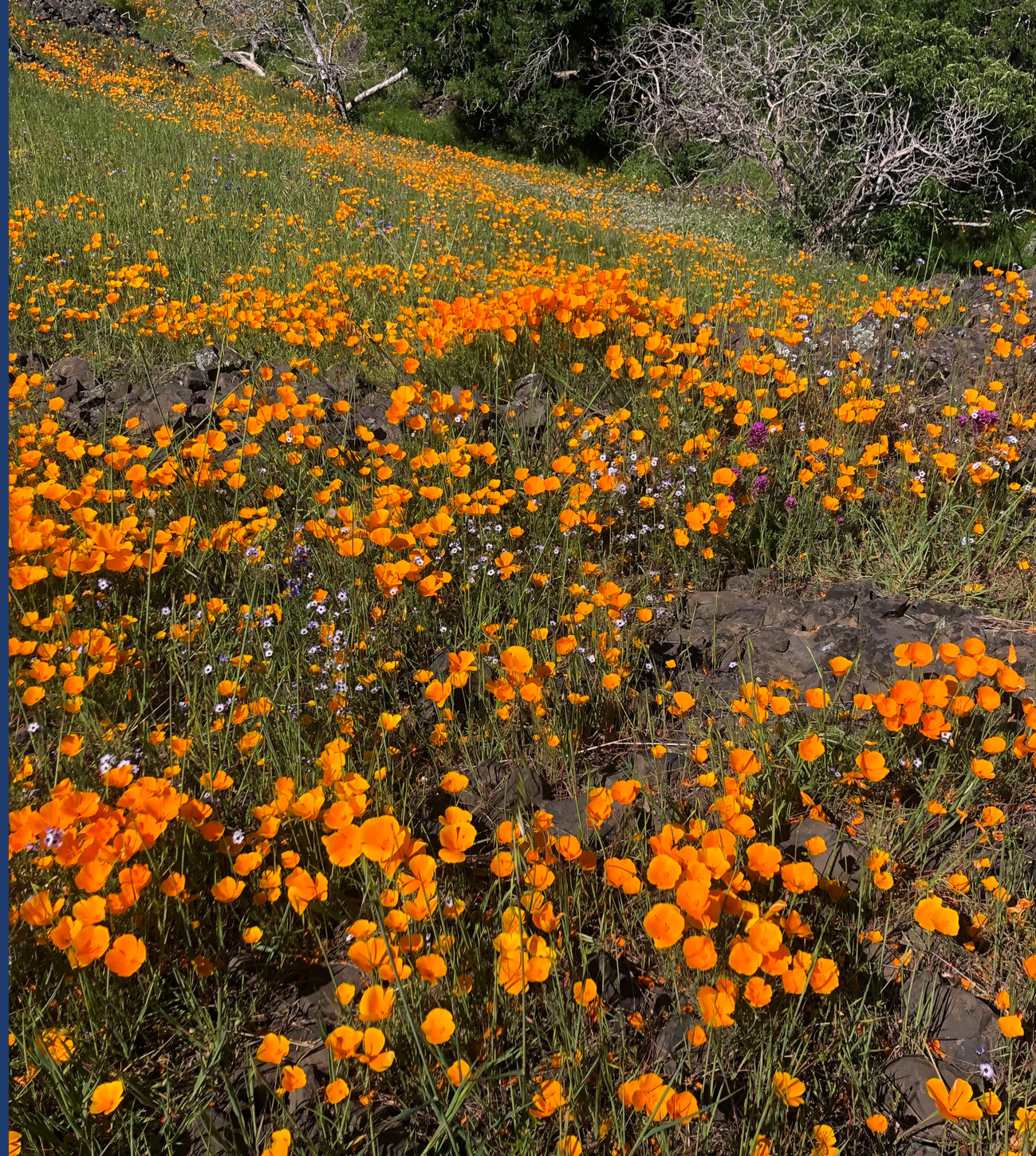
Spatially Selected & Dependent Random Effects for Small Area Estimation with Application to Rent Burden

Sho Kawano
Co-Authors: Paul Parker, Zehang Richard Li
UC Santa Cruz

INNOVATIVE MODELING APPROACHES FOR SMALL AREA ESTIMATION IN THE PRESENCE OF COMPLEX DEPENDENCE STRUCTURES

JSM, AUGUST 4TH 2024

BACKGROUND & MOTIVATION



SMALL AREA ESTIMATION

- Survey is conducted in study area with n subregions
- Goal: estimate θ_i from each subregion $i = 1, \cdots, n$ using…
  - y_i: direct estimate
  - d_i: survey variance
- Problem: y_i can be unreliable with high d_i for subregions with small sample size (“small area”)

FAY-HERRIOT MODEL

- Problem: y_i can be unreliable with high d_i for small sample regions
- Solution: use a model that "borrows strength" across areas

FAY-HERRIOT MODEL

- Problem: y_i can be unreliable with high d_i for small sample regions
- Solution: use a model that "borrows strength" across areas
- *Fay & Herriot (1979)* proposed

$$[\boldsymbol{y} \mid \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}), \quad \boldsymbol{\theta} = \boldsymbol{X\beta} + \boldsymbol{u}, \text{ and } \boldsymbol{u} \sim N_n(0, \sigma^2 I)$$

- where $\boldsymbol{D} = diag\{d_i\}_{i=1}^n$ and survey variances d_i,
- $n \times j$ covariate matrix $\boldsymbol{X}$, coefficients $\boldsymbol{\beta}$
- $\boldsymbol{u}$ the random effects

FAY-HERRIOT MODEL

- Problem: y_i can be unreliable with high d_i for small sample regions
- Solution: use a model that "borrows strength" across areas
- *Fay & Herriot (1979)* proposed

$$[\boldsymbol{y} \mid \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}), \quad \boldsymbol{\theta} = \boldsymbol{X\beta} + \boldsymbol{u}, \text{ and } \boldsymbol{u} \sim N_n(0, \sigma^2 I)$$

- where $\boldsymbol{D} = diag\{d_i\}_{i=1}^n$ and survey variances d_i,
- $n \times j$ covariate matrix $\boldsymbol{X}$, coefficients $\boldsymbol{\beta}$
- $\boldsymbol{u}$ the random effects

Notation: $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Multivariate Normal Distribution of dimension m
$\boldsymbol{\mu}$ *mean,* covariance $\boldsymbol{\Sigma}$

RESEARCH MOTIVATION

FH *Model:* $[\boldsymbol{y}\,|\,\boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D})$, with $\boldsymbol{\theta} = \boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{u}$ and $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$

Issues with IID Normal Assumption

RESEARCH MOTIVATION

FH *Model:* $[\boldsymbol{y}\,|\,\boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D})$, with $\boldsymbol{\theta} = \boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{u}$ and $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$

Issues with IID Normal Assumption

1. Spatial dependence is not modeled

RESEARCH MOTIVATION

FH *Model:* $[\boldsymbol{y}\,|\,\boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D})$, with $\boldsymbol{\theta} = \boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{u}$ and $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$

Issues with IID Normal Assumption

1. Spatial dependence is not modeled

2. Not all random effects may be needed

RESEARCH MOTIVATION

FH *Model:* $[\boldsymbol{y}\,|\,\boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D})$, with $\boldsymbol{\theta} = \boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{u}$ and $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$

Issues with IID Normal Assumption

1. Spatial dependence is not modeled

2. Not all random effects may be needed

Could we address both of these issues?

EXISTING APPROACHES & PROPOSED MODEL

THE FAY-HERRIOT MODEL: ISSUES

- *Model:* $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Issues with IID Normal Assumption $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$

  1. Spatial dependence is not modeled

  2. Not all random effects may be needed

1. MODELING SPATIAL DEPENDENCE

- Solution: assume $\boldsymbol{u} \sim N_n(0, \sigma^2 \mathbf{Q}^{-1})$ with spatial precision matrix $\mathbf{Q}$

1. MODELING SPATIAL DEPENDENCE

- Solution: assume $\boldsymbol{u} \sim N_n(0, \sigma^2 \mathbf{Q}^{-1})$ with spatial precision matrix $\mathbf{Q}$
- Common Choice: Conditional Autoregressive (CAR) prior
- $\mathbf{Q} = diag\{n_i\}_{i=1}^n - \rho \boldsymbol{A}$ with spatial corr. parameter ρ, A adjacency matrix, n_i no. neighbors for area i
- Set an adjacency matrix $\boldsymbol{A}$ such that $A_{ij} = 1$ if areas i & j are neighbors ($i \sim j$) and $A_{ij} = 0$ otherwise.
- Then the conditional distribution of u_i given the others u_{-i} is:

$$[u_i \,|\, u_{-i}, \sigma^2, \rho] \sim N\left(\frac{\rho}{n_i} \sum_{i \sim j} u_j, \frac{\sigma^2}{n_i}\right)$$

- $\rho = 1$ results in the improper intrinsic CAR (ICAR) prior

1. MODELING SPATIAL DEPENDENCE: BYM EFFECTS

- Problem: Using CAR when there is no spatial correlation can yield misleading results (Lereoux et al., 2000; Wakefield, 2007)

1. MODELING SPATIAL DEPENDENCE: BYM EFFECTS

- Problem: Using CAR when there is no spatial correlation can yield misleading results (Lereoux et al., 2000; Wakefield, 2007)
- Solution: random effects with BYM structure *(Besag, York, Mollié. 1991)*

1. MODELING SPATIAL DEPENDENCE: BYM EFFECTS

- Problem: Using CAR when there is no spatial correlation can yield misleading results (Lereoux et al., 2000; Wakefield, 2007)
- Solution: random effects with BYM structure *(Besag, York, Mollié. 1991)*

$$\boldsymbol{u} = \boldsymbol{v}_1 + \boldsymbol{v}_2 \textit{ with}$$

1. MODELING SPATIAL DEPENDENCE: BYM EFFECTS

- Problem: Using CAR when there is no spatial correlation can yield misleading results (Lereoux et al., 2000; Wakefield, 2007)
- Solution: random effects with BYM structure *(Besag, York, Mollié. 1991)*

$$\boldsymbol{u} = \boldsymbol{v}_1 + \boldsymbol{v}_2 \text{ with}$$

IID effect

$$[\boldsymbol{v_1}] \sim N_n(\mathbf{0}, \sigma_1^2 \boldsymbol{I})$$

1. MODELING SPATIAL DEPENDENCE: BYM EFFECTS

- Problem: Using CAR when there is no spatial correlation can yield misleading results (Lereoux et al., 2000; Wakefield, 2007)
- Solution: random effects with BYM structure *(Besag, York, Mollié. 1991)*

$$\boldsymbol{u} = \boldsymbol{v}_1 + \boldsymbol{v}_2 \text{ with}$$

IID effect

$$[\boldsymbol{v_1}] \sim N_n(\mathbf{0}, \sigma_1^2 \boldsymbol{I})$$

Spatial effect

$$[\boldsymbol{v_2}] \sim N_n(\mathbf{0}, \sigma_2^2 \mathbf{Q}^-)$$

$\mathbf{Q}^-$ *: gen. inverse of ICAR precision matrix.*

THE FAY-HERRIOT MODEL: ISSUES

- *Model:* $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Issues with IID Normal Assumption $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$

  1. Spatial dependence is not modeled

  2. Not all random effects may be needed

THE FAY-HERRIOT MODEL: ISSUES

- *Model:* $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Issues with IID Normal Assumption $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$

  1. Spatial dependence is not modeled
  2. Not all random effects may be needed
     - Datta, Hall, and Mandal (2011) showed this via a hypothesis test
     - Problem: Can only test including all random effects or none.

2. NOT ALL EFFECTS MAY BE NEEDED

- Solution: Use Bayesian variable selection

2. NOT ALL EFFECTS MAY BE NEEDED

- Solution: Use Bayesian variable selection
- Datta and Mandal (2015) proposed a model with a spike-and-slab prior

$$u_i = \delta_i \cdot v_i$$

where $[v_i \mid \delta_i = 1, \sigma^2] \overset{ind}{\sim} N(0, \sigma^2)$

and $[v_i \mid \delta_i = 0] = 0$

and $[\delta_i \mid p] \overset{iid}{\sim} Bernoulli(p)$ with $[p] \sim Beta(a, b)$.

2. NOT ALL EFFECTS MAY BE NEEDED

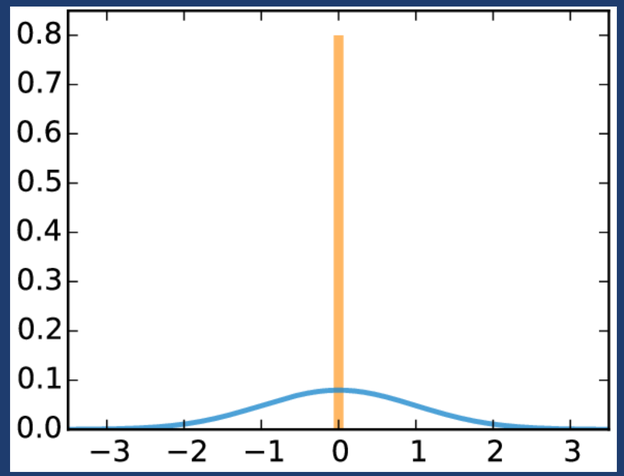
- Solution: Use Bayesian variable selection
- Datta and Mandal (2015) proposed a model with a spike-and-slab prior

$$u_i = \delta_i \cdot \boldsymbol{v}_i$$

where $[v_i \,|\, \delta_i = 1, \sigma^2] \overset{ind}{\sim} N(0, \sigma^2)$

and $[v_i \,|\, \delta_i = 0] = 0$

and $[\delta_i \,|\, p] \overset{iid}{\sim} Bernoulli(p)$ with $[p] \sim Beta(a, b)$.



2. NOT ALL EFFECTS MAY BE NEEDED

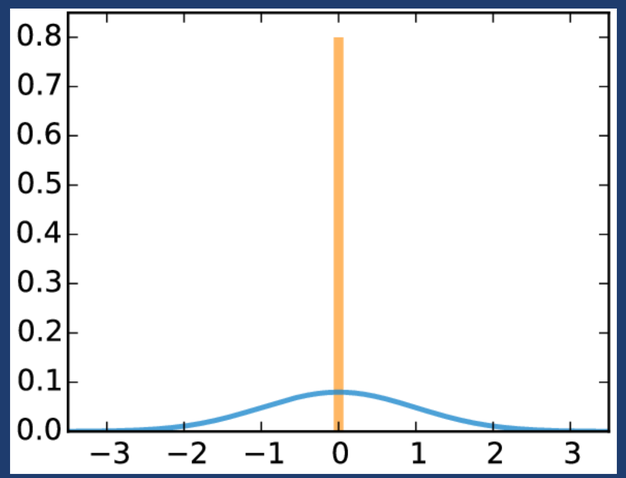
- Solution: Use Bayesian variable selection
- Datta and Mandal (2015) proposed a model with a spike-and-slab prior

$$u_i = \delta_i \cdot \boldsymbol{v}_i$$

where $[v_i \,|\, \delta_i = 1, \sigma^2] \overset{ind}{\sim} N(0, \sigma^2)$ effect is selected

and $[v_i \,|\, \delta_i = 0] = 0$

and $[\delta_i \,|\, p] \overset{iid}{\sim} Bernoulli(p)$ with $[p] \sim Beta(a, b)$.



2. NOT ALL EFFECTS MAY BE NEEDED

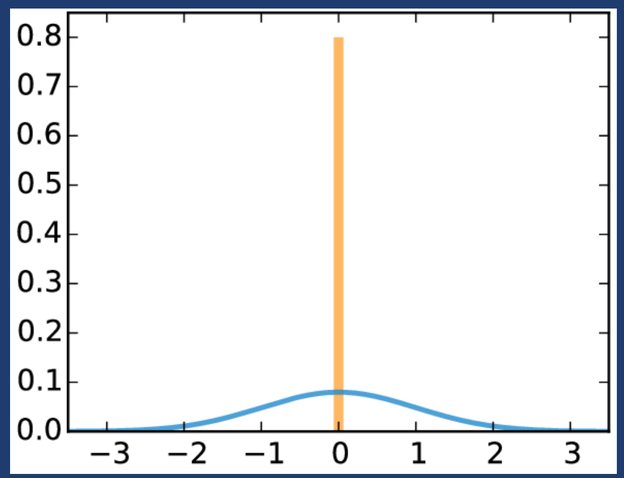
- Solution: Use Bayesian variable selection
- Datta and Mandal (2015) proposed a model with a spike-and-slab prior

$$u_i = \delta_i \cdot v_i$$

where $[v_i \mid \delta_i = 1, \sigma^2] \overset{ind}{\sim} N(0, \sigma^2)$ effect is selected

and $[v_i \mid \delta_i = 0] = 0$ effect is not selected

and $[\delta_i \mid p] \overset{iid}{\sim} Bernoulli(p)$ with $[p] \sim Beta(a, b)$.



COMBINING SPATIAL MODELING WITH SELECTION

- Datta-Mandal Model

$$u_i = \delta_i \cdot v_i$$

$$[v_i \mid \delta_i = 1, \sigma^2] \overset{ind}{\sim} N(0, \sigma^2)$$
$$[v_i \mid \delta_i = 0] = 0$$

Random Effects

$$[\delta_i \mid p] \overset{iid}{\sim} Bernoulli(p) \text{ with}$$
$$[p] \sim Beta(a, b).$$

Selection Process

COMBINING SPATIAL MODELING WITH SELECTION

- Datta-Mandal Model

$$u_i = \delta_i \cdot v_i$$

Could spatial dependence be incorporated into both levels?

$$[v_i \mid \delta_i = 1, \sigma^2] \overset{ind}{\sim} N(0, \sigma^2)$$
$$[v_i \mid \delta_i = 0] = 0$$

Random Effects

$$[\delta_i \mid p] \overset{iid}{\sim} Bernoulli(p) \text{ with}$$
$$[p] \sim Beta(a, b).$$

Selection Process

PROPOSED MODEL

- Data Model: $[\boldsymbol{y}\,|\,\boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{u}$

PROPOSED MODEL

- Data Model: $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

$$u_i = \delta_i \cdot (v_{1i} + v_{2i})$$

$$[\boldsymbol{v}_1 \,|\, \sigma_1^2] \sim N_n(\mathbf{0}, \sigma_1^2 \boldsymbol{I}) \text{ and } [\boldsymbol{v}_2 \,|\, \sigma_2^2] \sim N_n(\mathbf{0}, \sigma_2^2 \mathbf{Q}^-)$$

PROPOSED MODEL

- Data Model: $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

$$u_i = \delta_i \cdot (v_{1i} + v_{2i})$$

$$[\boldsymbol{v}_1 \,|\, \sigma_1^2] \sim N_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I}) \text{ and } [\boldsymbol{v}_2 \,|\, \sigma_2^2] \sim N_n(\boldsymbol{0}, \sigma_2^2 \mathbf{Q}^-)$$

- Selection Process:

$$[\delta_i \,|\, p_i] \overset{ind}{\sim} Bernoulli(p_i) \text{ with } logit(\boldsymbol{p}) = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$$

$$[\boldsymbol{\psi}_1 \,|\, s_1^2] \sim N_n(\boldsymbol{0}, s_1^2 \boldsymbol{I}) \text{ and } [\boldsymbol{\psi}_2 \,|\, s_2^2] \sim N_n(\boldsymbol{0}, s_2^2 \boldsymbol{Q}^-)$$

PROPOSED MODEL

- Data Model: $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

  $u_i = \delta_i \cdot (v_{1i} + v_{2i})$

  $[\boldsymbol{v}_1 \,|\, \sigma_1^2] \sim N_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$ and $[\boldsymbol{v}_2 \,|\, \sigma_2^2] \sim N_n(\boldsymbol{0}, \sigma_2^2 \mathbf{Q}^-)$

- Selection Process:

  $[\delta_i \,|\, p_i] \overset{ind}{\sim} Bernoulli(p_i)$ with $logit(\boldsymbol{p}) = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$

  $[\boldsymbol{\psi}_1 \,|\, s_1^2] \sim N_n(\boldsymbol{0}, s_1^2 \boldsymbol{I})$ and $[\boldsymbol{\psi}_2 \,|\, s_2^2] \sim N_n(\boldsymbol{0}, s_2^2 \boldsymbol{Q}^-)$

✅ Incorporates Spatial Dependence on both levels

PROPOSED MODEL

Spatially Selected & Dependent Random Effects (SSD) Model

- Data Model: $[\boldsymbol{y} \mid \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

  $u_i = \delta_i \cdot (v_{1i} + v_{2i})$

  $[\boldsymbol{v}_1 \mid \sigma_1^2] \sim N_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$ and $[\boldsymbol{v}_2 \mid \sigma_2^2] \sim N_n(\boldsymbol{0}, \sigma_2^2 \mathbf{Q}^-)$

✅ Incorporates Spatial Dependence on both levels

- Selection Process:

  $[\delta_i \mid p_i] \overset{ind}{\sim} Bernoulli(p_i)$ with $logit(\boldsymbol{p}) = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$

  $[\boldsymbol{\psi}_1 \mid s_1^2] \sim N_n(\boldsymbol{0}, s_1^2 \boldsymbol{I})$ and $[\boldsymbol{\psi}_2 \mid s_2^2] \sim N_n(\boldsymbol{0}, s_2^2 \boldsymbol{Q}^-)$

PRIOR SPECIFICATION

- Coefficients: $\boldsymbol{\beta} \sim N_j(\mathbf{0},\ 100^2\boldsymbol{I})$
- Random effect variances: $\sigma_1^2 \sim IG(a_1, b_1)$ and $\sigma_2^2 \sim IG(a_2, b_2)$
- Logit effect variances: $s_1^2 \sim IG(c_1, d_1)$ and $s_2^2 \sim IG(c_2, d_2)$

PRIOR SPECIFICATION

- Coefficients: $\boldsymbol{\beta} \sim N_j(\mathbf{0},\ 100^2\boldsymbol{I})$

Notation: $IG(A, B)$ is a inverse-gamma distribution with shape A and scale B

- Random effect variances: $\sigma_1^2 \sim IG(a_1, b_1)$ and $\sigma_2^2 \sim IG(a_2, b_2)$
- Logit effect variances: $s_1^2 \sim IG(c_1, d_1)$ and $s_2^2 \sim IG(c_2, d_2)$

PRIOR SPECIFICATION

Notation: $IG(A, B)$ is a inverse-gamma distribution with shape A and scale B

- Coefficients: $\boldsymbol{\beta} \sim N_j(\mathbf{0},\ 100^2\boldsymbol{I})$
- Random effect variances: $\sigma_1^2 \sim IG(a_1, b_1)$ and $\sigma_2^2 \sim IG(a_2, b_2)$
- Logit effect variances: $s_1^2 \sim IG(c_1, d_1)$ and $s_2^2 \sim IG(c_2, d_2)$
- Recommend:
  - Scale the data $\{\boldsymbol{y}, \boldsymbol{D}\}$, like Bayesian Lasso
  - Scale ICAR precision matrix $\mathbf{Q}$ so that σ_1^2, σ_2^2 / s_1^2, s_2^2 are comparable (Sørbye and Rue, 2014).
  - Set priors for σ_1^2, σ_2^2 so that slab distribution is wide

PROPOSED MODEL

Notation: $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Multivariate Normal Distribution of dimension m
$\boldsymbol{\mu}$ *mean,* covariance $\boldsymbol{\Sigma}$

- Data Model: $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

  $u_i = \delta_i \cdot (v_{1i} + v_{2i})$

  with $[\boldsymbol{v}_1 \,|\, \sigma_1^2] \sim N_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$ and $[\boldsymbol{v}_2 \,|\, \sigma_2^2] \sim N_n(\boldsymbol{0}, \sigma_2^2 \mathbf{Q}^-)$
- Selection Process:

  $[\delta_i \,|\, p_i] \overset{ind}{\sim} Bernoulli(p_i)$ and $logit(\boldsymbol{p}) = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$

  with $[\boldsymbol{\psi}_1 \,|\, s_1^2] \sim N_n(\boldsymbol{0}, s_1^2 \boldsymbol{I})$ and $[\boldsymbol{\psi}_2 \,|\, s_2^2] \sim N_n(\boldsymbol{0}, s_2^2 \boldsymbol{Q}^-)$
- Priors: $\boldsymbol{\beta} \sim N_j(\boldsymbol{0},\ 100^2 \boldsymbol{I})$; σ_1^2 and $\sigma_2^2 \sim IG(a, b)$;
  s_1^2 and $s_2^2 \sim IG(c, d)$

PROPOSED MODEL

Notation: $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Multivariate Normal Distribution of dimension m
$\boldsymbol{\mu}$ *mean,* covariance $\boldsymbol{\Sigma}$

- Data Model: $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

  $u_i = \delta_i \cdot (v_{1i} + v_{2i})$

  with $[\boldsymbol{v}_1 \,|\, \sigma_1^2] \sim N_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$ and $[\boldsymbol{v}_2 \,|\, \sigma_2^2] \sim N_n(\boldsymbol{0}, \sigma_2^2 \mathbf{Q}^-)$
- Selection Process:

  $[\delta_i \,|\, p_i] \overset{ind}{\sim} Bernoulli(p_i)$ and $logit(\boldsymbol{p}) = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$

  with $[\boldsymbol{\psi}_1 \,|\, s_1^2] \sim N_n(\boldsymbol{0}, s_1^2 \boldsymbol{I})$ and $[\boldsymbol{\psi}_2 \,|\, s_2^2] \sim N_n(\boldsymbol{0}, s_2^2 \boldsymbol{Q}^-)$
- Priors: $\boldsymbol{\beta} \sim N_j(\boldsymbol{0},\ 100^2 \boldsymbol{I})$; σ_1^2 and $\sigma_2^2 \sim IG(a, b)$; s_1^2 and $s_2^2 \sim IG(c, d)$

Pólya-Gamma Data Augmentation (Polson et al., 2013)

PROPOSED MODEL

- Data Model: $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

  $u_i = \delta_i \cdot (v_{1i} + v_{2i})$

  with $[\boldsymbol{v}_1 \,|\, \sigma_1^2] \sim N_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$ and $[\boldsymbol{v}_2 \,|\, \sigma_2^2] \sim N_n(\boldsymbol{0}, \sigma_2^2 \mathbf{Q}^-)$
- Selection Process:

  $[\delta_i \,|\, p_i] \overset{ind}{\sim} Bernoulli(p_i)$ and $logit(\boldsymbol{p}) = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$

  with $[\boldsymbol{\psi}_1 \,|\, s_1^2] \sim N_n(\boldsymbol{0}, s_1^2 \boldsymbol{I})$ and $[\boldsymbol{\psi}_2 \,|\, s_2^2] \sim N_n(\boldsymbol{0}, s_2^2 \boldsymbol{Q}^-)$
- Priors: $\boldsymbol{\beta} \sim N_j(\boldsymbol{0},\ 100^2 \boldsymbol{I})$; σ_1^2 and $\sigma_2^2 \sim IG(a, b)$; s_1^2 and $s_2^2 \sim IG(c, d)$

PROPOSED MODEL

- Data Model: $[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}),\ \boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{u}$
- Random Effects:

  $u_i = \delta_i \cdot (v_{1i} + v_{2i})$

  with $[\boldsymbol{v}_1 \,|\, \sigma_1^2] \sim N_n(\boldsymbol{0}, \sigma_1^2 \boldsymbol{I})$ and $[\boldsymbol{v}_2 \,|\, \sigma_2^2] \sim N_n(\boldsymbol{0}, \sigma_2^2 \mathbf{Q}^-)$
- Selection Process:

  $[\delta_i \,|\, p_i] \overset{ind}{\sim} Bernoulli(p_i)$ and $logit(\boldsymbol{p}) = \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2$

  with $[\boldsymbol{\psi}_1 \,|\, s_1^2] \sim N_n(\boldsymbol{0}, s_1^2 \boldsymbol{I})$ and $[\boldsymbol{\psi}_2 \,|\, s_2^2] \sim N_n(\boldsymbol{0}, s_2^2 \boldsymbol{Q}^-)$
- Priors: $\boldsymbol{\beta} \sim N_j(\boldsymbol{0},\ 100^2 \boldsymbol{I})$; σ_1^2 and $\sigma_2^2 \sim IG(a, b)$; s_1^2 and $s_2^2 \sim IG(c, d)$

Posterior Inference with Gibbs Sampler

ESTIMATING MEDIAN RENT BURDEN

DATA ANALYSIS

DATA ANALYSIS: SETTING

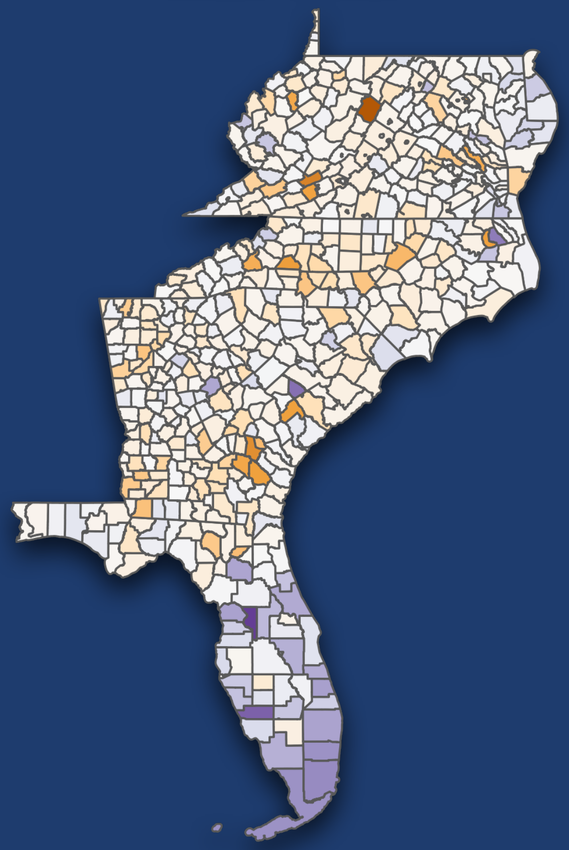
- Variable of interest θ_i: rent burden, share of household income used to pay rent
- Policy-Relevance: recent ACS estimates suggest that over half of U.S. renters are paying >30% of their income in rent (highest in record)

DATA ANALYSIS: SETTING

- Variable of interest θ_i: rent burden, share of household income used to pay rent
- Policy-Relevance: recent ACS estimates suggest that over half of U.S. renters are paying >30% of their income in rent (highest in record)
- Data: 5-year estimates from the 2019 ACS, South Atlantic Census Division ($n = 588$)
- Models fit with log-transformed direct estimates, delta method used to estimate variance
- *Covariates $\boldsymbol{X}$ also from ACS (education, race, poverty related)*

DATA ANALYSIS: COMPARING DATTA-MANDAL VS. PROPOSED MODEL

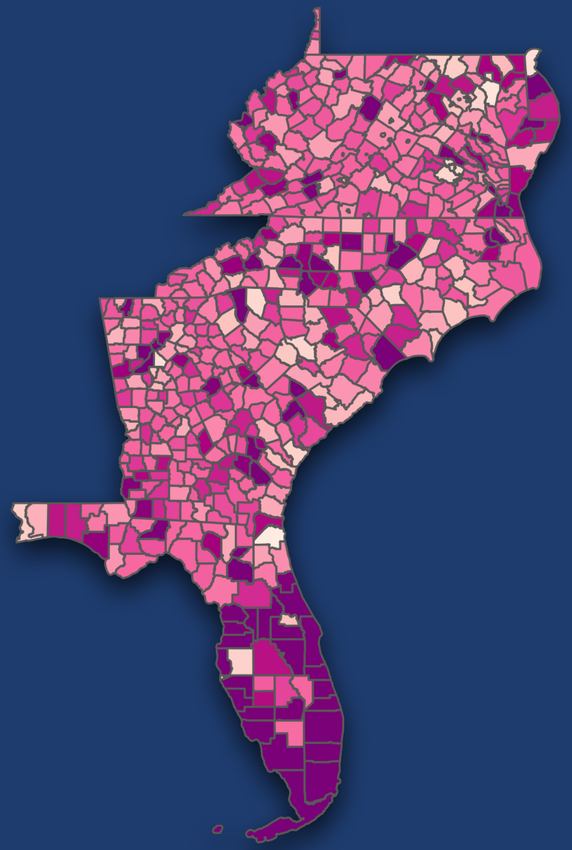
Datta-Mandal



Random Effects

-0.2 -0.1 0.0 0.1 0.2

Datta-Mandal

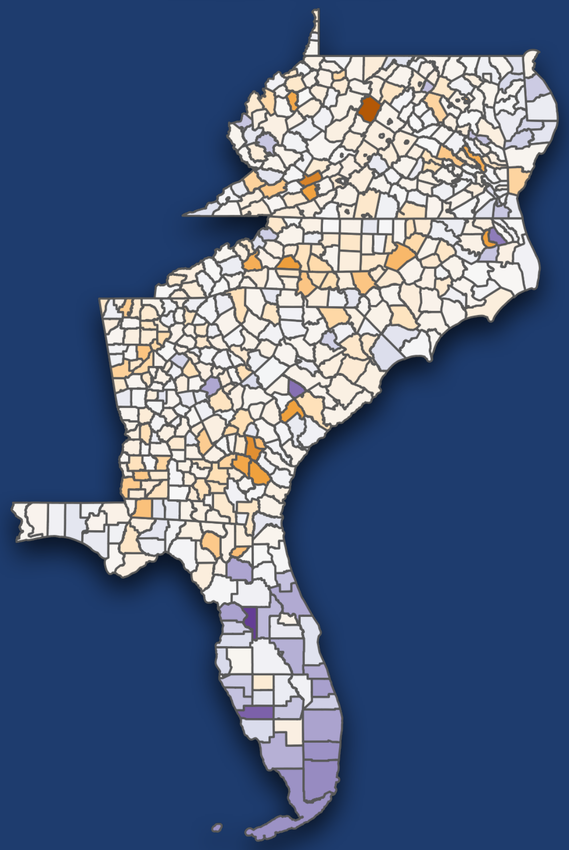


Selection Probabilities

0.2 0.4 0.6 0.8 1.0

DATA ANALYSIS: COMPARING DATTA-MANDAL VS. PROPOSED MODEL

Datta-Mandal

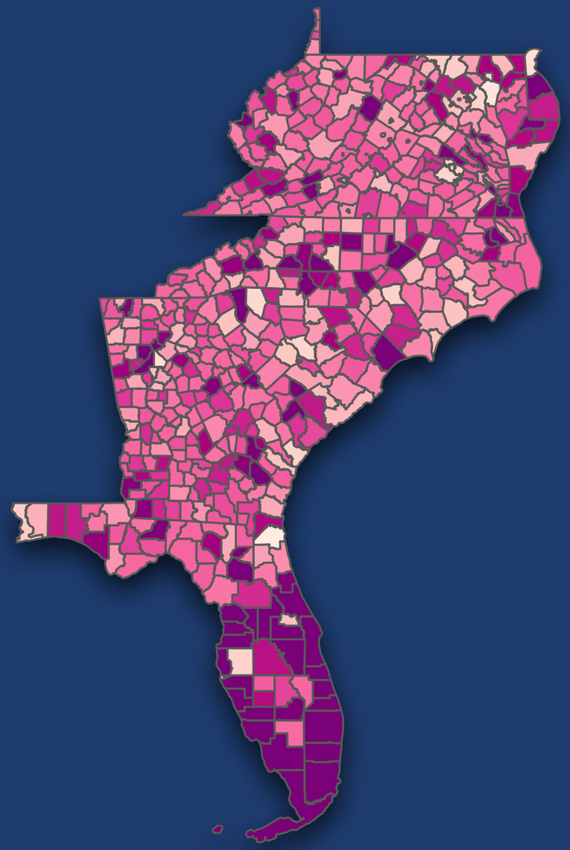


Spatial dependence!

Random Effects

-0.2 -0.1 0.0 0.1 0.2

Datta-Mandal

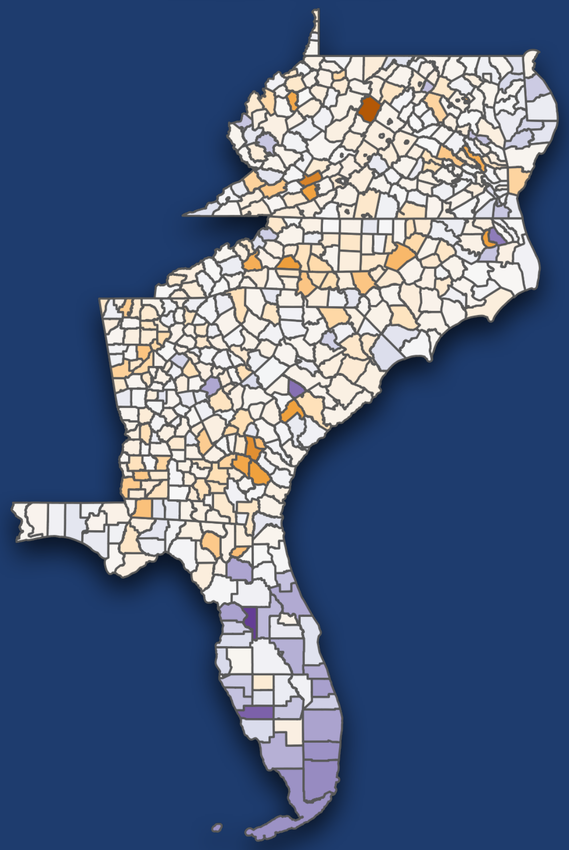


Selection Probabilities

0.2 0.4 0.6 0.8 1.0

DATA ANALYSIS: COMPARING DATTA-MANDAL VS. PROPOSED MODEL

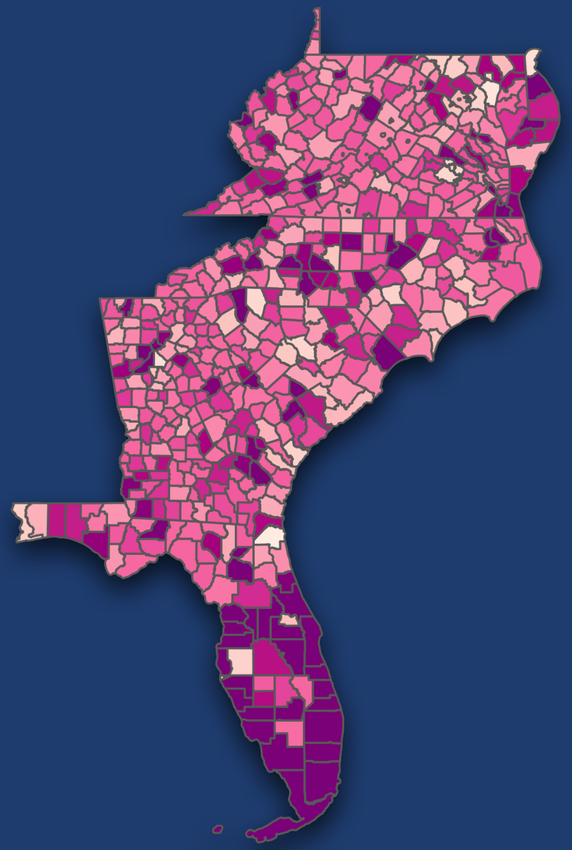
Datta-Mandal



Random Effects

-0.2 -0.1 0.0 0.1 0.2

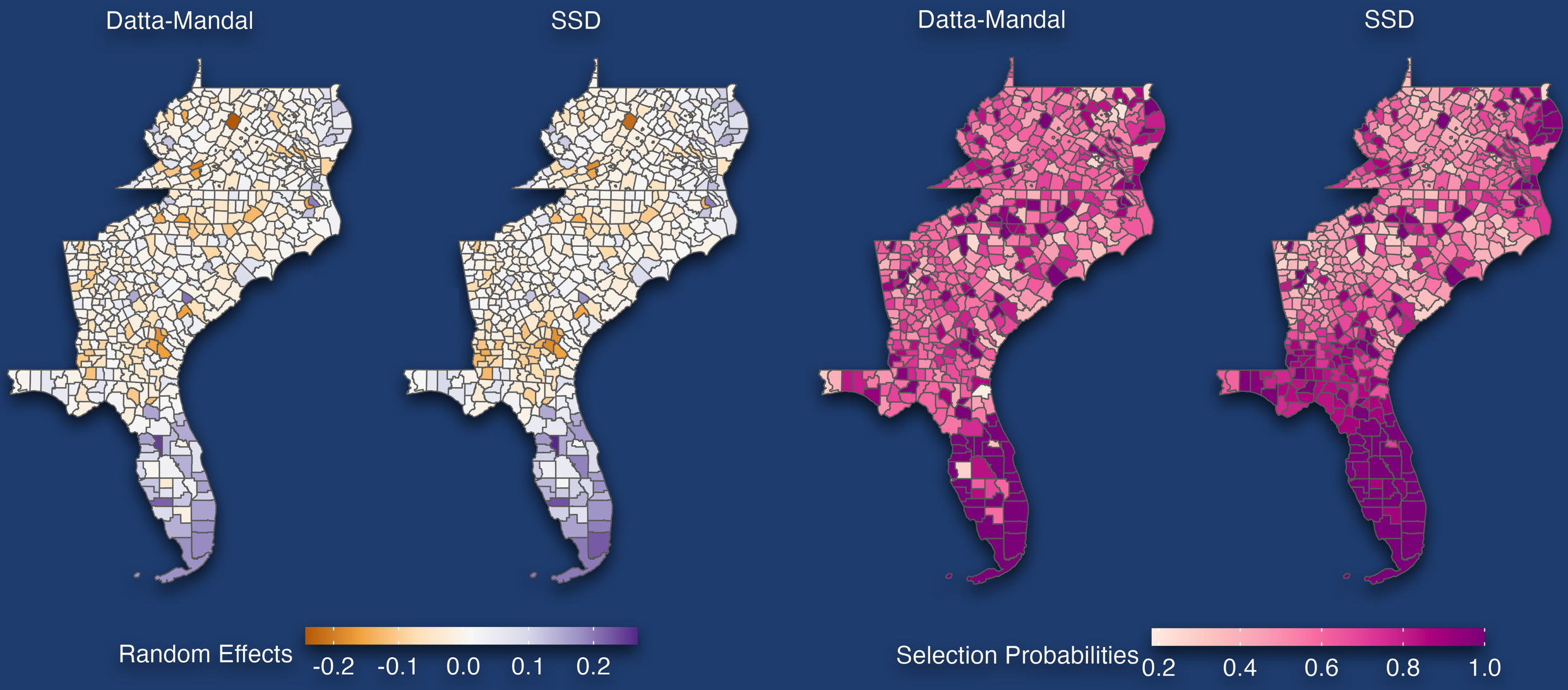
Datta-Mandal



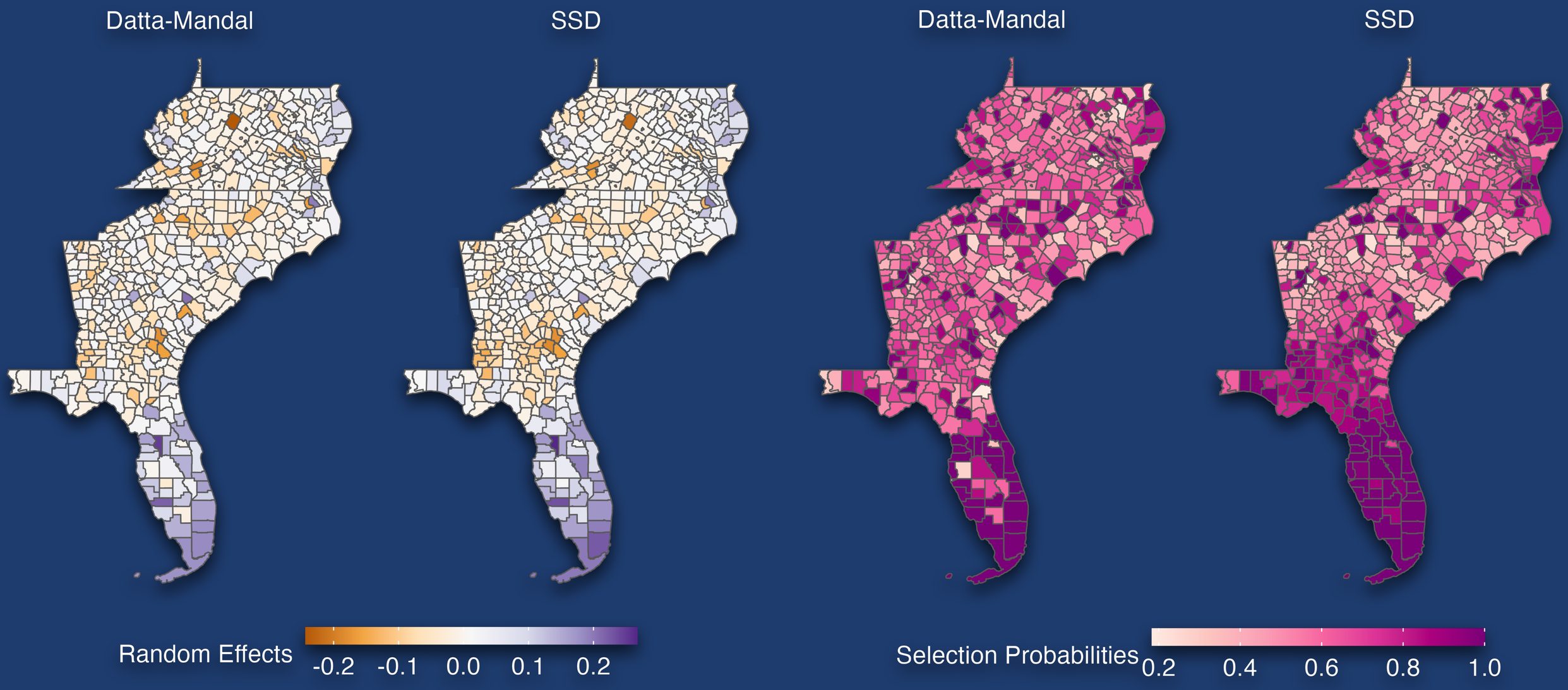
Selection Probabilities

0.2 0.4 0.6 0.8 1.0

DATA ANALYSIS: COMPARING DATTA-MANDAL VS. PROPOSED MODEL



DATA ANALYSIS: COMPARING DATTA-MANDAL VS. PROPOSED MODEL



EMPIRICAL SIMULATION STUDY

SIMULATION STUDY: SETUP

- Goal: create "empirical" simulation study that emulates real data
- *Data:* $\{\boldsymbol{z}, \boldsymbol{D}\}$ *from ACS, North Carolina (*$n = 100$*)*
- $\boldsymbol{\theta}$*: rent burden*
- *Covariates* $\boldsymbol{X}$ *also from ACS (education, race, poverty related)*

SIMULATION STUDY: SETUP

- Goal: create "empirical" simulation study that emulates real data
- *Data:* $\{\boldsymbol{z}, \boldsymbol{D}\}$ *from ACS, North Carolina (*$n = 100$*)*
- $\boldsymbol{\theta}$*: rent burden*
- *Covariates* $\boldsymbol{X}$ *also from ACS (education, race, poverty related)*

z_i

The "true"
small area means
(direct estimates)

SIMULATION STUDY: SETUP

- Goal: create "empirical" simulation study that emulates real data
- *Data:* $\{\boldsymbol{z}, \boldsymbol{D}\}$ *from ACS, North Carolina (*$n = 100$*)*
- $\boldsymbol{\theta}$*: rent burden*
- *Covariates* $\boldsymbol{X}$ *also from ACS (education, race, poverty related)*

z_i

The "true"
small area means
(direct estimates)

$y_i^{(g)} \overset{ind}{\sim} N\left(\log(z_i), d_i/z_i^2\right)$

Add noise using
survey variance
(log transformed)

SIMULATION STUDY: SETUP

- Goal: create "empirical" simulation study that emulates real data
- *Data:* $\{\boldsymbol{z}, \boldsymbol{D}\}$ *from ACS, North Carolina (*$n = 100$*)*
- $\boldsymbol{\theta}$*: rent burden*
- *Covariates* $\boldsymbol{X}$ *also from ACS (education, race, poverty related)*

z_i

The "true"
small area means
(direct estimates)

$y_i^{(g)} \overset{ind}{\sim} N\left(\log(z_i), d_i/z_i^2\right)$

Add noise using
survey variance
(log transformed)

$y_i^{(g)}$

Synthetic
"direct estimates"

SIMULATION STUDY: SETUP

- Goal: create "empirical" simulation study that emulates real data
- *Data:* $\{\boldsymbol{z}, \boldsymbol{D}\}$ *from ACS, North Carolina (*$n = 100$*)*
- $\boldsymbol{\theta}$*: rent burden*
- *Covariates* $\boldsymbol{X}$ *also from ACS (education, race, poverty related)*

z_i

The "true"
small area means
(direct estimates)

$y_i^{(g)} \overset{ind}{\sim} N\left(\log(z_i), d_i/z_i^2\right)$

Add noise using
survey variance
(log transformed)

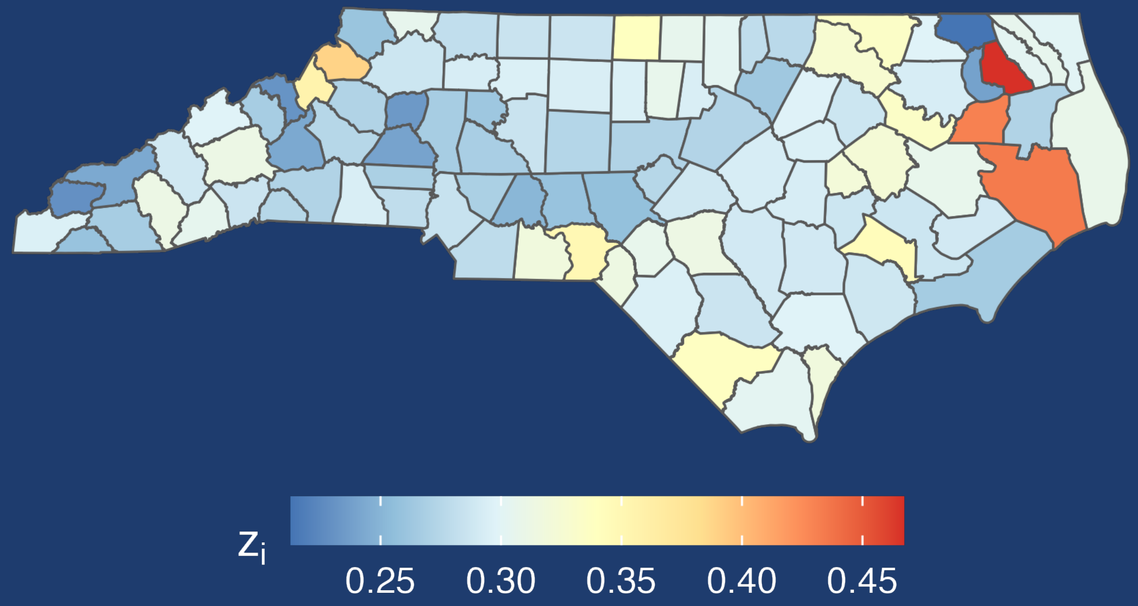
$y_i^{(g)}$

Synthetic
"direct estimates"

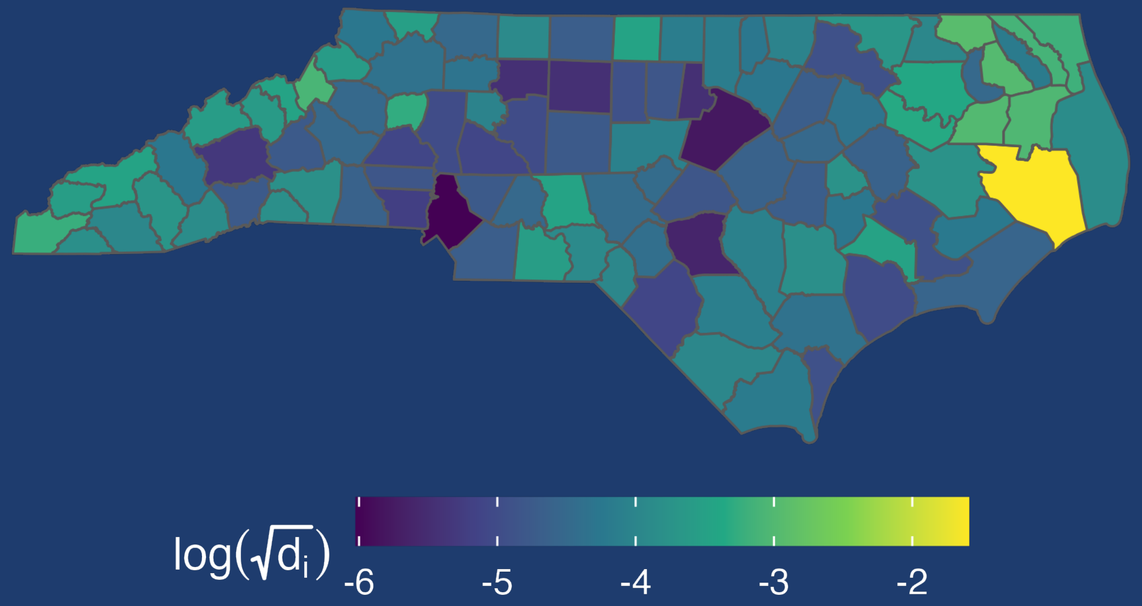
for simulations $g = 1, \cdots, 100$ *&* *areas* $i = 1, \cdots, n.$

SIMULATION STUDY: DATA

The 'True' Small Area Means: z_i

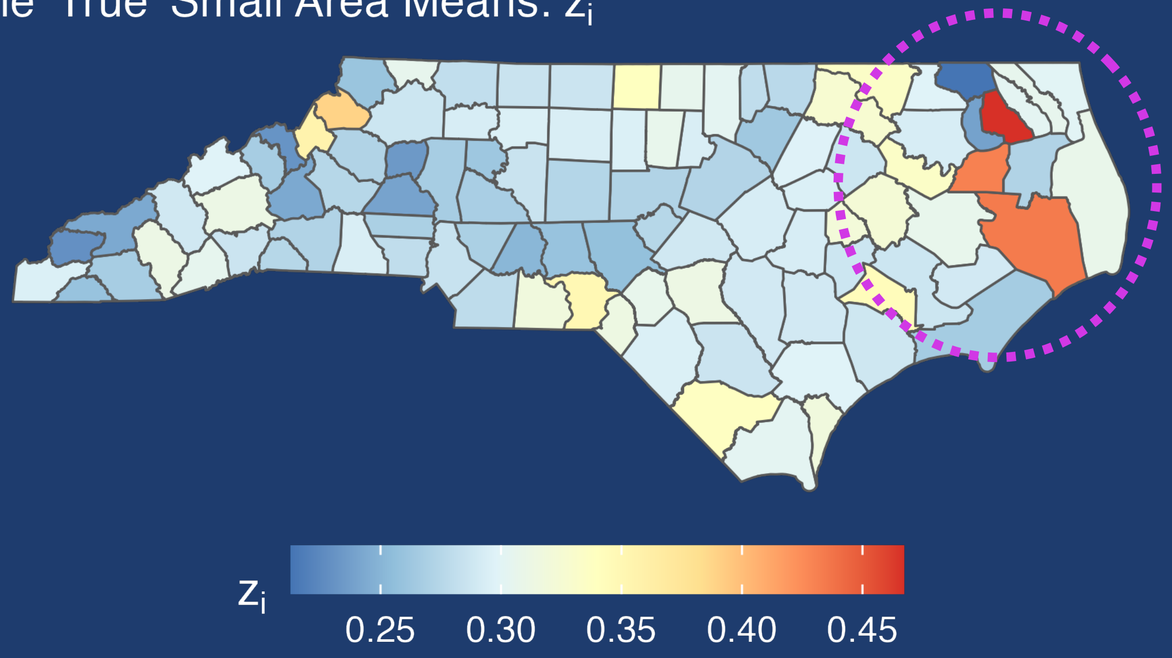


Log Survey Standard Errors

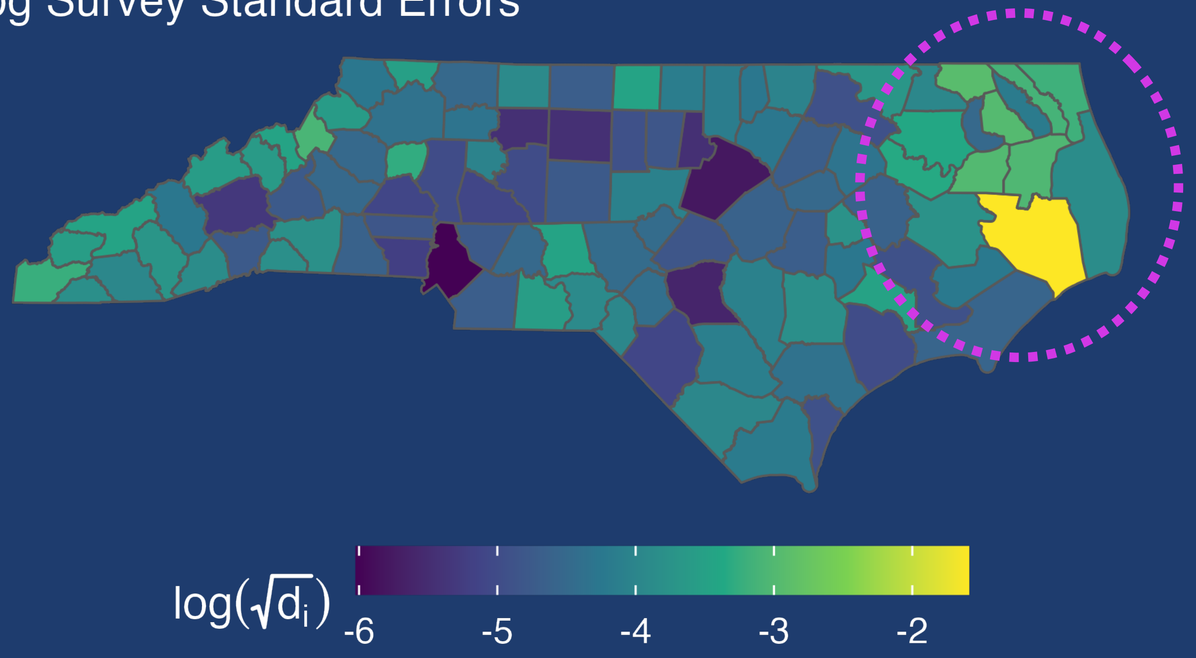


SIMULATION STUDY: DATA

The 'True' Small Area Means: z_i



Log Survey Standard Errors



SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE				
FAY-HERRIOT				
BYM				
DATTA-MANDAL				
SSD (PROPOSED)				

From $G = 100$ simulations

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	12.3×10^{-4}			
FAY-HERRIOT	6.8×10^{-4}			
BYM	6.9×10^{-4}			
DATTA-MANDAL	6.5×10^{-4}			
SSD (PROPOSED)	5.3×10^{-4}			

$$\frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_i^{(g)} - z_i)^2$$

From $G = 100$ simulations

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE REDUCTION SSD MODEL	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	57 %			
FAY-HERRIOT	22 %			
BYM	23 %			
DATTA-MANDAL	18 %			
SSD (PROPOSED)	*NA*			

$$\frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_i^{(g)}-z_i)^2$$

From $G = 100$ *simulations*

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	12.3×10^{-4}			
FAY-HERRIOT	6.8×10^{-4}			
BYM	6.9×10^{-4}			
DATTA-MANDAL	6.5×10^{-4}			
SSD (PROPOSED)	5.3×10^{-4}			

From $G = 100$ simulations

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	12.3×10^{-4}	*NA*		
FAY-HERRIOT	6.8×10^{-4}	0.836		
BYM	6.9×10^{-4}	0.831		
DATTA-MANDAL	6.5×10^{-4}	0.793		
SSD (PROPOSED)	5.3×10^{-4}	0.896		

$$\frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}\{\hat{l}_i^{(g)} < z_i\} \cdot \boldsymbol{I}\{z_i < \hat{u}_i^{(g)}\}$$

From $G = 100$ *simulations*

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	12.3×10^{-4}	*NA*		
FAY-HERRIOT	6.8×10^{-4}	0.836		
BYM	6.9×10^{-4}	0.831		
DATTA-MANDAL	6.5×10^{-4}	0.793		
SSD (PROPOSED)	5.3×10^{-4}	0.896		

$$\frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i^{(g)}-\hat{l}_i^{(g)})+\frac{2}{\alpha}(\hat{l}_i^{(g)}-z_i)\boldsymbol{I}\{\hat{l}_i^{(g)}>z_i\}+\frac{2}{\alpha}(z_i-\hat{u}_i^{(g)})\boldsymbol{I}\{z_i>\hat{u}_i^{(g)}\}$$

From $G = 100$ simulations

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	12.3×10^{-4}	*NA*	*NA*	
FAY-HERRIOT	6.8×10^{-4}	0.836	0.109	
BYM	6.9×10^{-4}	0.831	0.115	
DATTA-MANDAL	6.5×10^{-4}	0.793	0.096	
SSD (PROPOSED)	5.3×10^{-4}	0.896	0.076	

$$\frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i^{(g)}-\hat{l}_i^{(g)})+\frac{2}{\alpha}(\hat{l}_i^{(g)}-z_i)\boldsymbol{I}\{\hat{l}_i^{(g)}>z_i\}+\frac{2}{\alpha}(z_i-\hat{u}_i^{(g)})\boldsymbol{I}\{z_i>\hat{u}_i^{(g)}\}$$

From $G = 100$ simulations

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	12.3×10^{-4}	*NA*	*NA*	2.1×10^{-3}
FAY-HERRIOT	6.8×10^{-4}	0.836	0.109	10.8×10^{-3}
BYM	6.9×10^{-4}	0.831	0.115	11.1×10^{-3}
DATTA-MANDAL	6.5×10^{-4}	0.793	0.096	11.0×10^{-3}
SSD (PROPOSED)	5.3×10^{-4}	0.896	0.076	8.6×10^{-3}

$$\frac{1}{n}\sum_{i=1}^{n}\left|z_i - \frac{1}{G}\sum_{g=1}^{G}\hat{z}_i^{(g)}\right|$$

From $G = 100$ simulations

SIMULATION STUDY: RESULTS

ESTIMATOR	MSE	90% CREDIBLE INTERVALS		ABSOLUTE BIAS
		COVERAGE RATE	INTERVAL SCORE	
DIRECT ESTIMATE	12.3×10^{-4}	*NA*	*NA*	2.1×10^{-3}
FAY-HERRIOT	6.8×10^{-4}	0.836	0.109	10.8×10^{-3}
BYM	6.9×10^{-4}	0.831	0.115	11.1×10^{-3}
DATTA-MANDAL	6.5×10^{-4}	0.793	0.096	11.0×10^{-3}
SSD (PROPOSED)	5.3×10^{-4}	0.896	0.076	8.6×10^{-3}

From $G = 100$ simulations

SUMMARY

In essence, we created a *Small Area Estimation* model that combines…

SUMMARY

In essence, we created a *Small Area Estimation* model that combines…

BAYESIAN VARIABLE SELECTION

SUMMARY

In essence, we created a *Small Area Estimation* model that combines…

BAYESIAN VARIABLE SELECTION

+

SPATIAL DATA MODELING

SUMMARY

In essence, we created a *Small Area Estimation* model that combines…

BAYESIAN VARIABLE SELECTION

+

SPATIAL DATA MODELING

Building in spatial selection & dependence can improve predictive performance of both point & interval estimates

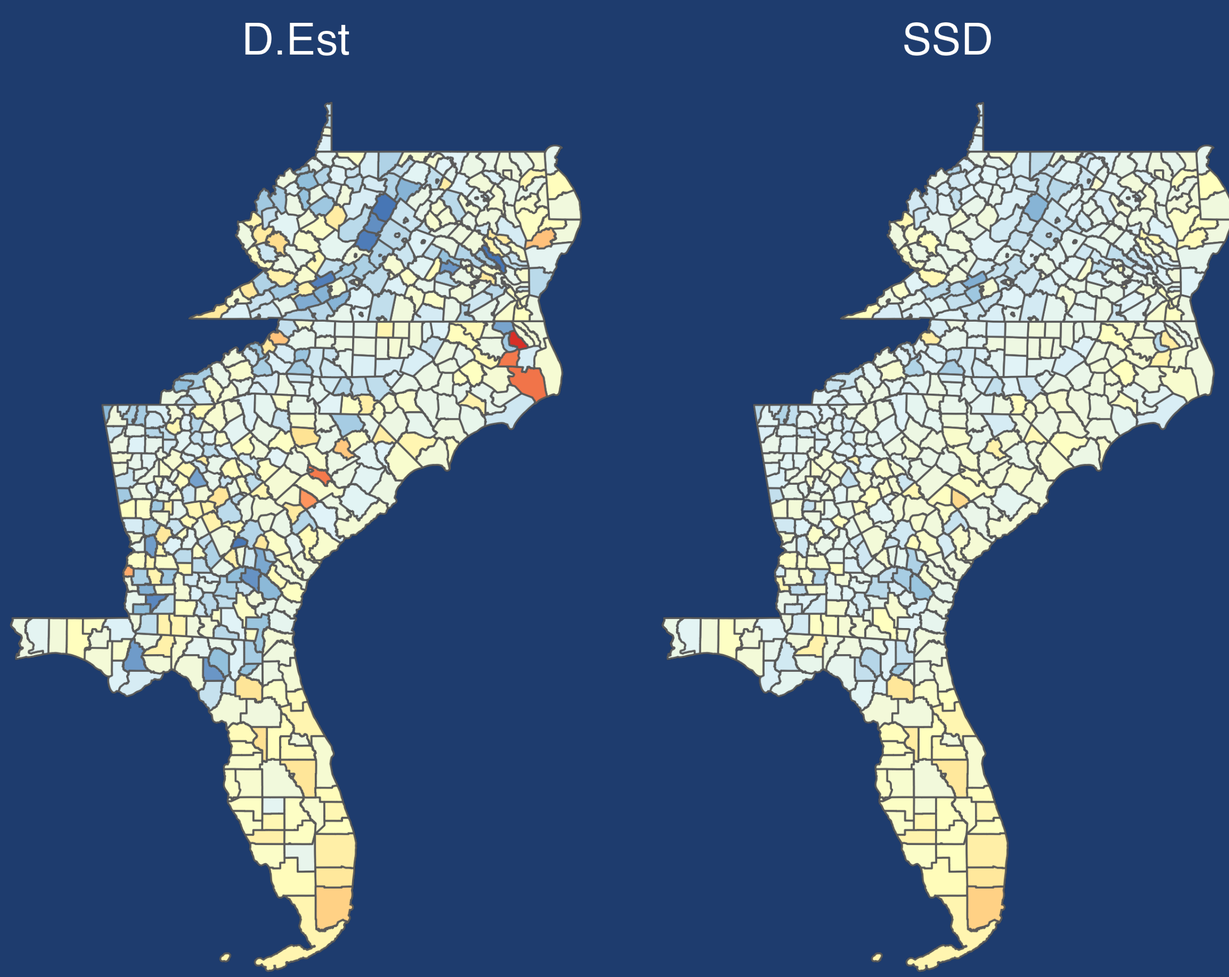
QUESTIONS?

THANK YOU!

REFERENCES

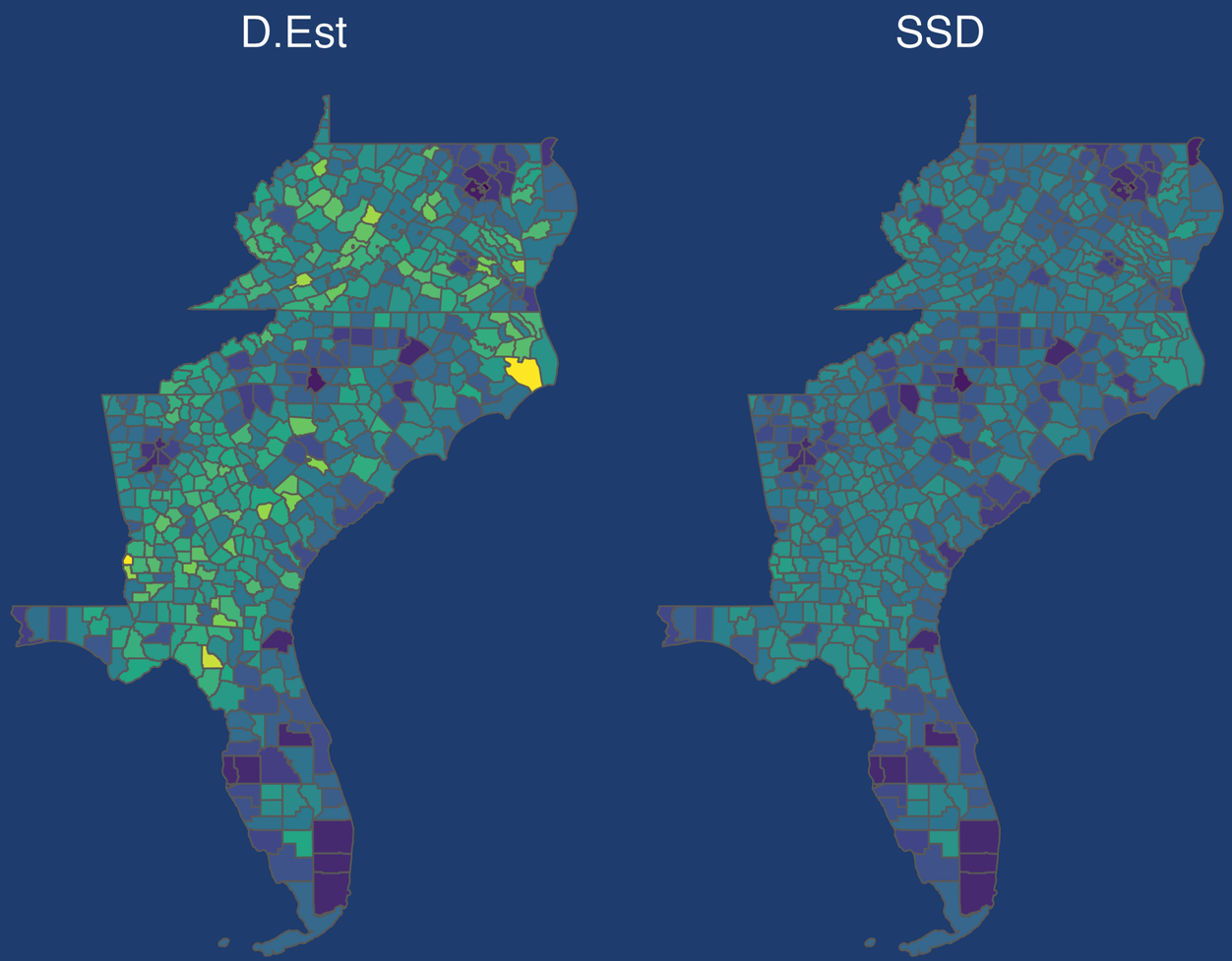
- Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," Journal of the Royal Statistical Society: Series B (Methodological), 36, 192–225. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x.
- Besag, J., York, J., and Mollié, A. (1991), "Bayesian image restoration, with two applications in spatial statistics," Annals of the Institute of Statistical Mathematics, 43, 1–20. https://doi.org/10.1007/BF00116466.
- Datta, G. S., Hall, P., and Mandal, A. (2011), "Model Selection by Testing for the Presence of Small-Area Effects, and Application to Area-Level Data," Journal of the American Statistical Association, 106, 362–374. https://doi.org/10.1198/jasa.2011.tm10036.
- Datta, G. S., and Mandal, A. (2015), "Small Area Estimation With Uncertain Random Effects," Journal of the American Statistical Association, 110, 1735–1744. https://doi.org/10.1080/01621459.2015.1016526.
- Fay, 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, 269–277. https://doi.org/10.1080/01621459.1979.10482505.
- Kawano, S., Parker, P. A., and Li, Z. R. (2024), "Spatially Selected and Dependent Random Effects for Small Area Estimation with Application to Rent Burden," arXiv.
- Leroux, B. G., Lei, X., and Breslow, N. (2000), "Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence," in Statistical Models in Epidemiology, the Environment, and Clinical Trials, eds. M. E. Halloran and D. Berry, New York, NY: Springer, pp. 179–191. https://doi.org/10.1007/978-1-4612-1284-3_4.
- 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, Taylor & Francis, 108, 1339–1349. https://doi.org/10.1080/01621459.2013.829001.
- Sørbye, S. H., and Rue, H. (2014), "Scaling intrinsic Gaussian Markov random field priors in spatial modelling," Spatial Statistics, Spatial Statistics Miami, 8, 39–51. https://doi.org/10.1016/j.spasta.2013.06.004.
- Wakefield, J. (2007), "Disease mapping and spatial regression with count data," Biostatistics (Oxford, England), 8, 158–183. https://doi.org/10.1093/biostatistics/kxl008.

ESTIMATES & SE: SURVEY VS. PROPOSED MODEL



Rent Burden Estimates

0.2 0.3 0.4

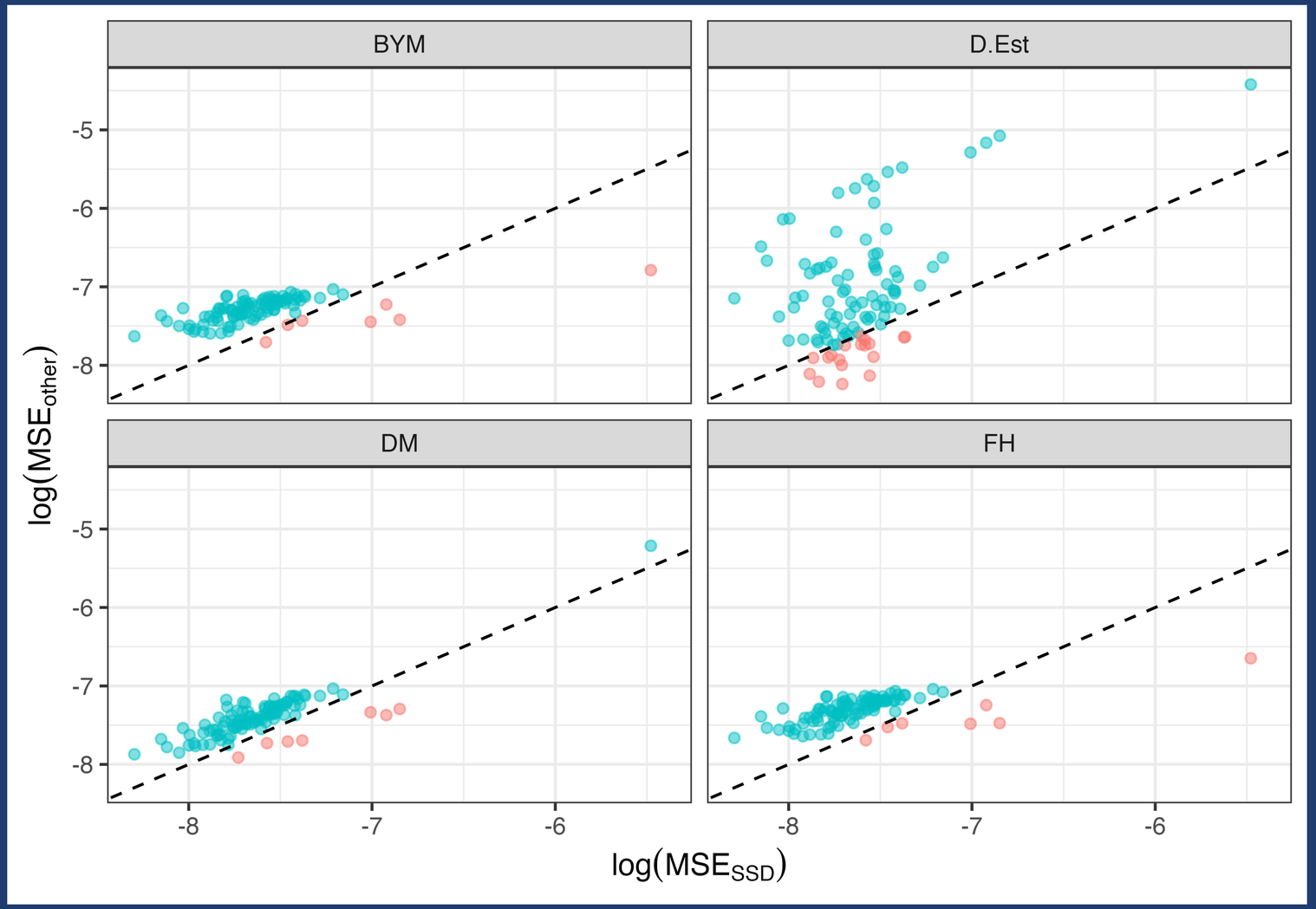


Log SE's

-6 -5 -4 -3 -2

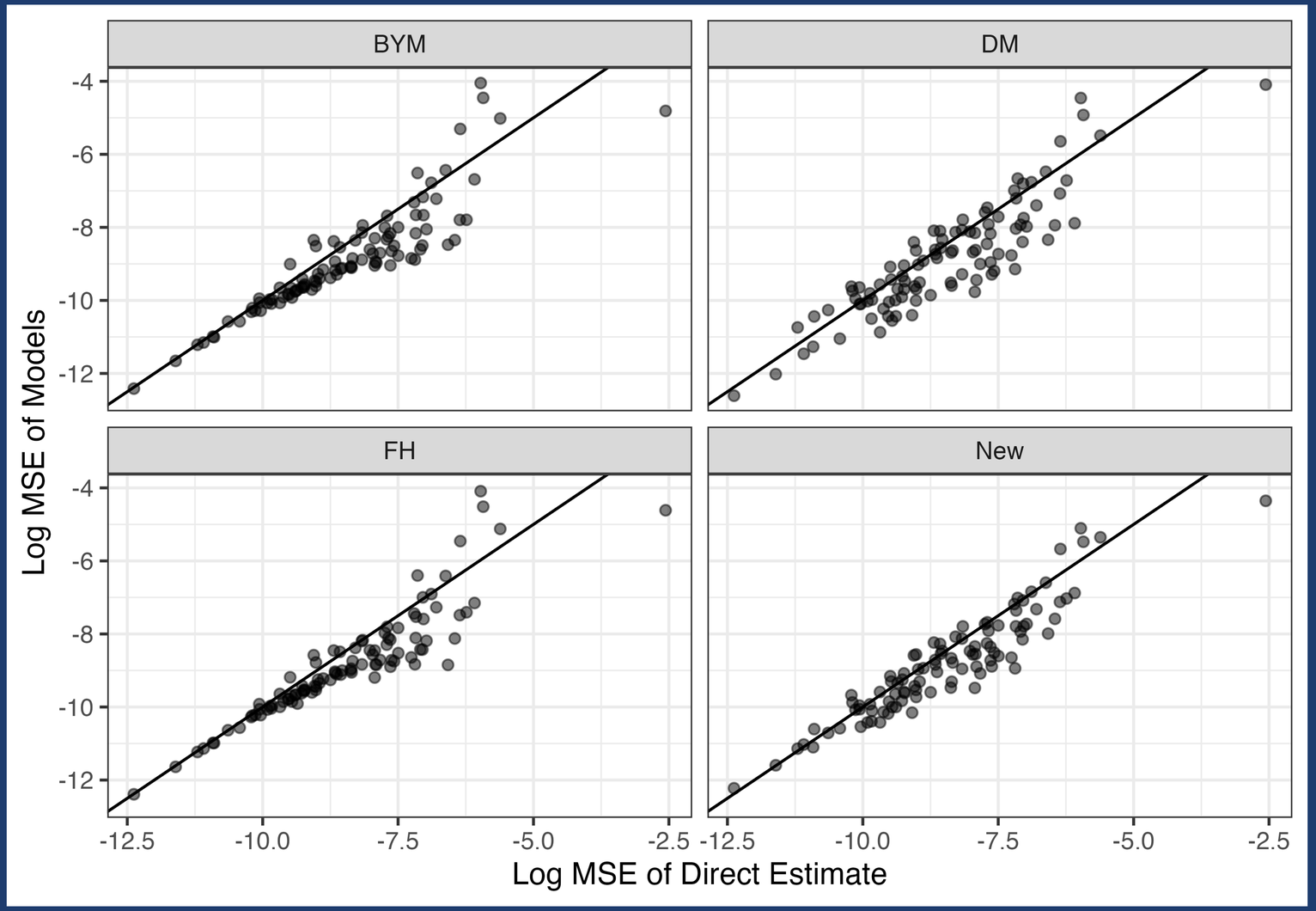
SIM. STUDY: LOG MSE ACROSS 100 **SIMULATIONS**

OTHER MODELS



PROPOSED SSD MODEL

SIMULATION STUDY: COMPARE MSE ACROSS **AREAS**



ADDRESSING ROBUSTNESS

- IID random effects $\boldsymbol{u} \sim N_n(0, \sigma^2 I)$ is not robust to outliers
  - Assume heavier-tailed distribution for $\boldsymbol{u}$ (ex: Datta & Lahiri, 1995)
  - Use a mixture model (Chakraborty et al., 2016)
  - Use Bayesian nonparametrics (Janicki et al., 2022)

GLOBAL-LOCAL RANDOM EFFECTS

- Tang et al. (2018):

  $[u_i \,|\, \lambda_i, \tau^2] \sim N(0, \lambda_i^2 \tau^2)$ with $\pi(\boldsymbol{\lambda}, \tau^2) \propto \pi(\tau^2) \prod_{i=1}^{n} \pi(\lambda_i^2)$ for areas $i = 1, \cdots, n$

  There are restrictions on the prior for λ_i —> requires a copula for modeling spatial dependence for $\boldsymbol{\lambda}$

- Tang et al. (2023) uses CAR for the random effects but not the shrinkage:

  $[\boldsymbol{u} \,|\, \rho, \boldsymbol{\lambda}, \tau] \sim N_n(\mathbf{0}, \tau^2 \boldsymbol{\Lambda} Q^{-1} \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = diag\{\lambda_i\}_{i=1}^{n}$ and Q is the CAR precision matrix

POSTERIOR SELECTION PROBABILITIES

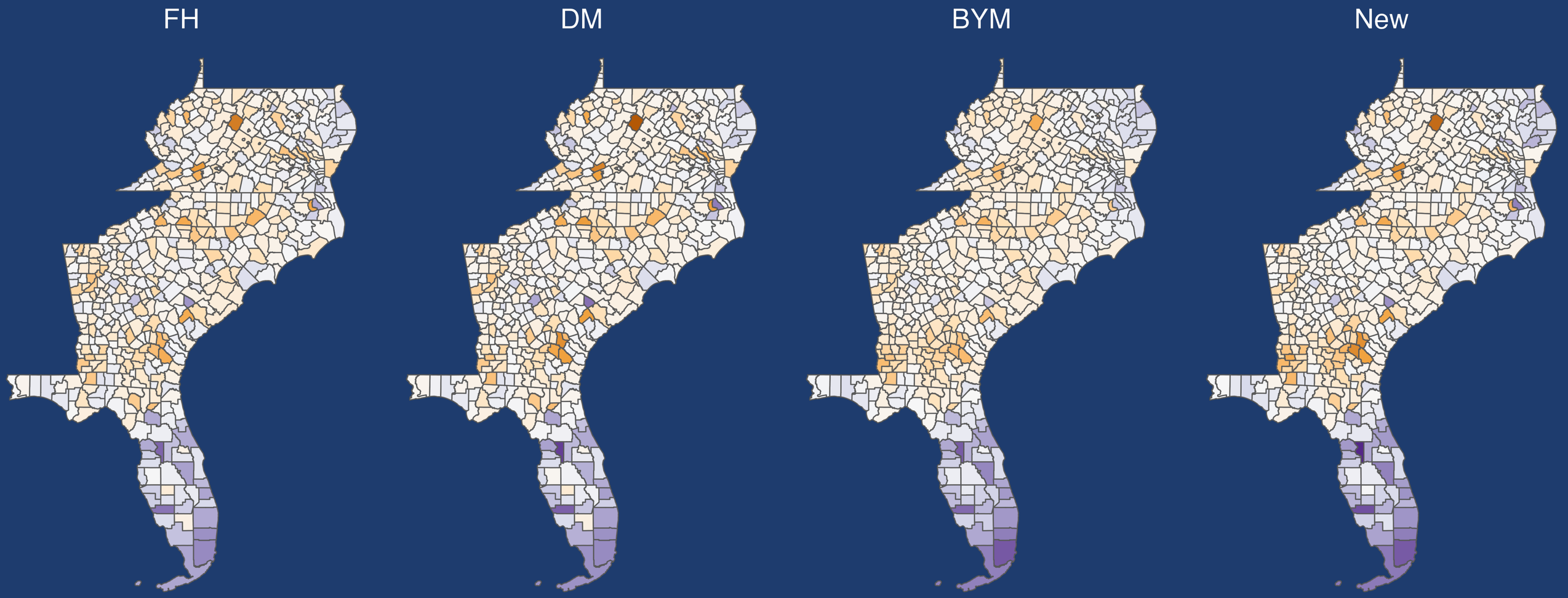
- Both models have posterior conditional: $[\delta_i \,|\, \tilde{p}_i, data] \sim Bern(\tilde{p}_i)$
- Datta-Mandal Spike-and-Slab:

$$\tilde{p}_i = \frac{p \cdot \phi\left(y_i \,|\, x_i^\top \boldsymbol{\beta} + v_i, d_i\right)}{p \cdot \phi\left(y_i \,|\, x_i^\top \boldsymbol{\beta} + v_i, d_i\right) + (1-p) \cdot \phi\left(y_i \,|\, x_i^\top \boldsymbol{\beta}, d_i\right)}$$

- Proposed SSD Model:

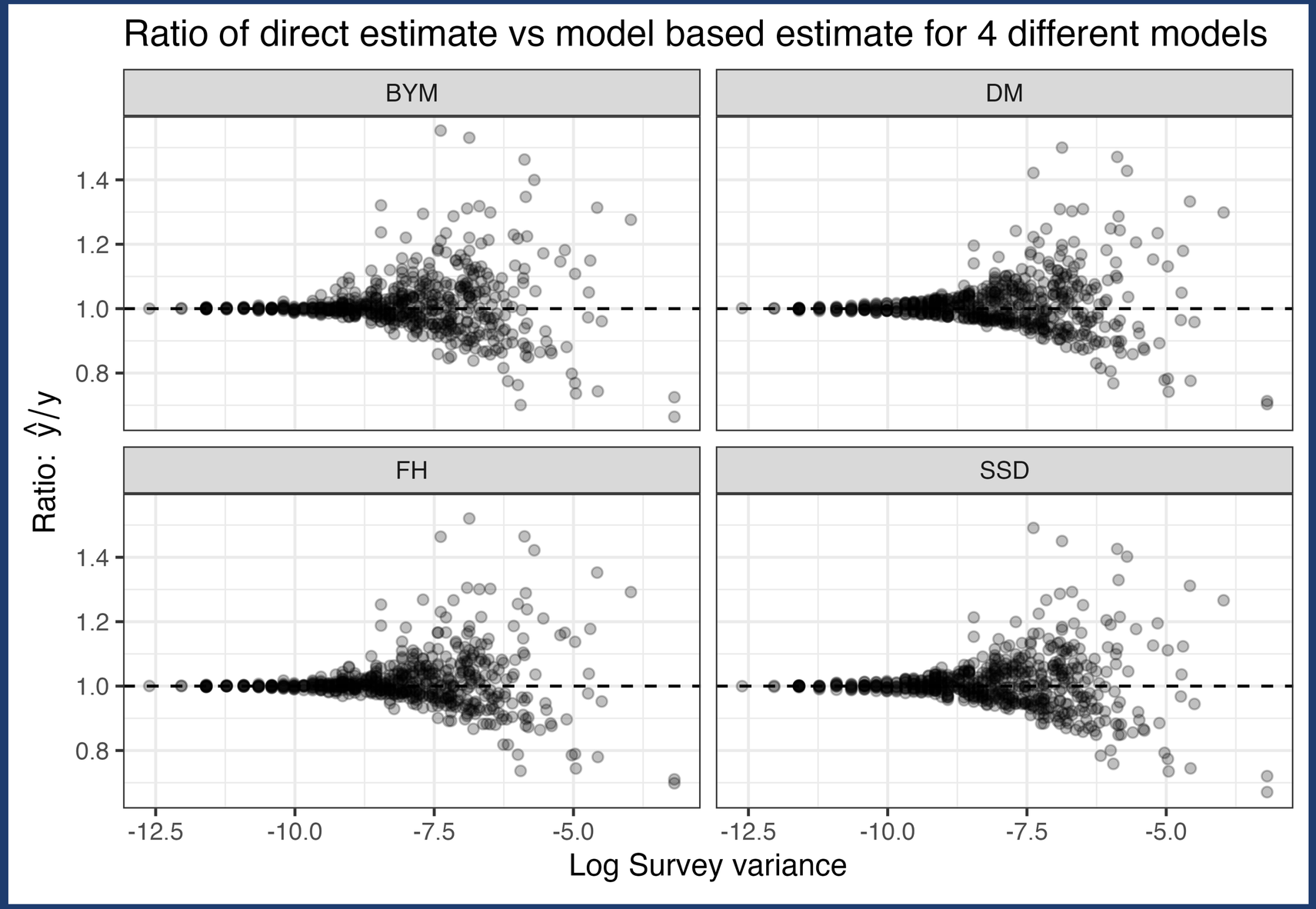
$$\tilde{p}_i = \frac{p_i \cdot \phi\left(y_i \,|\, x_i^\top \boldsymbol{\beta} + v_{1i} + v_{2i}, d_i\right)}{p_i \cdot \phi\left(y_i \,|\, x_i^\top \boldsymbol{\beta} + v_{1i} + v_{2i}, d_i\right) + (1-p_i) \cdot \phi\left(y_i \,|\, x_i^\top \boldsymbol{\beta}, d_i\right)}$$

RANDOM EFFECTS: ALL MODELS

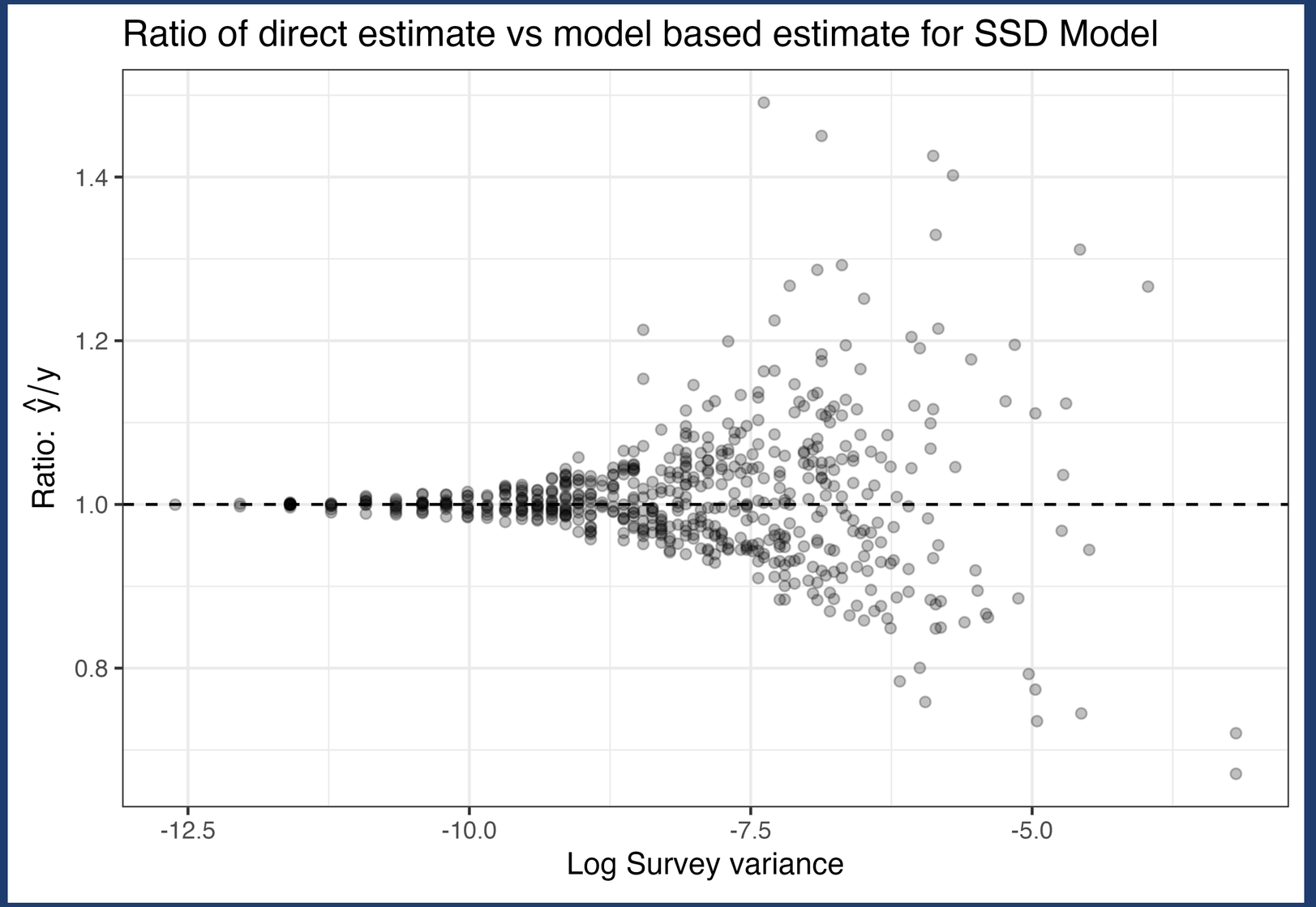


Comparing posterior means

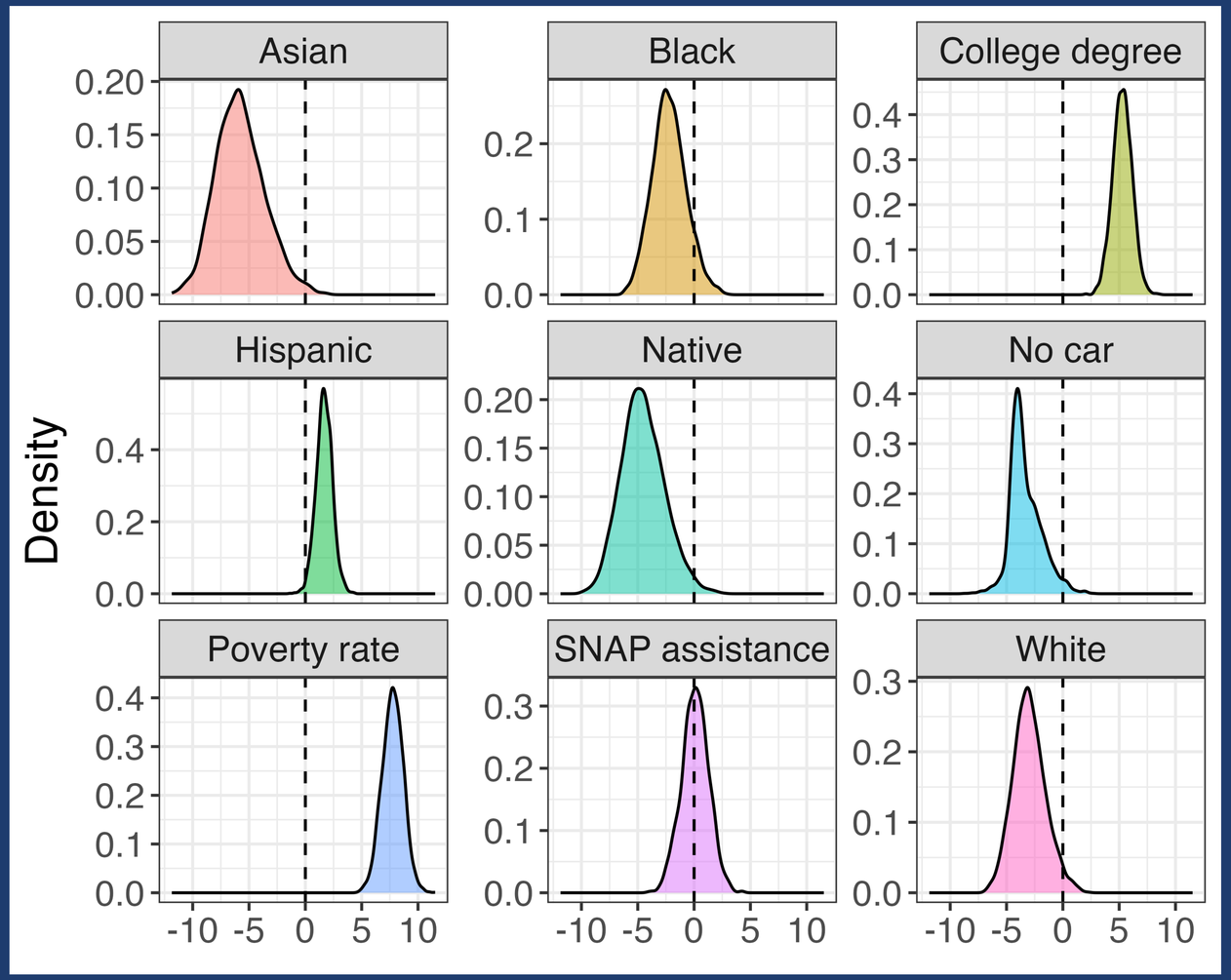
Data: 2015-2019 ACS from South Atlantic Division



Data: 2015-2019 ACS from South Atlantic Division



POSTERIOR INFERENCE ON β WITH SSD MODEL



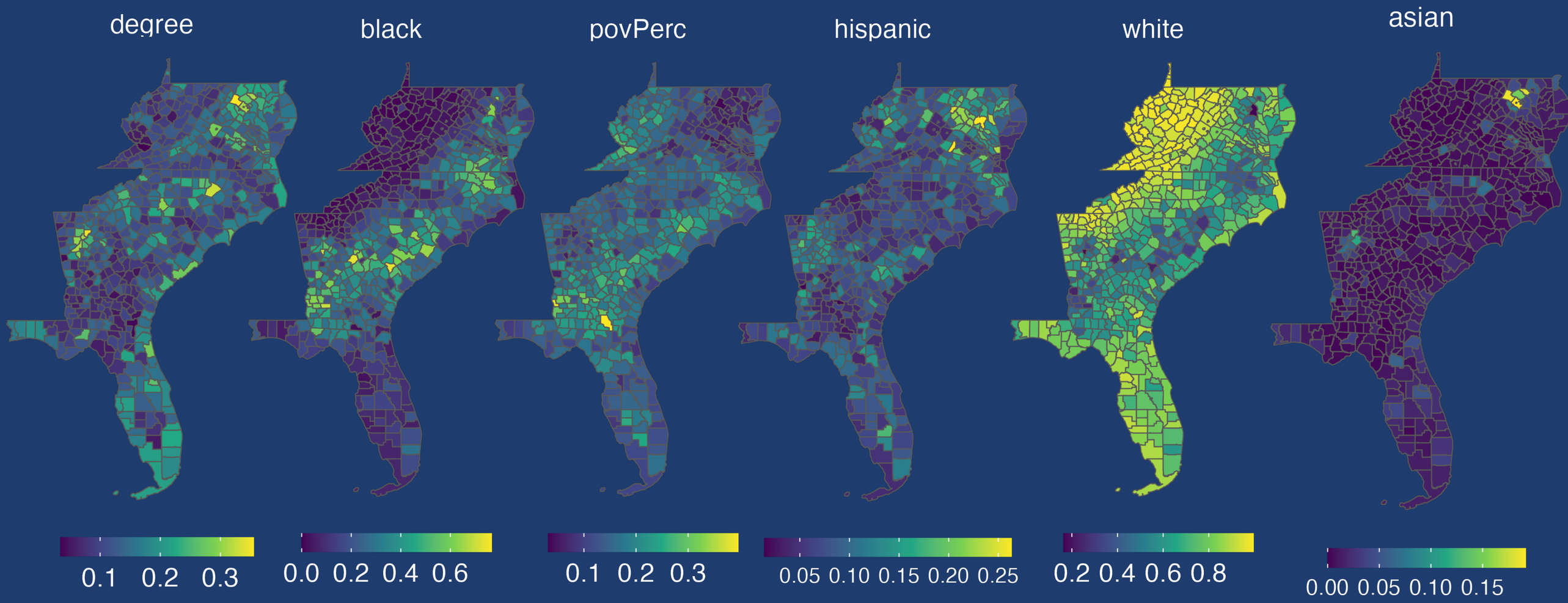
Posterior Densities of β for the 9 coefficients

Covariate	CI.lwr	Mean	CI.upr
Asian	-9.0515	-5.7632	-1.9794
Black	-4.5666	-2.2085	0.3748
College degree	3.8490	5.2784	6.6507
Hispanic	0.4862	1.6799	2.8459
Native	-7.5545	-4.5383	-1.2449
No car	-5.0267	-3.3149	-0.6819
Poverty rate	6.2090	7.7522	9.2575
SNAP assistance	-1.9553	0.0977	2.0137
White	-5.2108	-2.9752	-0.4263

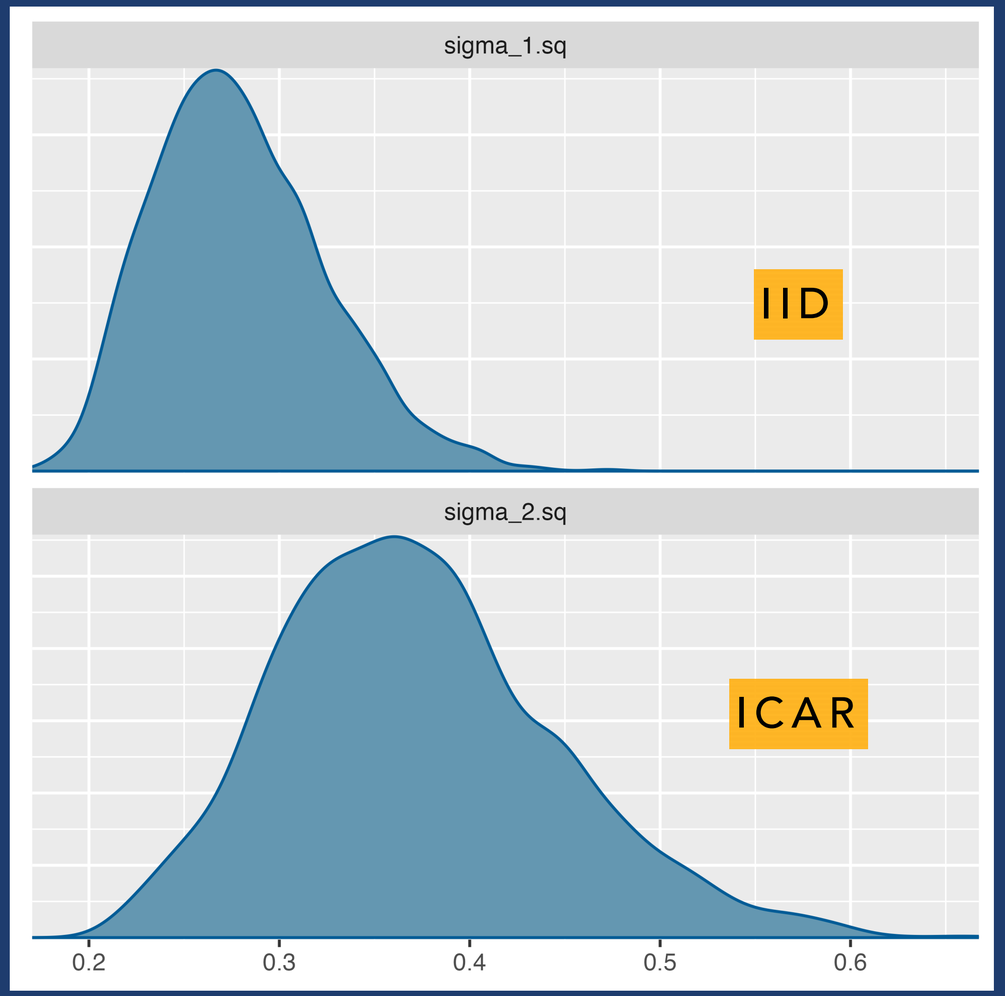
Posterior Summary (90% Credible Interval)

MAP OF SELECTED COVARIATES

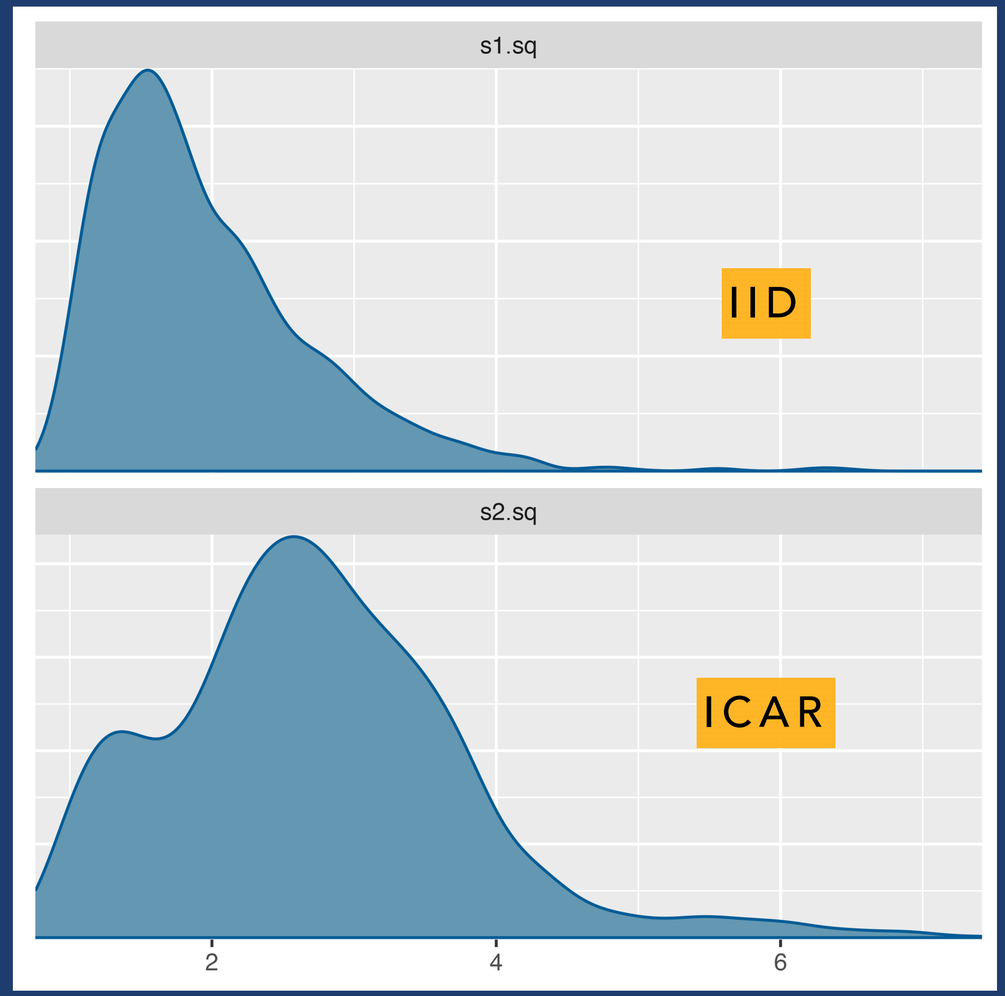
ACS Data from South Atlantic Census Division ($n = 588$)



POSTERIOR INFERENCE: IID VS. SPATIAL VARIANCES



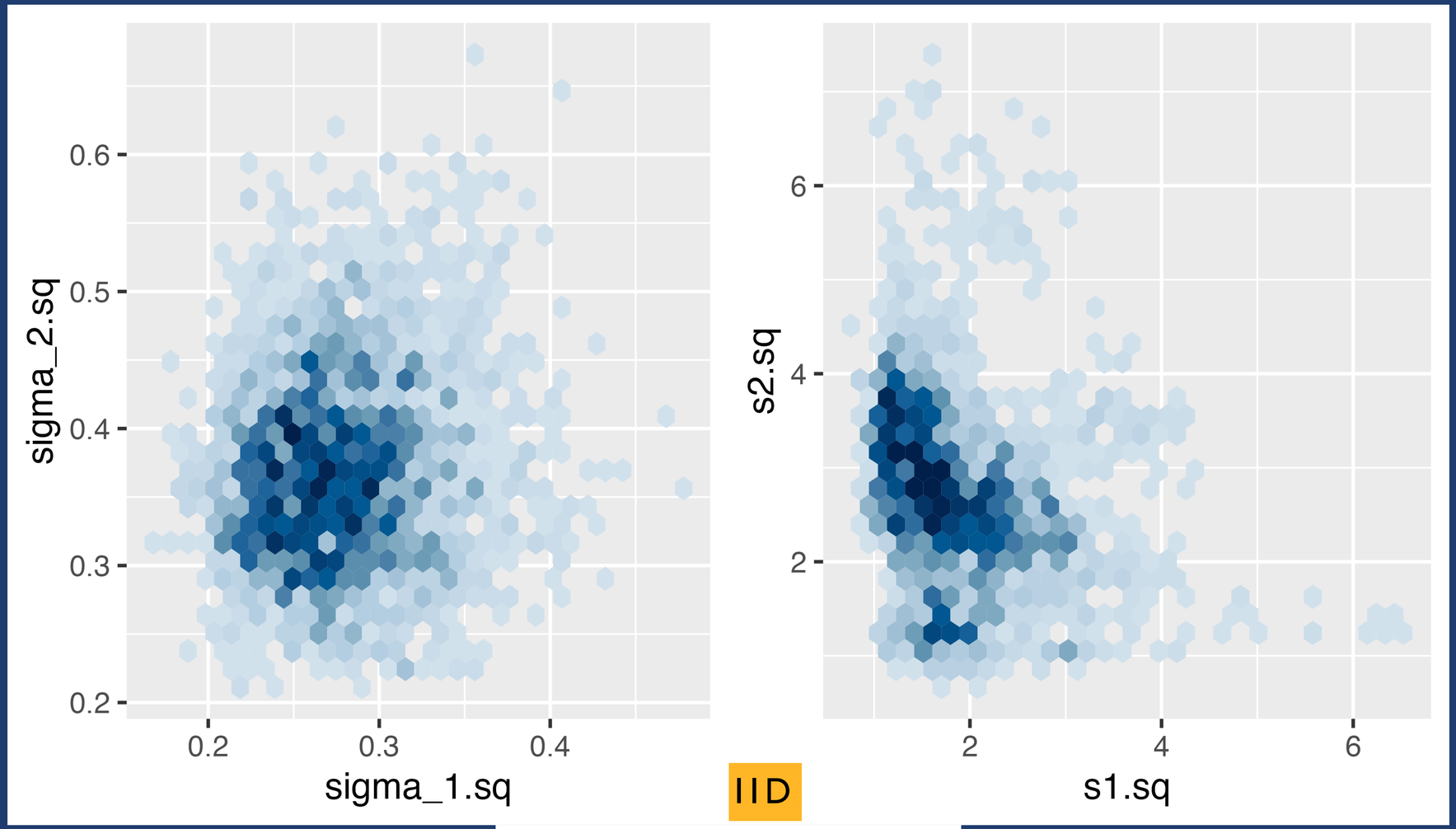
Posterior Densities of Random Effect Variances: σ_1^2, σ_2^2



Posterior Densities of Logit Variances: s_1^2, s_2^2

POSTERIOR INFERENCE: IID VS. SPATIAL VARIANCES

ICAR



ACS Data from South Atlantic Census Division ($n = 588$)

Random Effect Variances: σ_1^2, σ_2^2

Logit Variances: s_1^2, s_2^2

FAY-HERRIOT MODEL: BORROWING STRENGTH

- Problem: y_i can be unreliable with high d_i for small sample regions
- Solution: use a model that "borrow strength" across areas
- Example: FH model with prior $p(\beta, \sigma^2) \propto 1$

$$E[\theta_i \mid \beta, \sigma^2] = \frac{\sigma^2}{d_i + \sigma^2} y_i + \frac{d_i}{d_i + \sigma^2} \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

When d_i is small / y_i is a good estimator → y_i is given more weight

FAY-HERRIOT MODEL: BORROWING STRENGTH

- Problem: y_i can be unreliable with high d_i for small sample regions
- Solution: use a model that "borrow strength" across areas
- Example: FH model with prior $p(\beta, \sigma^2) \propto 1$

$$E[\theta_i \mid \beta, \sigma^2] = \frac{\sigma^2}{d_i + \sigma^2} y_i + \frac{d_i}{d_i + \sigma^2} \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

When d_i is large
y_i is unreliable

$\boldsymbol{x}_i^\top \boldsymbol{\beta}$ is given more weight