Spatial Change of Support Models for Differentially Private Decennial Census Counts of Persons by Detailed Race and Ethnicity

Andrew Raim[1], James Livsey[1], Kyle Irimata[1], Ryan Janicki[1], and Scott H. Holan[23]

[1]Center for Statistical Research and Methodology, U.S. Census Bureau
[2]University of Missouri
[3]Associate Directorate for Research and Methodology, U.S. Census Bureau

2022 FCSM Research and Policy Conference
October 27, 2022

Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors, and not those of the U. S. Census Bureau.

Decennial Census

According to Abowd (2018), JASON (2020),

- In 2010, the U. S. Census counted a total population of over 308 million people
- At least 7.7 billion statistics were published from the collected data
- At least 25 published statistics per person
- In 2018, Census succeeded in reconstructing, from published 2010 census data, geographic location, sex, age and ethnicity for 46% of the U. S. population
- Census was able to link 38% of the reconstructed micro data to information in commercial databases

Differential Privacy

The U. S. Census Bureau will implement differential privacy for the 2020 Census.

A statistic, T, is (ϵ, δ)-differentially private if for any two data sets X and X' differing by a single element and any $\mathcal{A}$ in the range of T,

$$P\left(T(X) \in \mathcal{A}\right) \leq e^{\epsilon} P\left(T(X') \in \mathcal{A}\right) + \delta$$

- Laplace noise: $(\epsilon, 0)-$differentially private
- Gaussian noise: $(\epsilon, \delta)-$differentially private

Consequences of differential privacy

Accuracy/privacy tradeoff

- Published estimates will be noisy
- Fewer estimates may be published

Our research goal: use model-based methods to

- Produce estimates which are more precise than those based on differentially private measurements
- Produce estimates when no differentially private measurement is available

Notation

Source support

- Let $A_1, \ldots, A_m$ be a set of non-overlapping geographies representing a "source": Counties in this example

Target support

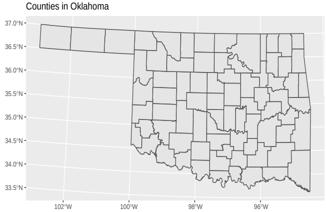
- Let $B_1, \ldots, B_n$ be a second set of geographies representing a "target": American Indian and Alaska Native (AIAN) areas in this example

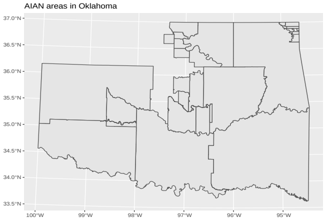
Notation, continued

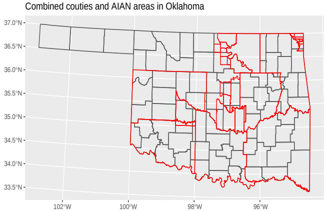
- For each area A_i in the source support we assume access to noisy measurements $Z(A_i)$ of an unobservable true value $Y(A_i)$, as well as a set of vector of predictors, $\boldsymbol{x}^T(A_i)$.
- For each area B_j in the target support, we have only knowledge of a set of predictors, $\boldsymbol{x}^T(B_j)$. We do not have noisy measurements, $Z(B_j)$, on the target support.
- Our goal is prediction of the true values, $Y(A_i)$ and $Y(B_j)$, on both the source support and the target support, using the observations $\{Z(A_i)\}$, and the predictors $\left\{\boldsymbol{x}^T(A_i)\right\}$ and $\left\{\boldsymbol{x}^T(B_j)\right\}$.

United States Census Bureau

Counties and AIAN areas in Oklahoma







United States Census Bureau

Statistical modeling framework

For a general region, A, the observed noisy measurements $Z(A)$ satisfy

$$Z(A) = Y(A) + \varepsilon(A),$$

where $\varepsilon(A)$ is a draw from a known differentially private distribution and $Y(A)$ is the unobservable true count.

Assume the true counts, $Y(A)$, can be aggregated from a point-level process

$$Y(A) = \int_A Y(s)ds,$$

Further assume that the point-level process can be decomposed as

$$Y(s) = \mu(s) + \gamma(s)$$

where $\mu(s)$ represents the fixed effects, and $\gamma(s)$ represents the random effects, which account for spatial dependencies and residual variation.

United States Census Bureau

Fixed effects

The model for the fixed effects, $\mu(s)$, is

$$\mu(s) = \boldsymbol{x}^T(s)\boldsymbol{\beta},$$

so that

$$\mu(A) = \int_A \mu(s)ds = \int_A \boldsymbol{x}^T(s)\boldsymbol{\beta}ds = \boldsymbol{x}^T(A)\boldsymbol{\beta}.$$

In our examples, $\boldsymbol{x}$ includes

- An intercept
- The count from the previous census
- The American Community Survey (ACS) 5-year estimate

United States Census Bureau

Random effects

We use a basis expansion for the random process, $\gamma(s)$ (Cressie and Johannesson, 2008; Bradley et al., 2017):

$$\gamma(s) = \sum_{k=1}^{\infty} \psi_k(s)\eta_k \approx \sum_{k=1}^{r} \psi_k(s)\eta_k + \xi(s),$$

where $\{\psi_k(s)\}$ is a collection of spatial basis functions, and η_k are independent, mean-zero Gaussian random variables and $\xi(s)$ is a residual random effect.

$$\gamma(A) = \int_A \left(\sum_{k=1}^{r} \psi_k(s)\eta_k + \xi(s) \right) ds = \sum_{k=1}^{r} \psi_k(A)\eta_k + \xi(A).$$

United States Census Bureau

Random coefficients

We assume a multivariate normal distribution for the random effects:

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)^T \sim N_r\left(\mathbf{0}, \sigma^2_{\boldsymbol{\eta}} \boldsymbol{K}\right)$$

- $\boldsymbol{K}$ is a known covariance matrix, constructed to induce spatial dependencies using a conditional autoregressive structure (Hughes and Haran, 2013)
- $\sigma^2_{\boldsymbol{\eta}}$ is an unknown parameter

Construction of $\boldsymbol{K}$

$\boldsymbol{K}$ is constructed to

- induce spatial dependencies
- reduce rank compared to a conditional auto regressive process

Let

- $\boldsymbol{P_X} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$
- $\boldsymbol{A}$ the adjacency matrix for counties
- $\boldsymbol{S}$ the first r eigenvectors of $\boldsymbol{P_X A P_X}$

United States Census Bureau

Construction of $\boldsymbol{K}$, continued

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

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

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

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

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

United States Census Bureau

Basis functions

Bisquare basis functions

$$\psi_k(s) = \left(1 - \frac{\|\boldsymbol{s} - \boldsymbol{c}_j\|^2}{w^2}\right)^2 I\left(\|\boldsymbol{s} - \boldsymbol{c}_j\| < w\right).$$

- $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_r)$ is a collection of equally-spaced knots
- w is 1.5 times the minimum distance between any two knots

The integral

$$\psi_k(A) = \int_A \psi_k(s) ds$$

is approximated numerically for each region A. (Bradley et al., 2017; Pebesma, 2018; Raim et al., 2021)

Source support model

Data model:

$$Z(A_i) = Y(A_i) + \varepsilon(A_i)$$

Process model:

$$Y(A_i) = \boldsymbol{x}^T(A_i)\boldsymbol{\beta} + \sum_{k=1}^{r} \psi_k(A_i)\eta_k + \xi(A_i),\ \ i = 1, \ldots, m,$$

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)^T \sim N_r\left(\mathbf{0}, \sigma_\eta^2 \boldsymbol{K}\right),\ \ \ \xi\left(A_i\right) \stackrel{i.i.d.}{\sim} N\left(0, \sigma_\xi^2\right)$$

Parameter model:

$$\boldsymbol{\beta} \sim N_p\left(\mathbf{0}, 10\boldsymbol{I}_{p\times p}\right),\ \ \ \sigma_{\boldsymbol{\eta}}^2 \sim IG(1,1),\ \ \ \sigma_\xi^2 \sim IG(1,1)$$

This model can be fit using a Gibbs sampler (Choi and Hobert, 2013).

United States Census Bureau

Change of support

The true count, $Y(B_j)$, in area B_j can be estimated using

$$\hat{Y}(B_j) = E\left(Y(B_j) \mid \{Z(A_i)\}_{i=1}^{n}, \boldsymbol{x}(B_j)\right).$$

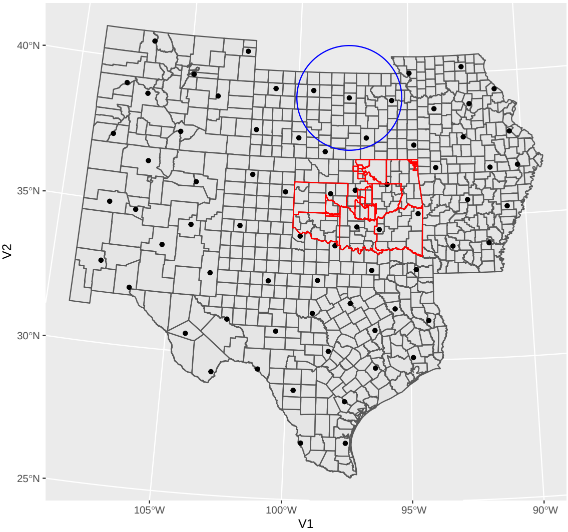
The covariates $\boldsymbol{x}^T(B_j)$ are assumed known, and the basis functions $\psi_k(B_j)$ are approximated using numerical integration. The distribution of

$$[Y(B_j) \mid \boldsymbol{Z}] = \int_{\boldsymbol{\theta}} [Y(B_j) \mid \boldsymbol{\theta}]\,[\boldsymbol{\theta} \mid \boldsymbol{Z}]\, d\boldsymbol{\theta}$$

can be approximated using the output of the Gibbs sampler for fitting the source support model.

Example: estimation of the number of Choctaw persons in counties and AIAN areas in Oklahoma

- Let $Y(A_i)$ be the Census 2010 count of the number of Choctaw persons in county i in Oklahoma
- Let $Y(B_j)$ be the Census 2010 count of the number of Choctaw persons in AIAN area j in Oklahoma
- $\boldsymbol{x}(s)$ includes an intercept, the Census 2000 count, and the 2009 ACS 5-year estimate.
- We generate $Z(A_i) = Y(A_i) + \varepsilon(A_i)$, where $\varepsilon(A_i) \overset{i.i.d.}{\sim} Lap(48)$
- 1000 data sets were created
- Our goal is estimation of $Y(A_i)$ and $Y(B_i)$ from the observed $Z(A_i)$



United States™ Census Bureau

Estimation of the number of Choctaw persons in counties in Oklahoma

	MOD	DP
RMSE	38	68
RMAE	4.5	10.1
MAX	241	346
Coverage	95%	95%

The metrics used are

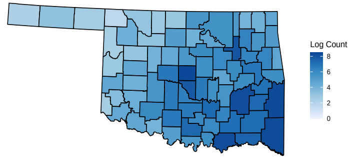
$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{n}\left(\hat{Y}_j - Y_j\right)^2}, \quad RMAE = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\mid \hat{Y}_i - Y_i \mid}{Y_i}\right).$$

$$MAX = \max_{i=1,\ldots,m}\left|\hat{Y}_j - Y_j\right|$$

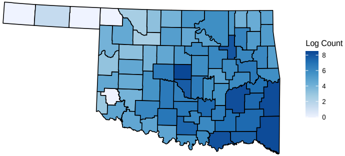
United States Census Bureau

Estimation of the number of Choctaw persons in counties in Oklahoma

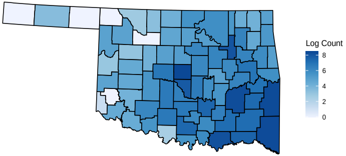
Log of 2010 Census count of Choctaw persons in counties in Oklahoma



Log of predicted number of persons in counties in Oklahoma



Log of DP measurements of number of persons in counties in Oklahoma



Estimation of the number of Choctaw persons in AIAN areas in Oklahoma

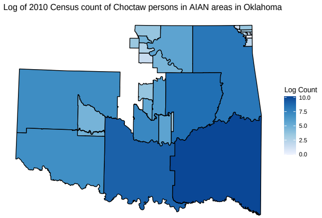
	MOD	AREAL
RMSE	145	1064
RMAE	0.35	1.97
MAX	592	5631
Coverage	91%	NA

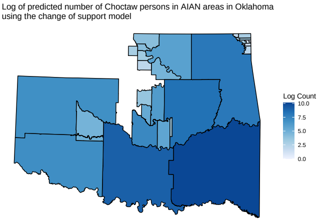
Comparison of model-based predictions with simple proportional allocation (Prener et al., 2019)

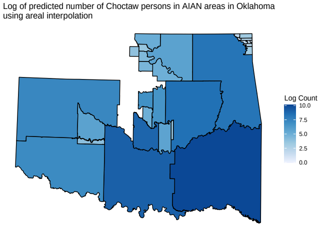
$$\hat{Y}_j = \sum_{i=1}^{n} Z(A_i) \frac{|B_j \cap A_i|}{|B_j|}$$

United States Census Bureau

Estimation of the number of Choctaw persons in AIAN areas in Oklahoma







Questions which need to be addressed

- Sensitivity analyses
  - Should we include additional covariates?
  - Class of basis functions
  - How many basis functions to use?
  - Choice of tuning parameters
  - How many geographic regions to include in the model?
- Log transformation gives a better fit to some data sets, but doesn't allow for change of support
- Test on other data sets

United States™ Census Bureau

References I

John M. Abowd. The U. S. Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*, New York, NY, USA, 2018. Association for Computing Machinery. URL https://doi.org/10.1145/3219819.3226070.

Jonathan R. Bradley, Christopher K. Wikle, and Scott H. Holan. Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error. *Journal of the Royal Statistical Society, Series B*, 79:815 – 832, 2017.

Hee Min Choi and James P. Hobert. Analysis of mcmc algorithms for bayesian linear regression with laplace errors. *Journal of Multivariate Analysis*, 117: 32–40, 2013. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2013.02.004. URL https://www.sciencedirect.com/science/article/pii/S0047259X13000183.

Noel Cressie and Gardar Johannesson. Fixed rank Kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70:209 – 226, 2008.

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

United States Census Bureau

References II

JASON. Formal privacy methods for the 2020 Census. Technical report, The MITRE Corportation, 2020.

Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018. doi: 10.32614/RJ-2018-009. URL `https://doi.org/10.32614/RJ-2018-009`.

Prener, Christopher, Revord, and Charles. areal: An R package for areal weighted interpolation. *Journal of Open Source Software*, 4(37), 2019. doi: 10.21105/joss.01221. URL `https://doi.org/10.21105/joss.01221`.

Andrew M. Raim, Scott H. Holan, Jonathan R. Bradley, and Christopher K. Wikle. Spatio-temporal change of support modeling with R. *Computational Statistics*, 36:749 – 780, 2021.

United States Census Bureau

Thank You!

- ryan.janicki@census.gov

United States Census Bureau