Bayesian Nonparametric Multivariate Spatial Mixture Mixed Effects Models

Scott H. Holan[1,2], Ryan Janicki[2], Andrew M. Raim[2], and Jerry Maples[2]

Department of Statistics
University of Missouri[1]
and
U.S. Census Bureau[2]

October 27, 2022

Outline

- Overview of the project goals, description of the datasets, and discussion of difficulties with the data
- Introduction of the multivariate spatial model
  - Summary of model-based predictions of counts within counties and comparison to direct estimates
  - Discussion of how the model can fail
- Extension of the multivariate spatial model to a mixture of multivariate spatial models
- Conclusion

American Community Survey (ACS)

- Nationwide survey collecting information about, for example, race, citizenship, employment, and other demographic and housing unit characteristics
- 3.5 million households sampled each year
- Billions of estimates produced
- 1-year estimates available for areas with a total population of 65,000+
- 5-year estimates available at geographies down to the block group

Examples of Requested ACS Special Tabulations

- Household income to poverty ratio
- Proportion of the population in poverty
- Total population by age, race/ethnicity, sex
- Counts of housing units by Tenure by Household Size by Age of Householder by Household Income

These tables are needed at different geographic levels, such as by State, County, Tract, or American Indian and Alaska Native (AIAN) area.

ACS Tabulations Used

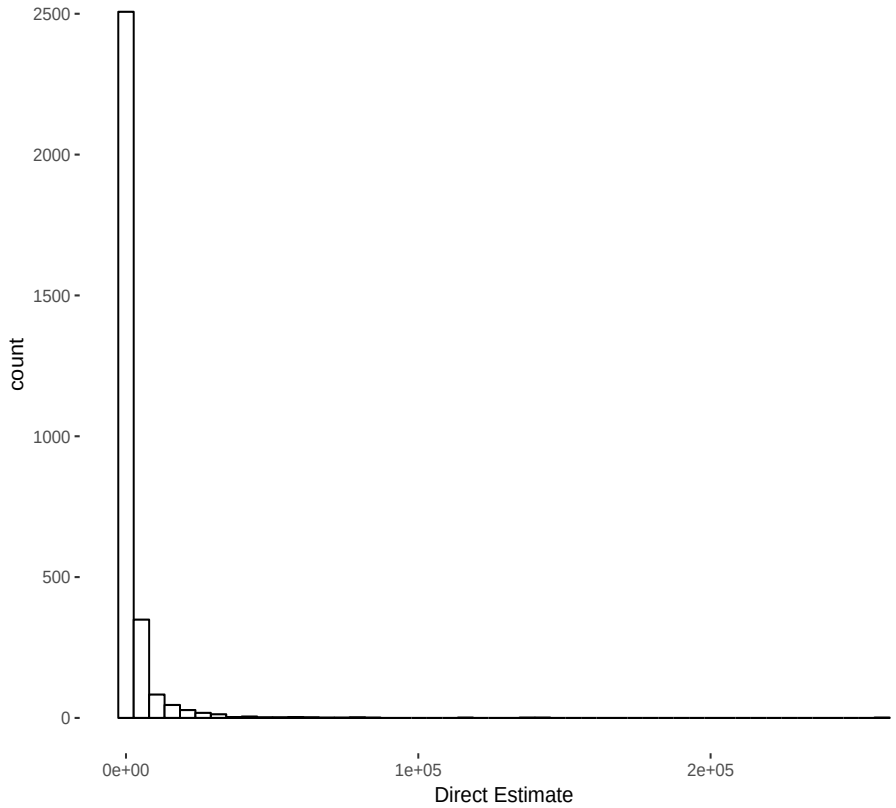
- Dataset 1 (AGE): Counts of children by age (0–1, 2–3, and 4–5) in counties in Minnesota
- Dataset 2 (AGE x RACE): Counts of children by age (0–1, 2–3, and 4–5) by Race (White Alone, Black Alone, Asian Alone, Native Hawaiian or Pacific Islander Alone, American Indian and Alaska Native Alone, Other Alone, or two or more races) in counties in Minnesota

Issues with the Data

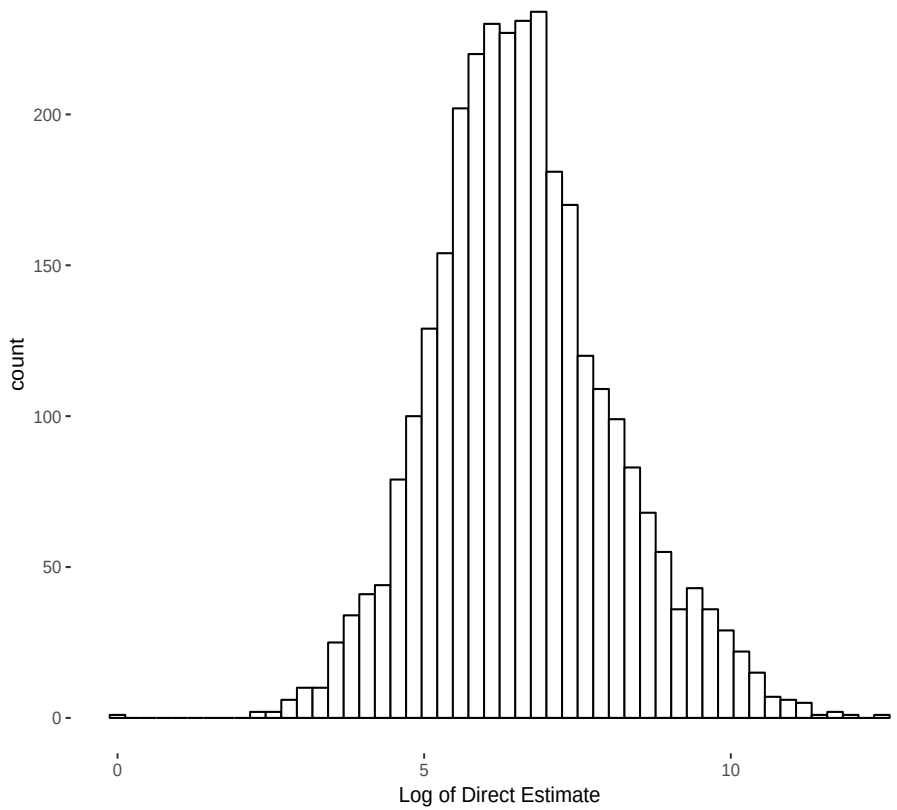
- Small or zero sample in some counties for some cells
- Skewed or multimodal distributions of direct estimates
- A large number of cells with direct estimates of zero
- Some sampling variances are either zero or undefined
- Multivariate dependence between the observations within county and multivariate spatial dependence between counties
- Limited predictor variables available

Data

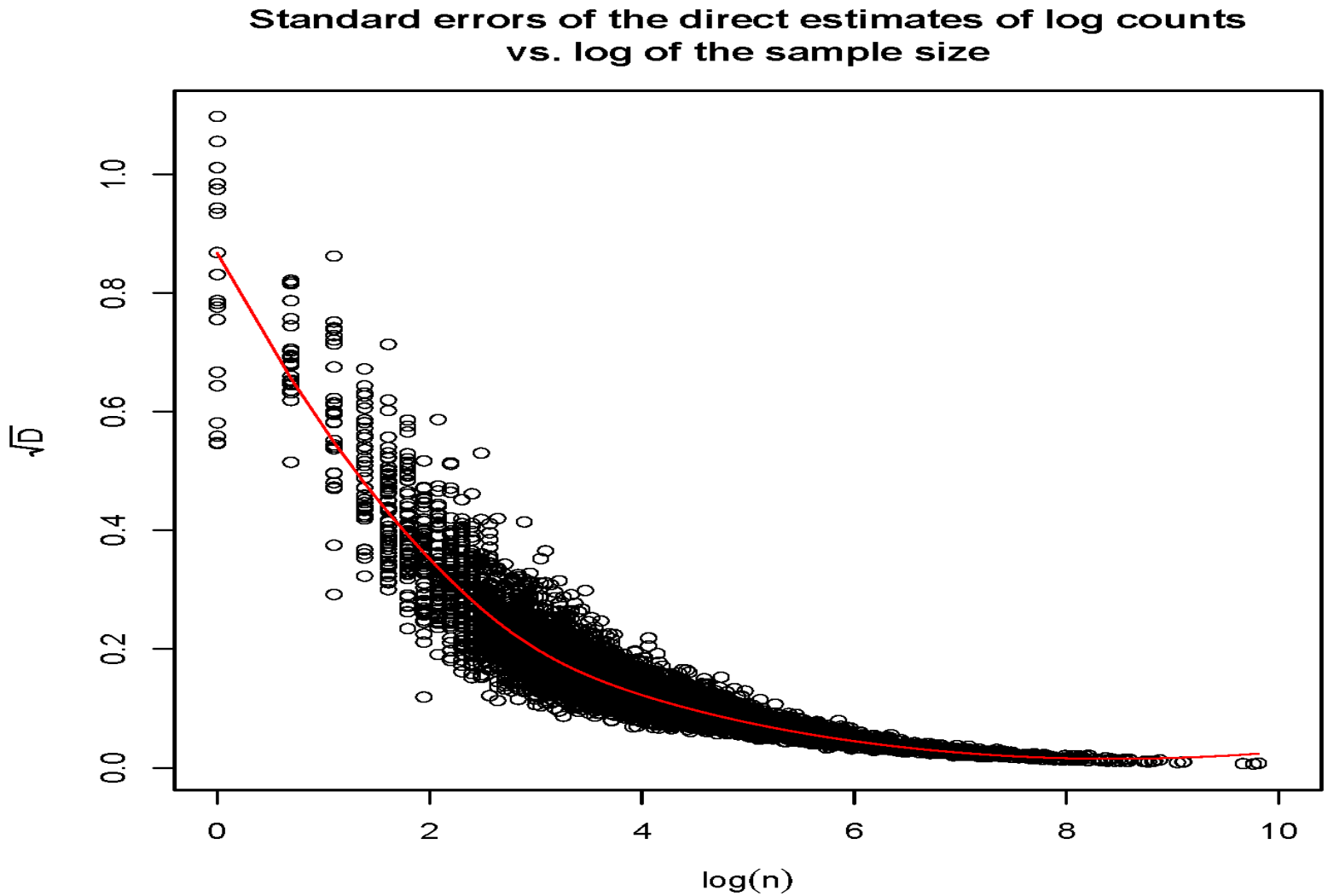
Histogram of the ACS 5 year estimates of total children ages 4 − 5 in all counties in the U.S.



Histogram of the log of the ACS 5 year estimates of total children ages 4 − 5 in all counties in the U.S.



Direct Variance Estimates



Notation

- Z_i^* is the direct estimate for the ith observation. For ease of presentation, i indexes the counties and demographic variables.
- $Z_i = \log(Z_i^* + 1)$
- $\boldsymbol{x}_i^T$ a vector of predictors, which includes the log of the county population size, and dummy variables corresponding to the demographic cross classifications
- $D_i = Var(Z_i)$ are the estimated (design-based) sampling variances

Multivariate Spatial Model

Cressie and Wikle (2011); Bradley et al. (2015):

- Data Model:

$$Z_i = Y_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, D_i)$$

- Process Model:

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{s}_i^T \boldsymbol{\eta}$$
$$\boldsymbol{\eta} \sim N_r\left(0, \sigma^2 \boldsymbol{K}\right)$$

- Goal: Prediction of $Y_i^* = e^{Y_i} - 1$

Remark About the Choice of Process Model

A popular choice of process model used in the spatial literature is

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \boldsymbol{s}_i^T \boldsymbol{\eta} + \xi_i,$$

where

- $\boldsymbol{x}_i^T \boldsymbol{\beta}$ captures trends in population and demographic cross classifications, as well as multivariate spatial dependencies
- $\boldsymbol{s}_i^T \boldsymbol{\eta}$ accounts for residual multivariate spatial dependence (random effects)
- ξ_i represents fine-scale variability

Note that including ξ_i resulted in overfitting to our datasets

Basis Functions

The matrix $\boldsymbol{S}$, with rows consisting of the vectors $\boldsymbol{s}_i^T$, are constructed via Moran's I basis expansion (Hughes and Haran, 2013):

- Let $\boldsymbol{P_X} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- Let $\boldsymbol{A}$ be the adjacency matrix of observations
- The columns of $\boldsymbol{S}$ are the first r eigenvectors of $\boldsymbol{P_X A P_X}$
- We used 50% of the available basis functions: eigenvectors corresponding to the largest positive eigenvalues

Construction of $\boldsymbol{K}$

- The matrix $\boldsymbol{K}$ is constructed to
  - induce multivariate dependencies
  - induce spatial dependencies using a conditional autoregressive (CAR) structure
  - be reduced rank compared to a CAR process
- Random effects may be shared across observations, inducing multivariate dependencies

Construction of $\boldsymbol{K}$ Cont.

Let $\boldsymbol{u}^{\boldsymbol{T}} = (u_1, \ldots, u_n)$ be an intrinsic conditional autoregressive (ICAR) process, with precision matrix $\frac{1}{\sigma^2}\boldsymbol{Q}$, so that

$$u_i | u_j, j \neq i, \sigma^2 \sim N\left(\sum_{j \sim i} \frac{j}{n_i}, \frac{\sigma^2}{n_i}\right),$$

where n_i is the number of neighbors of area i. Then

$$\boldsymbol{K} = \arg\min_{\boldsymbol{C}} \left\| \boldsymbol{Q} - \boldsymbol{S}\boldsymbol{C}^{-1}\boldsymbol{S}^{\boldsymbol{T}} \right\|_F,$$

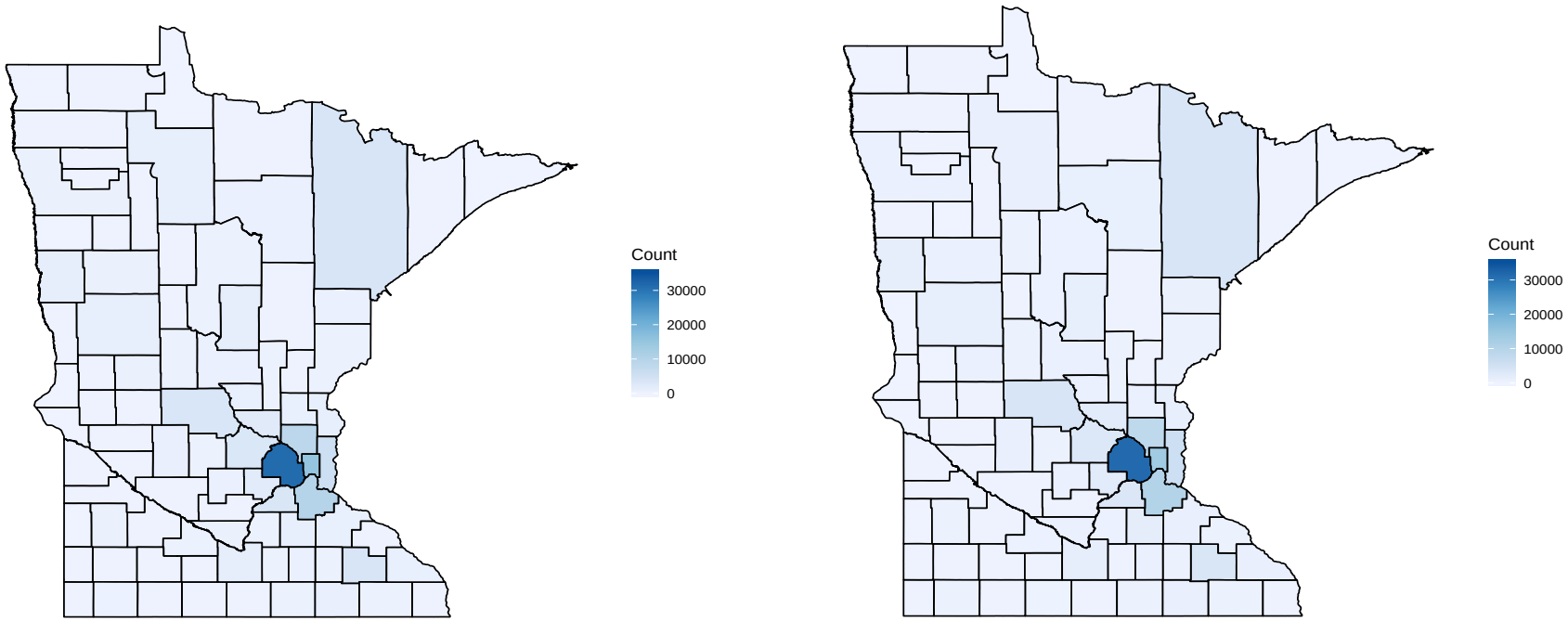
where the minimization is over the space of $r \times r$ positive definite matrices.

Estimation of the number of children by age in counties in Minnesota using the multivariate spatial model

Prediction of the Number of Children, Ages 0 – 1, 2 – 3, 4 – 5, in Counties

- Our target is predictions of counts in counties in Minnesota.
- We fit the model to counties in Minnesota and adjacent states (Iowa, North Dakota, South Dakota, Wisconsin)
- We wrote code in R and STAN (Carpenter et al., 2017) to fit the multivariate spatial mixture model to each dataset

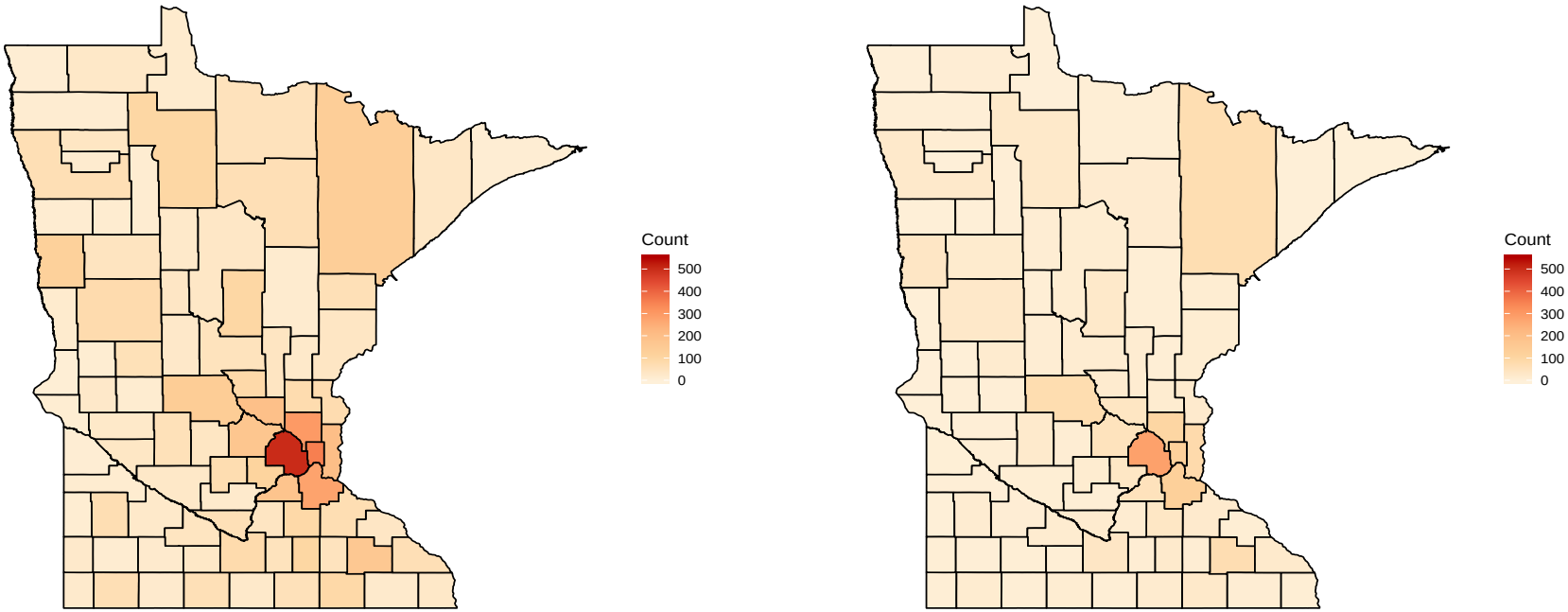
Comparison of Direct Estimates and Predicted Values



(a) (b)

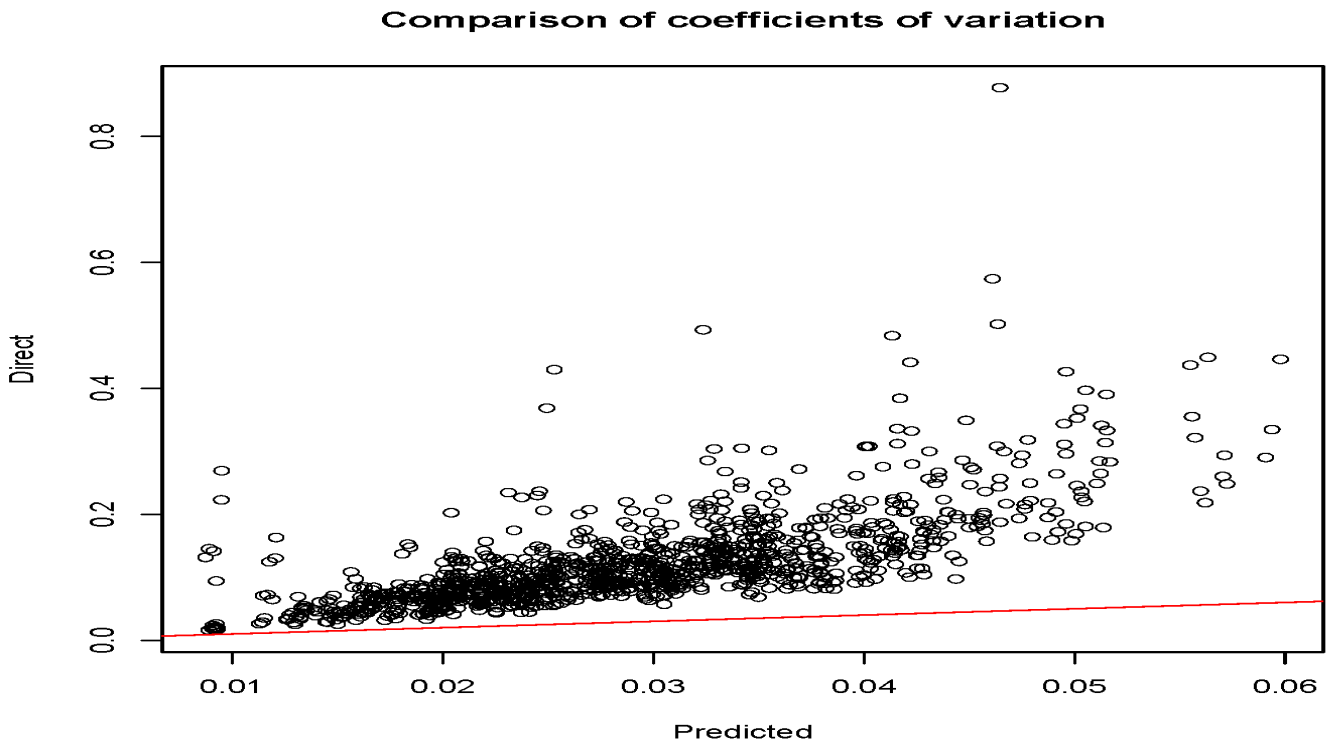
Figure: (a) Direct estimates of the number of children, ages 0–1.
(b) Model-based predictions of the number of children, ages 0–1.

Comparison of the Standard Errors of Direct Estimates and Standard Errors of Predicted Values



(a) (b)

Figure: (a) Standard errors of direct estimates of the number of children, ages 0–1. (b) Posterior standard error model-based predictions of the number of children, ages 0–1.



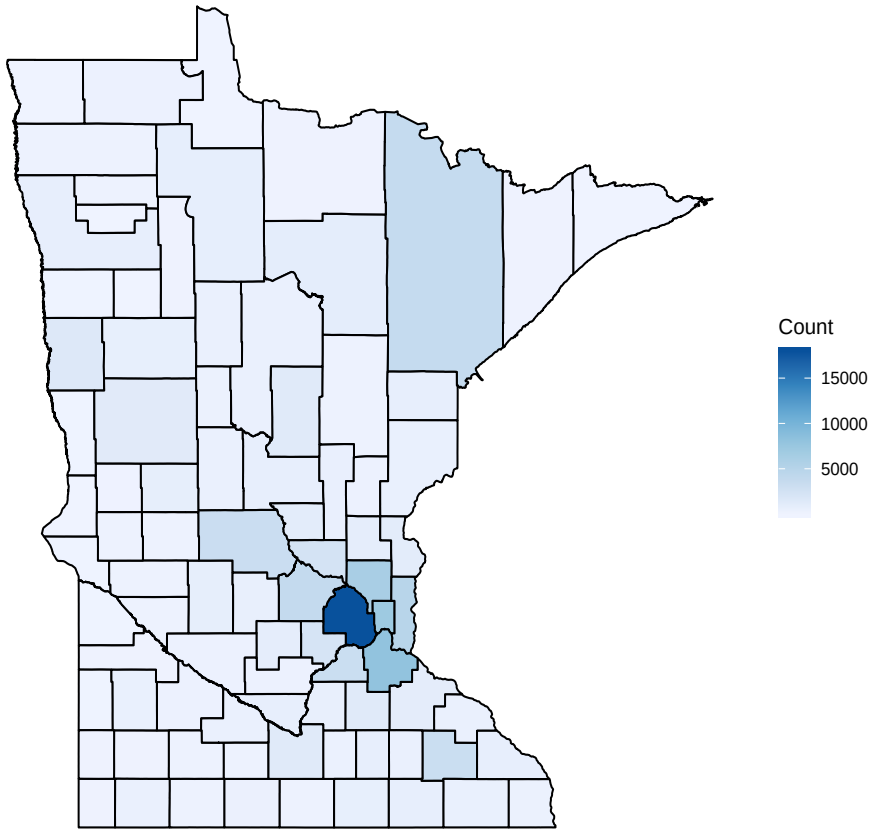
Min	Q1	Median	Q3	Max
96	78	74	68	41

Table: Percent reduction in coefficients of variation

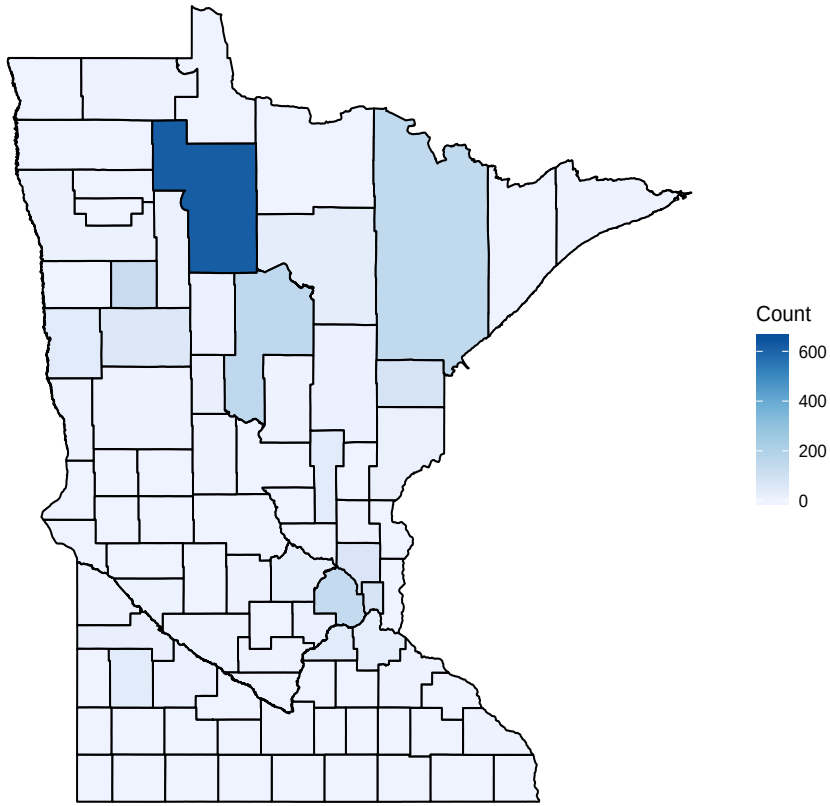
Estimation of the number of children by age and race in counties in Minnesota using the multivariate spatial model

Example of Varying Spatial Patterns in the Data

Direct estimates of the number of White children, ages 2 − 3, in counties in Minnesota



Direct estimates of the number of American Indian or Alaska Native children, ages 2 − 3, in counties in Minnesota



- Great results fitting the multivariate spatial model to a wide variety of ACS special tabulations
- Can have poor results when fitting to more complicated datasets with more demographic cross classifications
- Issues:
  - Varying spatial patterns for different Race/Ethnicity groups.
  - The spatial model can be too aggressive in smoothing the data
- Possible solutions:
  - Manually break up difficult datasets if we know the clustering structure, and fit separate models on each cluster of data
  - Use a model-based approach to simultaneously cluster the data, and predict counts

Multivariate Spatial Model with Dirichlet Process Mixing

- Data Model:

$$Z_i = Y_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, D_i)$$

- Process Model:

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_i + \boldsymbol{s}_i^T \boldsymbol{\eta}_i$$
$$\boldsymbol{\theta}_i^T = \left(\boldsymbol{\beta}_i^T, \boldsymbol{\eta}_i^T\right) \mid G \overset{i.i.d.}{\sim} G$$
$$G \sim DP\left(\alpha G_0\right)$$

- α is a concentration parameter and G_0 the base measure

Multivariate Spatial Model with Dirichlet Process Mixing Cont.

- Parameter Model:

$$
\begin{aligned}
G_0 &= N_p(0, \sigma^2_\beta I_{p\times p}) N_r(0, \sigma^2_\eta \boldsymbol{K}) \\
\alpha &\sim Gamma(a_\alpha, b_\alpha) \\
\sigma^2_\beta &\sim InvGamma(a_\beta, b_\beta) \\
\sigma^2_\eta &\sim InvGamma(a_\eta, b_\eta)
\end{aligned}
$$

- With this parameterization we have conjugacy. However, the Gibbs sampler can be slow to converge.

Stick-breaking Representation of the Dirichlet Process Prior

Sethuraman (1994):

$$G(\cdot) = \sum_{j=1}^{\infty} \pi_j \delta_{\boldsymbol{\theta}_j}(\cdot)$$

$$\boldsymbol{\theta}_j \overset{i.i.d.}{\sim} G_0$$

$$V_i \overset{i.i.d.}{\sim} Beta(1, \alpha)$$

$$\pi_1 = V_1$$

$$\pi_i = V_i \prod_{k \leq i} (1 - V_k)$$

Implementation: Gibbs sampler in R/C++ (Neal, 2000)

Multivariate Spatial Mixture Model – Finite-dimensional Approximation

We can truncate G to get a finite-dimensional approximation

- Data Model:

$$Z_i \sim \sum_{j=1}^{K} \pi_j f(z_i \mid Y_{ij}, D_i),$$

  where $f(\cdot \mid Y_{ij}, D_i)$ is the normal density with mean Y_{ij} and variance D_i.

- Process Model:

$$Y_{ij} = \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \boldsymbol{s}_i^T \boldsymbol{\eta}_j$$
$$\boldsymbol{\eta}_j \sim N_r\left(0, \sigma_j^2 \boldsymbol{K}\right)$$

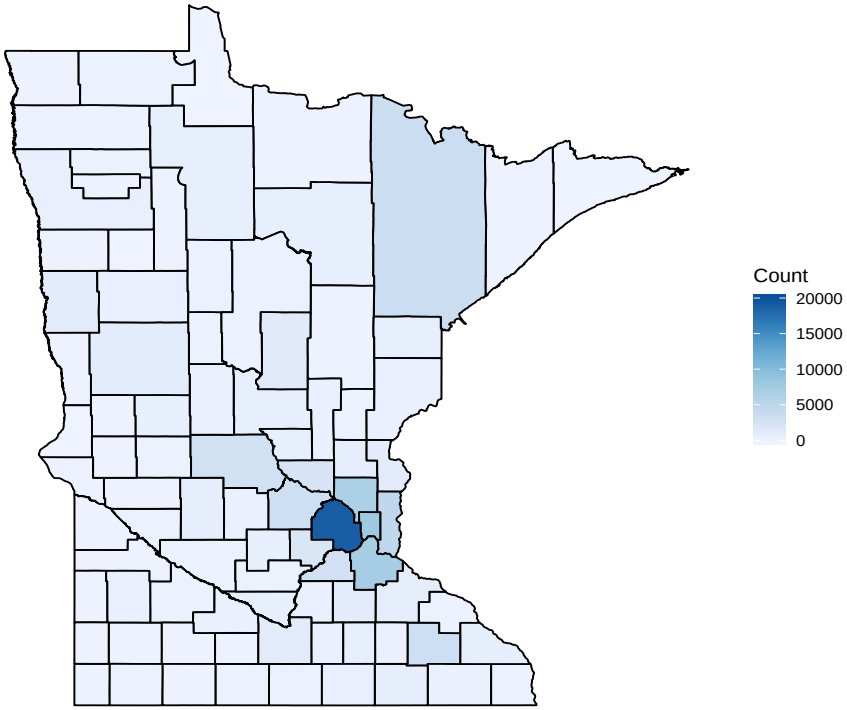
Multivariate Spatial Mixture Model Cont.

- π_j are the probabilities of belonging to cluster j
- The number of clusters is not random
- If the truncation level is sufficiently large, many clusters will be empty
- Nearly identical results using a Dirichlet process prior or a stick breaking prior, as long as the truncation is sufficiently large.
- Faster convergence using the stick breaking prior.

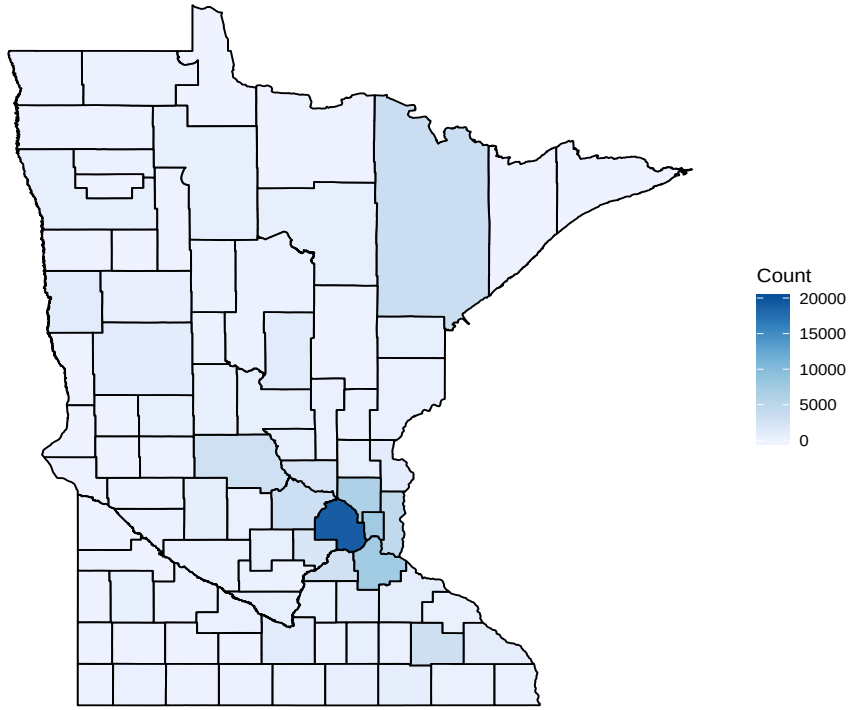
Estimation of the number of children by age and race using the multivariate spatial mixture model

Comparison of Direct Estimates and Predicted Values of the Number of White Children, Ages 2 – 3

Direct estimates of the number of White children, ages 2 − 3, in counties in Minnesota

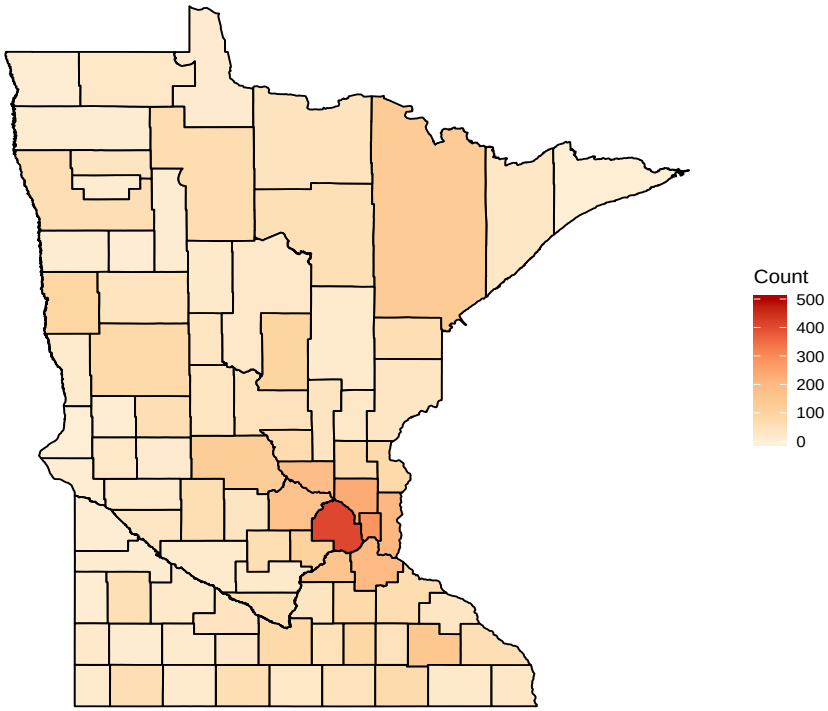


Model−based predictions of the number of White children, ages 2 − 3, in counties in Minnesota

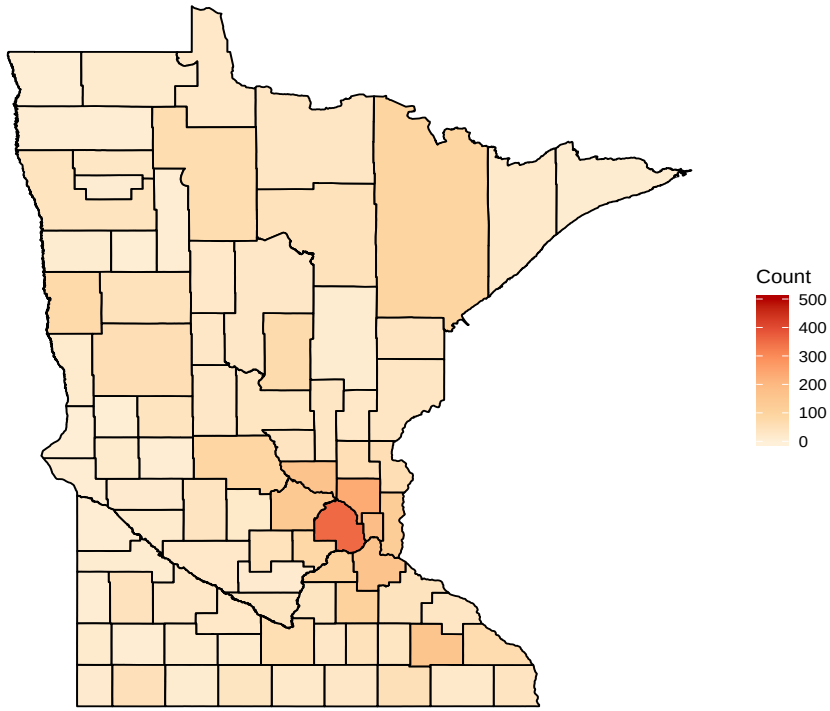


Comparison of the Standard Errors of Direct Estimates and Standard Errors of Predicted Values

Standard errors of direct estimates of the number of White children, ages 2 − 3, in counties in Minnesota

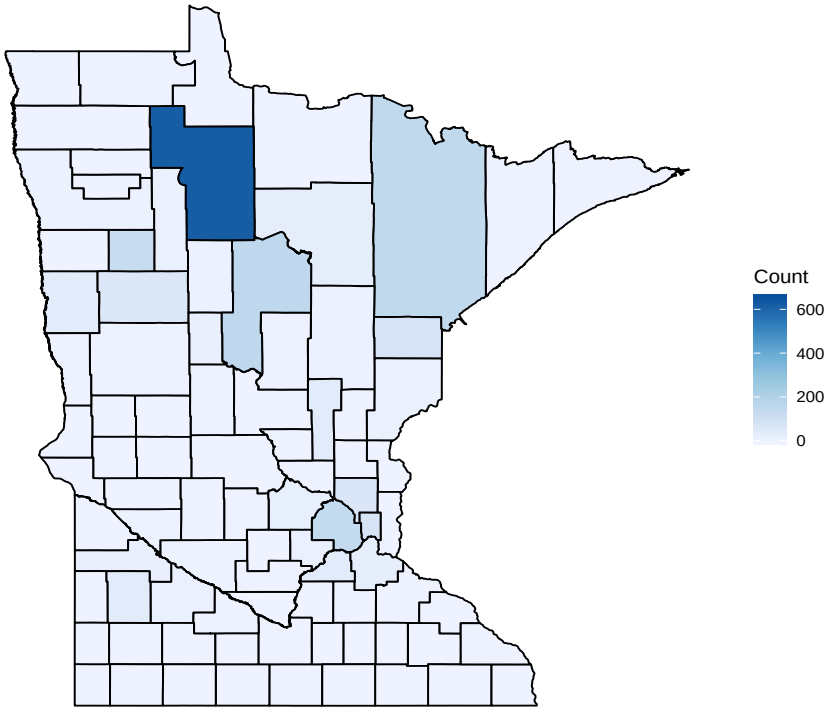


Standard errors of model−based predictions of the number of White children, ages 2 − 3, in counties in Minnesota

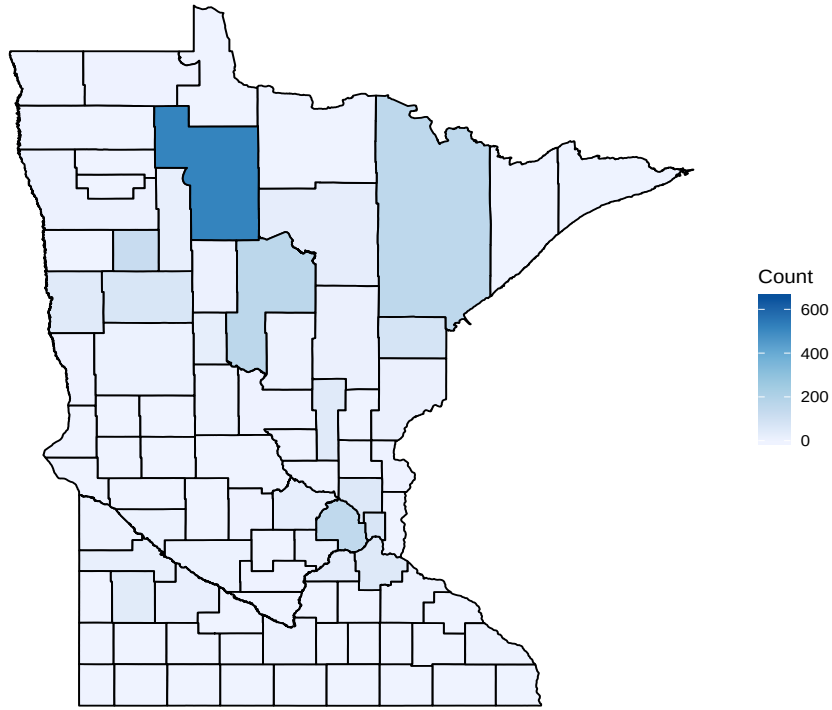


Comparison of Direct Estimates and Predicted Values of the Number of American Indian and Alaska Native Children, Ages 2 – 3

Direct estimates of the number of American Indian or Alaska Native children, ages 2 − 3, in counties in Minnesota

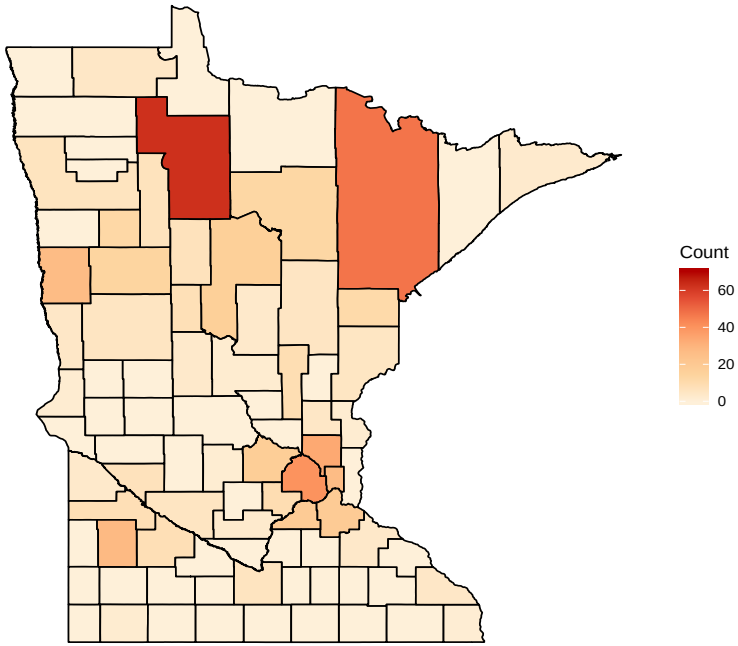


Model−based predictions of the number of American Indian or Alaska Native children, ages 2 − 3, in counties in Minnesota

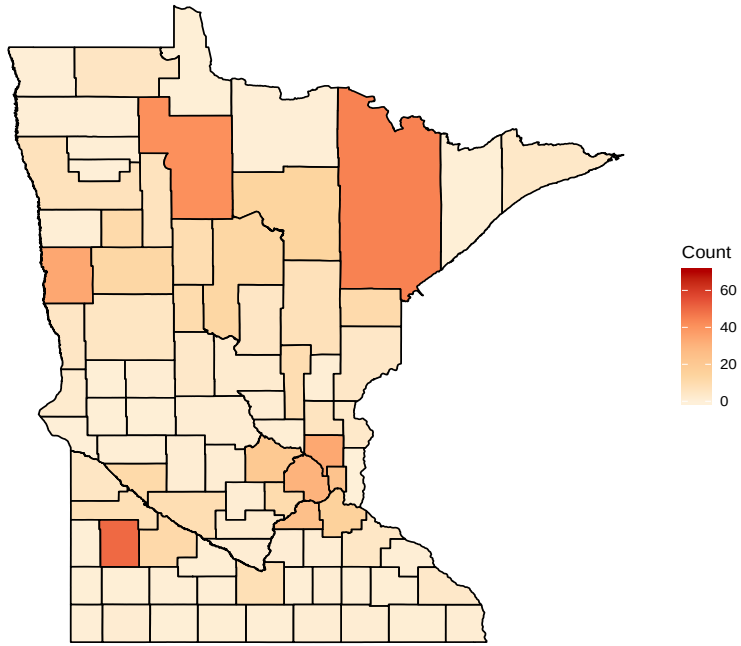


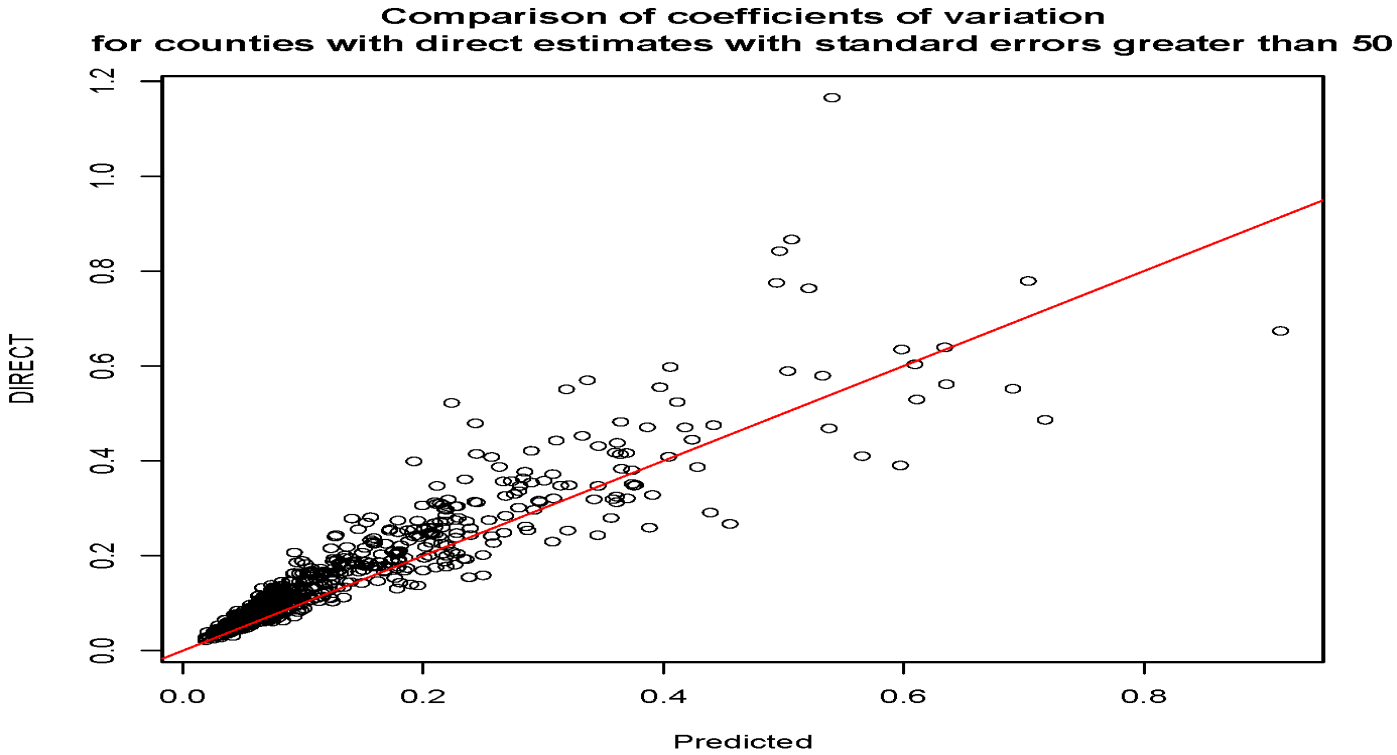
Comparison of the Standard Errors of Direct Estimates and Standard Errors of Predicted Values

Standard errors of direct estimates of the number of American Indian or Alaska Native children, ages 2 − 3, in counties in Minnesota



Standard errors of model−based predictions of the number of American Indian or Alaska Native children, ages 2 − 3, in counties in Minnesota





Min	Q1	Median	Q3	Max
57	30	20	7	-70

Table: Percent reduction in coefficients of variation for counties with direct estimates with standard errors greater than 50

Conclusions

- We were able to produce model-based estimates with greater precision than the corresponding direct estimates, despite the lack of strong covariates
- A multivariate mixed effects spatial model can be used for a wide variety of ACS special tabulations
- For more complicated datasets with varying spatial and/or multivariate characteristics, we developed a multivariate spatial mixture model which provides greater flexibility
- The model is of independent interest and is applicable to a broad set of subject-matter scientific problems

Thank you!

holans@missouri.edu

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2015). "Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics." *Annals of Applied Statistics*, 9, 4, 1761–1791.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles*, 76, 1, 1–32.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

Hughes, J. and Haran, M. (2013). "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models." *Journal of the Royal Statistical Society, Series B*, 75, 1, 139 – 159.

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9, 2, 249 – 265.

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4, 639 – 650.