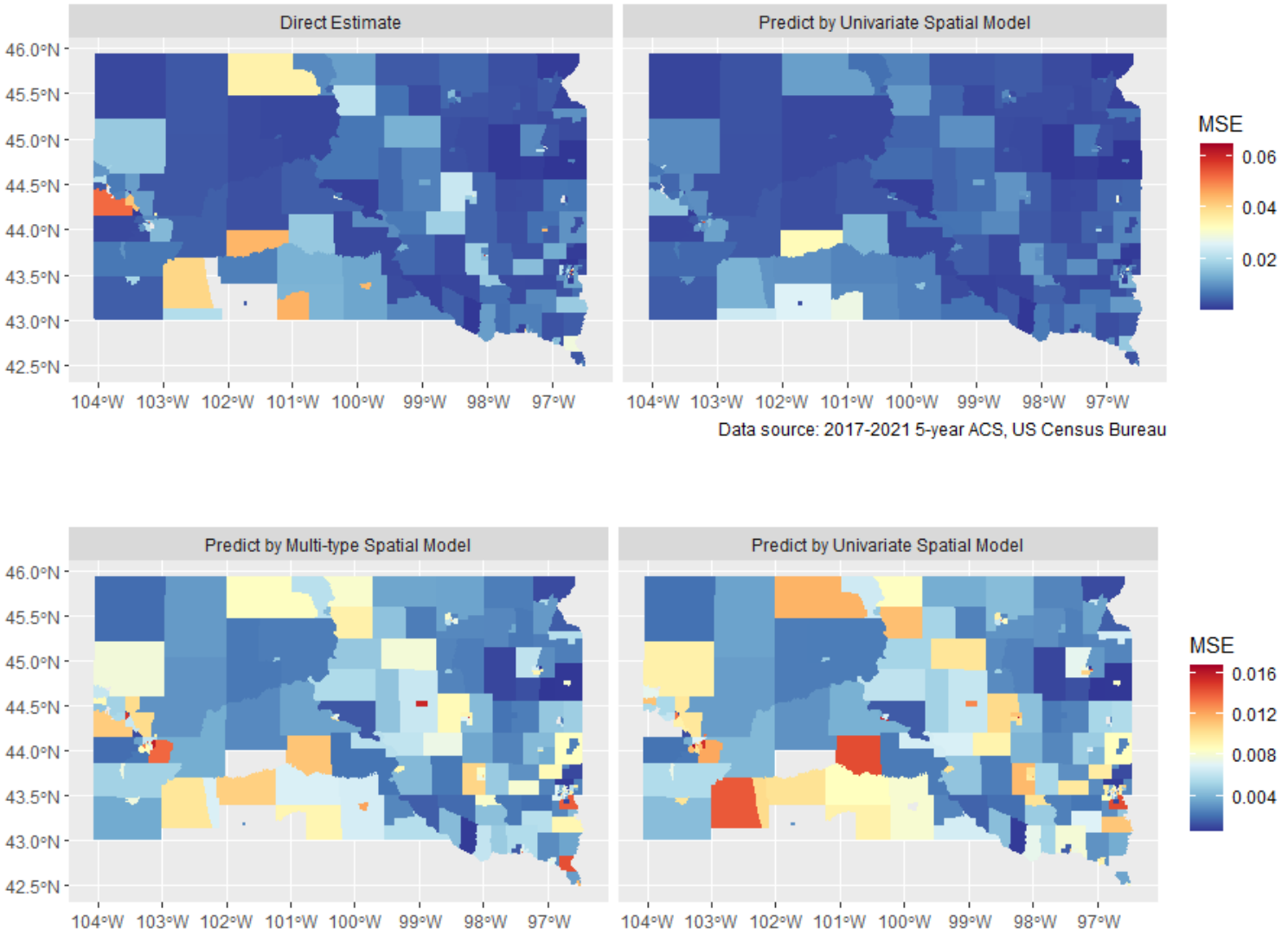
- Spatial plot for interval score



Simulation Study: MSE for Gaussian Response Comparison



Simulation Study: MSE for Binomial Response Comparison



Simulation Study: Results

- Empirical simulation study results for Gaussian response

Type	MSE	Coverage	IS	MSE Red (%)
Direct Estimate	0.0152	-	-	-
UNIS Model	0.0109	94.1%	0.550	**27.90%**
MUTS Model	0.0094	95.0%	0.498	**38.15%**

- Empirical simulation study results for Binomial response

Type	MSE	Coverage	IS	MSE Red (%)
Direct Estimate	0.00160	-	-	-
UNIS Model	0.00113	93.8%	0.170	**29.40%**
MUTS Model	0.00102	94.6%	0.157	**36.42%**

- MSE reduction between MUTS and UNIS model:**14.21%** and **10.0%**.

Census Region Level Analysis

- The West North Central Division in the Midwest Region for data analysis with a total of 5,834 tracts
- Variance and variance reduction for Gaussian Response

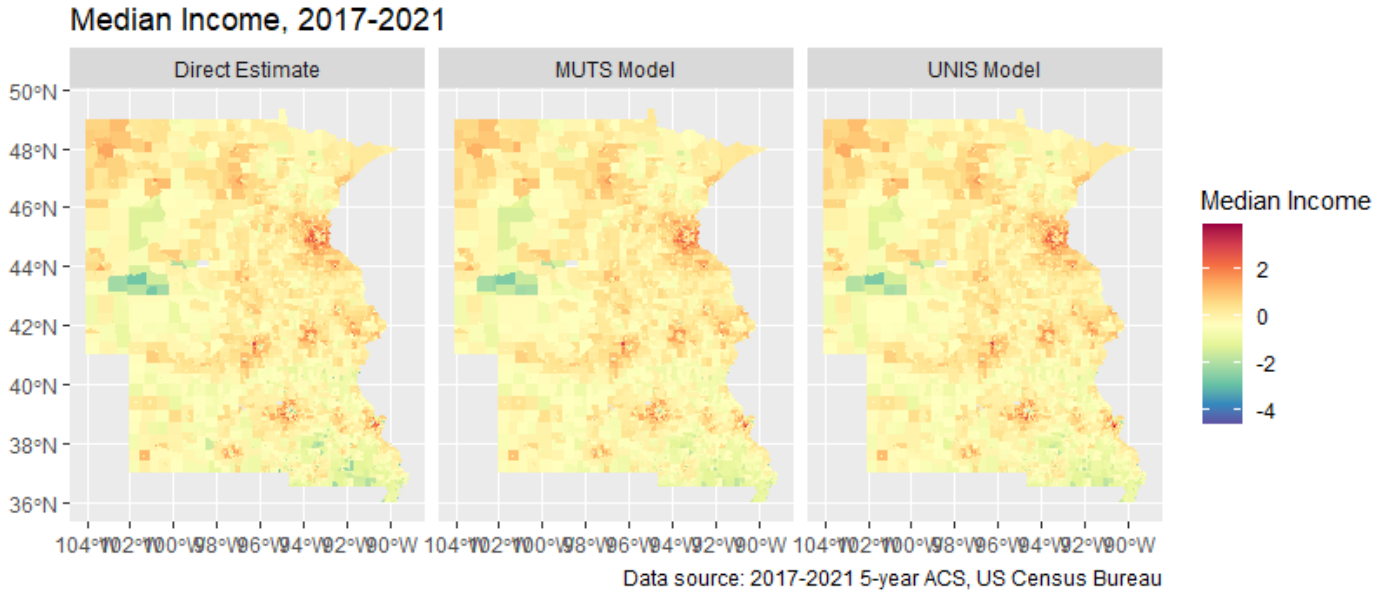
Type	Avg Var	Var Red (%)
Direct Estimate	0.121	-
UNIS Model	0.083	**30.93%**
MUTS Model	0.065	**45.79%**

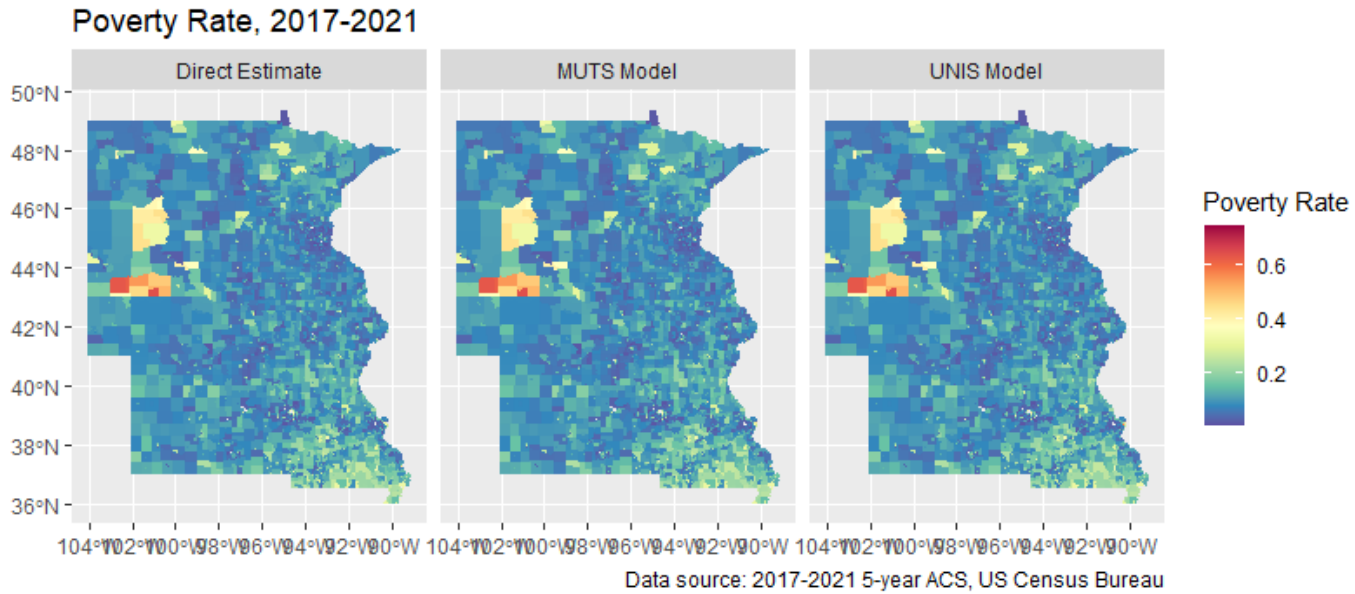
- Variance and variance reduction for Binomial Response

Type	Avg Var	Var Red (%)
Direct Estimate	0.00169	-
UNIS Model	0.00115	**32.31%**
MUTS Model	0.00095	**43.66%**

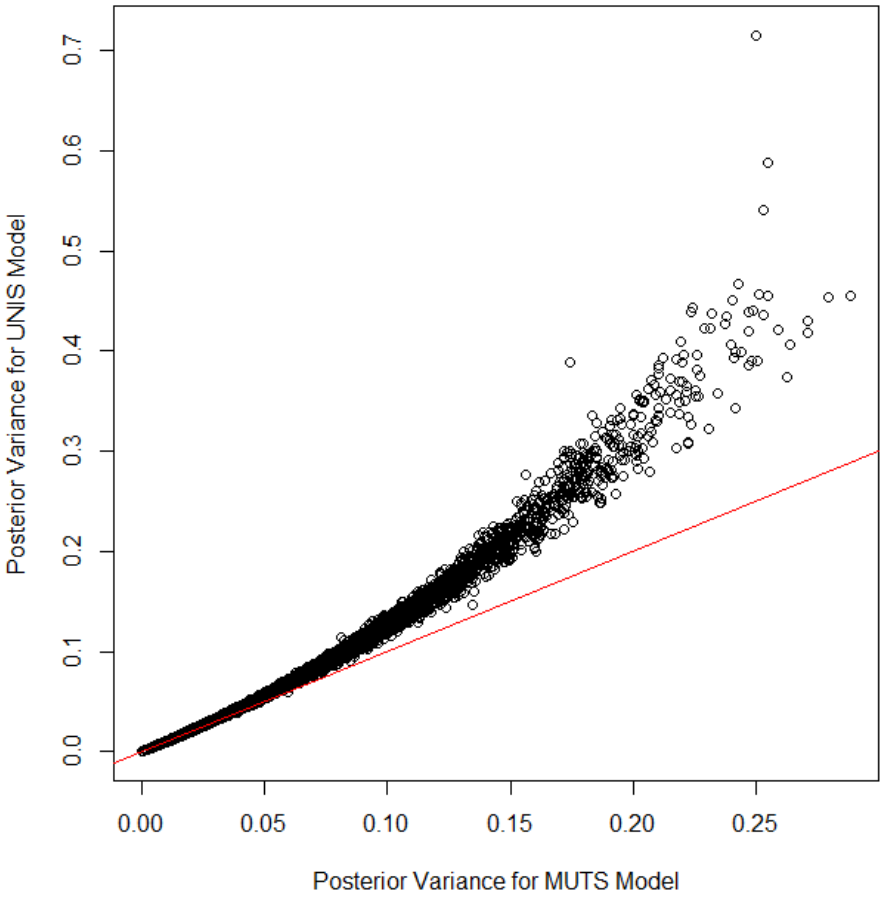
- Variance reduction between MUTS and UNIS model:**21.51%** and **16.77%**.

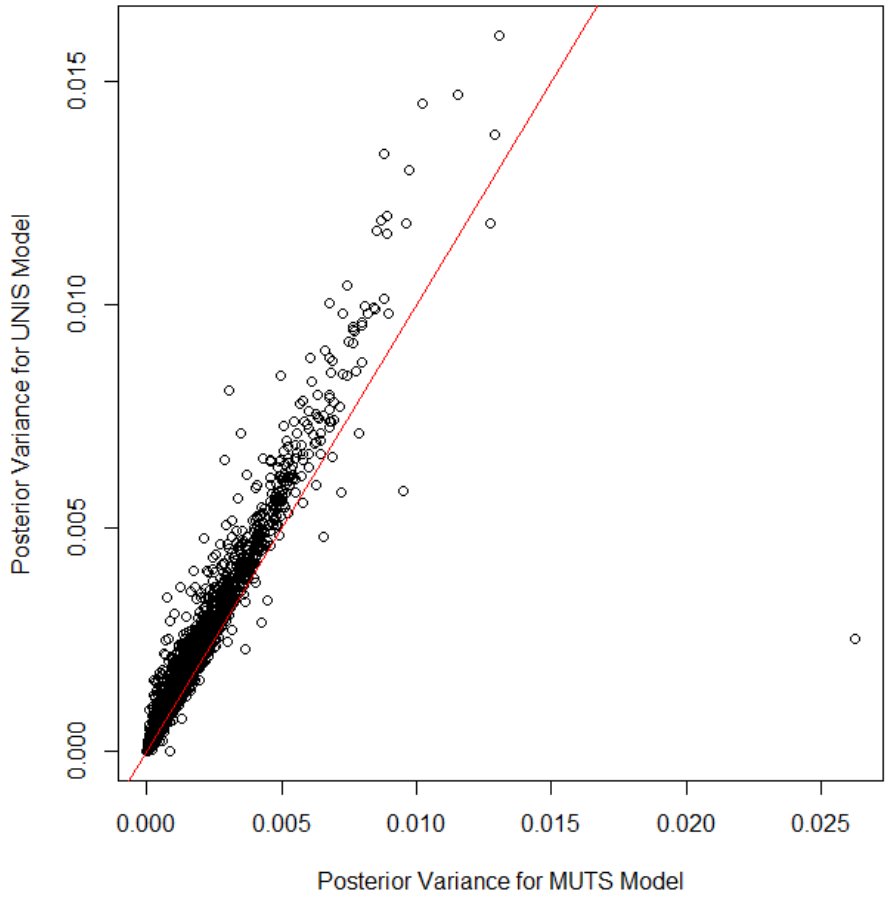
Census Region Level Analysis





Census Region Level Analysis





Discussion

- The MUTS model outperforms both direct estimate and the UNIS model in predictive accuracy.
- Data analysis at the regional level shows a significant improvement in variance.
- The choice of basis functions can effectively reduce the dimensionality of spatial information.
- In the future, we plan to extend the model to the unit level.

References

Chen, C., Wakefield, J., and Lumely, T. (2014). The use of sampling weights in bayesian hierarchical models for small area estimation. *Spatial and spatio-temporal epidemiology*, 11:33–43.

Fay III, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.

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

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.