A Socio-Demographic Latent Space Approach to Spatial Data When Geography is Important but not All-Important

Are physical neighbors also social neighbors?

Saikat Nandy [1], **Scott H. Holan** [2,3], and **Michael Schweinberger** [4]

[1]Department of Biostatistics, St. Jude Children's Research Hospital.
[2]Department of Statistics, University of Missouri-Columbia
[3]Office of the Associate Director for Research and Methodology, U.S. Census Bureau
[4]Department of Statistics, The Pennsylvania State University

Outline

Neighborhood Network in Spatial Statistics

First Law of Geography

"Everything is related to everything else. But near things are more related than distant things." (Tobler, 1970).

Existing Network Structures

- Popular methods to model spatial-temporal dependency
  I. Conditionally Autoregressive (CAR) models (Besag, 1974),
  II. Intrinsic Autoregressive (IAR) models (Besag et al., 1991),
  III. Vector Autoregressive (VAR) models (Cressie and Wikle, 2011).
- These methods are dependent on adjacency matrices based on nearest neighbor approach.
- In a regular lattice, i and j are nearest neighbors, if they share an edge between them.
- Are geographic neighbors also social neighbors?

A Motivating Example - Single State

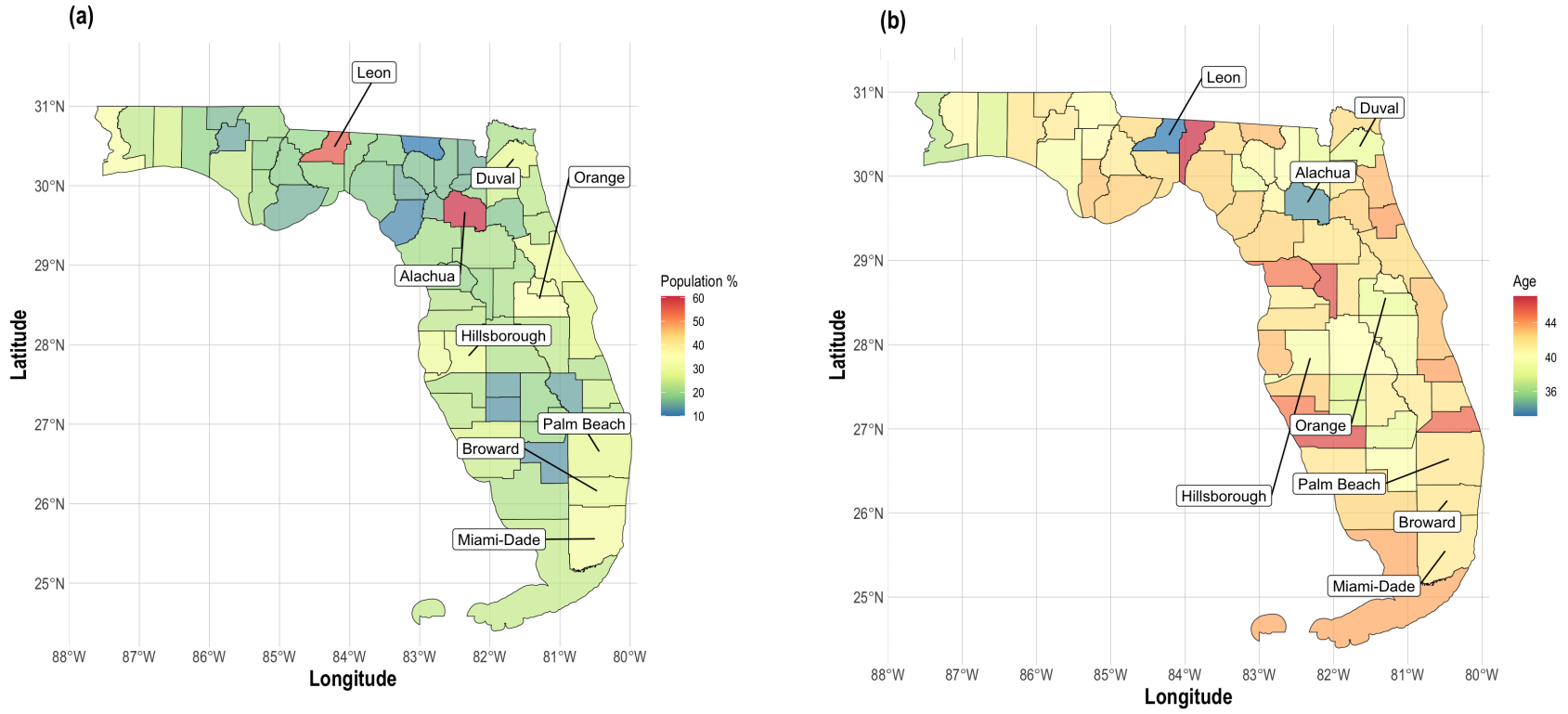


Figure 1: 2019 ACS design-based estimates from counties in Florida: (a) the percentage of the population enrolled in college or graduate school, (b) median age of the labor force.

A Motivating Example - Multiple States

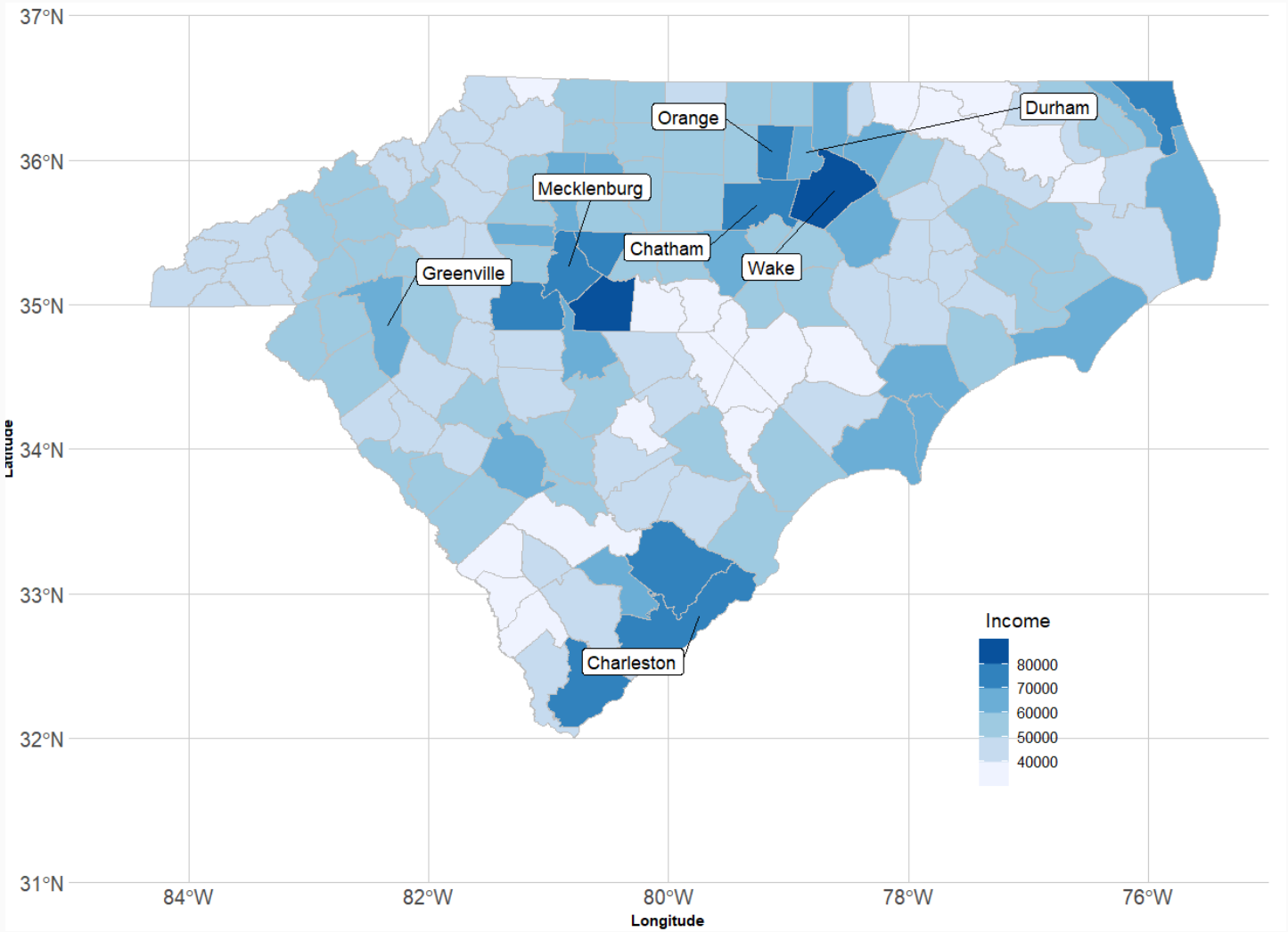


Figure 2: 2022 ACS design-based estimates of median household income in the past 12 months from counties in North and South Carolina.

New Neighborhood Network

A New Neighborhood Model

Conditional on the positions $Z_i = z_i$ and $Z_j = z_j$ of spatial units $i \in \mathcal{D}$ and $j \in \mathcal{D}$ in $\mathbb{R}^2$, neighborhood indicators $B_{i,j} \in \{0, 1\}$ are generated by drawing

$$B_{i,j} \mid \alpha, \gamma, Z_i = z_i, Z_j = z_j \overset{ind.}{\sim} \text{Bernoulli}(p_{i,j}), \tag{1}$$

where the log odds of the conditional probability $p_{i,j}$ of the event $\{B_{i,j} = 1\}$ is given by

$$\text{logit}(p_{i,j}) \quad = \quad \alpha - \gamma\, d_1(i,j) - (1-\gamma)\, d_2(i,j). \tag{2}$$

- $d_1(i,j)$ and $d_2(i,j)$ represent the geographical and socio-demographic distance between spatial units
- $\gamma \in [0, 1]$ specifies the relative importance of $d_1(i,j)$ relative to $d_2(i,j)$.
- intercept $\alpha \in \mathbb{R}$ controls the expected number of neighbors when the two terms $d_1(i,j)$ and $d_2(i,j)$ vanish,

A DAG representation

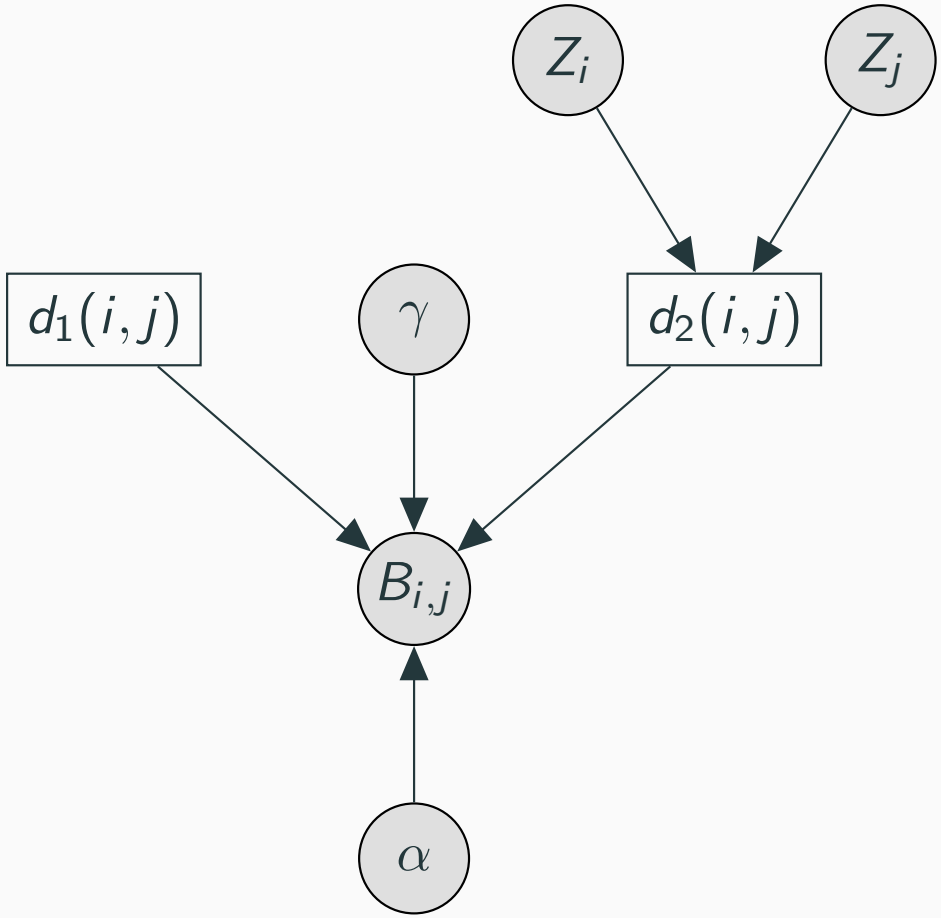


Figure 3: The conditional independence structure of the neighborhood model that generates neighborhood indicators $B_{i,j}$. Unshaded circles indicate observable random variables, while shaded circles indicate unobservable random variables. Rectangles represent either known constants or known functions of random variables.

NNSD Network model

The neighborhood model is defined as follows:

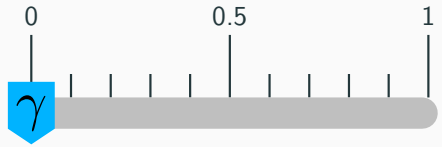
$$
\begin{aligned}
&B_{i,j} \mid \alpha,\, \gamma,\, Z_i = z_i,\, Z_j = z_j \overset{ind.}{\sim} \text{Bernoulli}(p_{i,j}),\\
&\text{logit}(p_{i,j}) = \alpha - \gamma\, d_1(i,j) - (1-\gamma)\, d_2(i,j)\\
&\alpha \mid \sigma_\alpha^2 \;\sim\; \text{Normal}(0,\, \sigma_\alpha^2)\\
&\gamma \;\sim\; \text{Uniform}(0,\, 1)\\
&\mathbf{Z}_i \mid \mathbf{S}_i,\, \boldsymbol{\delta},\, \sigma_z^2 \overset{ind.}{\sim} \text{MVN}_2(\mathbf{S}_i\, \boldsymbol{\delta},\, \sigma_z^2\, \mathbf{I}_2)\\
&\boldsymbol{\delta} \mid \sigma_\delta^2 \;\sim\; \text{MVN}_k(\mathbf{0}_k,\, \sigma_\delta^2\, \mathbf{I}_k),
\end{aligned}
\tag{3}
$$

The geographical distance $d_1(i,j)$ is the Euclidean distance between the centroids of areal polygons i and j in a two-dimensional geographical space and $d_2(i,j) = \|\mathbf{z}_i - \mathbf{z}_j\|_2$ is the Euclidean distance between the positions $\mathbf{z}_i$ and $\mathbf{z}_j$ of areal units i and j in a two-dimensional latent socio-demographic space.

The Weight Parameter

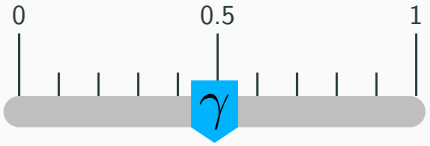
$$\text{logit}(p_{i,j}) = \alpha - \gamma\, d_1(i,j) - (1-\gamma)\, d_2(i,j)$$

SD network:



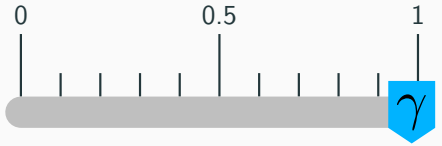
- Disregards the geographical proximity.

NNSD network:



- Puts equal weights to both distance components.

NN network:



- Disregards the socio-demographic distance.

Spatial Mixed Effect Model

$$
\begin{aligned}
&\text{Data Model}: \mathbf{Y} \mid \boldsymbol{\mu}, \sigma_y^2 \sim \text{MVN}_N(\boldsymbol{\mu}, \sigma_y^2 \mathbf{I}_N) \\
&\text{Process Model}: \boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\epsilon}, \sigma_\mu^2 \sim \text{MVN}_N(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \sigma_\mu^2 \mathbf{I}_N) \\
&\text{Process Model}: \boldsymbol{\epsilon} \mid \mathbf{B}, \sigma_\epsilon^2 \sim \text{MVN}_N(\mathbf{0}_N, \sigma_\epsilon^2 (\mathbf{D}(\mathbf{B}) - \mathbf{B})^{-1}) \\
&\text{Prior 1:} \quad \boldsymbol{\beta} \mid \sigma_\beta^2 \sim \text{MVN}_p(\mathbf{0}_p, \sigma_\beta^2 \mathbf{I}_p) \\
&\text{Prior 2:} \quad \sigma_\mu^2 \mid a_\mu, b_\mu \sim \text{Inverse-Gamma}(a_\mu, b_\mu) \\
&\text{Prior 3:} \quad \sigma_\epsilon^2 \mid a_\epsilon, b_\epsilon \sim \text{Inverse-Gamma}(a_\epsilon, b_\epsilon)
\end{aligned}
\tag{4}
$$

Analysis Results

Let Y_l be the 2019 ACS 5-year period estimates of median household income for county 'l' ($l = 1, \ldots, N$) in the state of Florida, then a pseudo data, $\mathbf{Y}^* = (Y_1^*, \ldots, Y_N^*)$ given $\mathbf{y} = (Y_1, \ldots, Y_N)$ was generated as

$$\mathbf{Y}^\star \mid \mathbf{Y} = \mathbf{y}, \boldsymbol{\Sigma}_y \sim \text{MVN}_N(\mathbf{y}, \boldsymbol{\Sigma}_y), \tag{5}$$

where the variance-covariance matrix $\boldsymbol{\Sigma}_y$ is a diagonal matrix comprising of the 2019 ACS sampling error variance-covariance matrix.

Simulation Study Results

Network Model	MSE	MAE
ICAR	7.18×10^{-4}	6.55×10^{-2}
NN	7.43×10^{-5}	1.58×10^{-2}
NNSD	4.60×10^{-5}	1.44×10^{-2}
SD	6.45×10^{-5}	1.57×10^{-2}

Table 1: Empirical simulation study.The NN network model is equivalent to the NNSD network model with $\gamma = 1$, while the SD network model is equivalent to the NNSD network model with $\gamma = 0$. ICAR refers to the traditional Intrinsic Conditional Autoregressive model. The estimates are produced based on the posterior mean of 5000 post burn-in samples from the Gibbs sampler. The mean squared error (MSE), and the mean absolute error (MAE) are used to compare the model performance with the naive model.

ACS Data Application

- We use the NNSD network to model the 2019 ACS design-based estimates of the median household income of counties in Florida
- We use the 2019 ACS design-based estimates of the median housing cost for counties in Florida on the log scale as a covariate $\mathbf{X}$.
- The latent positions $\mathbf{Z}_i$ of county i in the socio-demographic space are estimated using the ACS 2019 5-year period estimates of the percentage of county population below the poverty line as a covariate $\mathbf{S}_i$.
- We assume that the sampling error variance σ_y^2 of the response variable $\mathbf{Y}$ is known, based on the sampling error variance reported by the ACS.

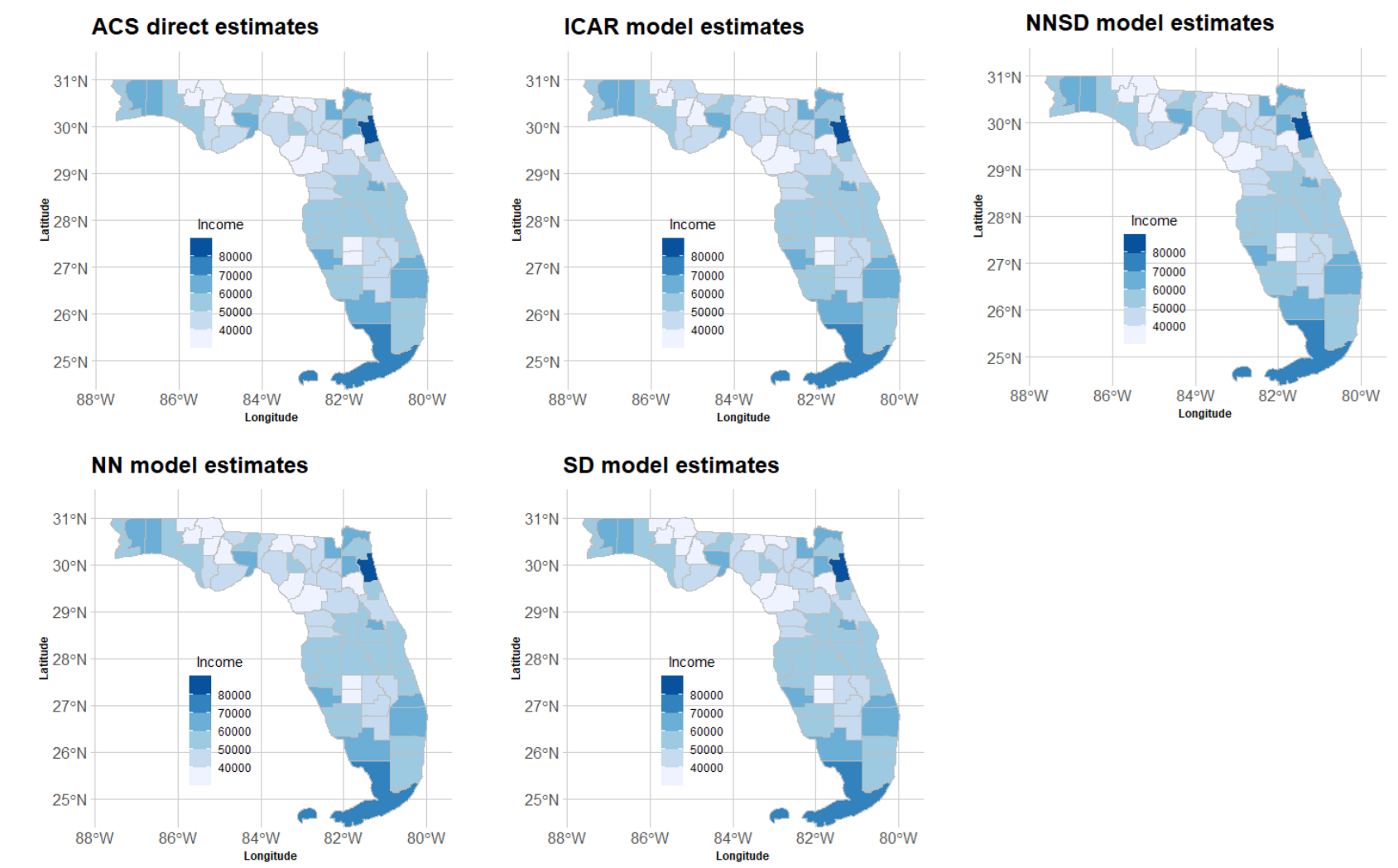
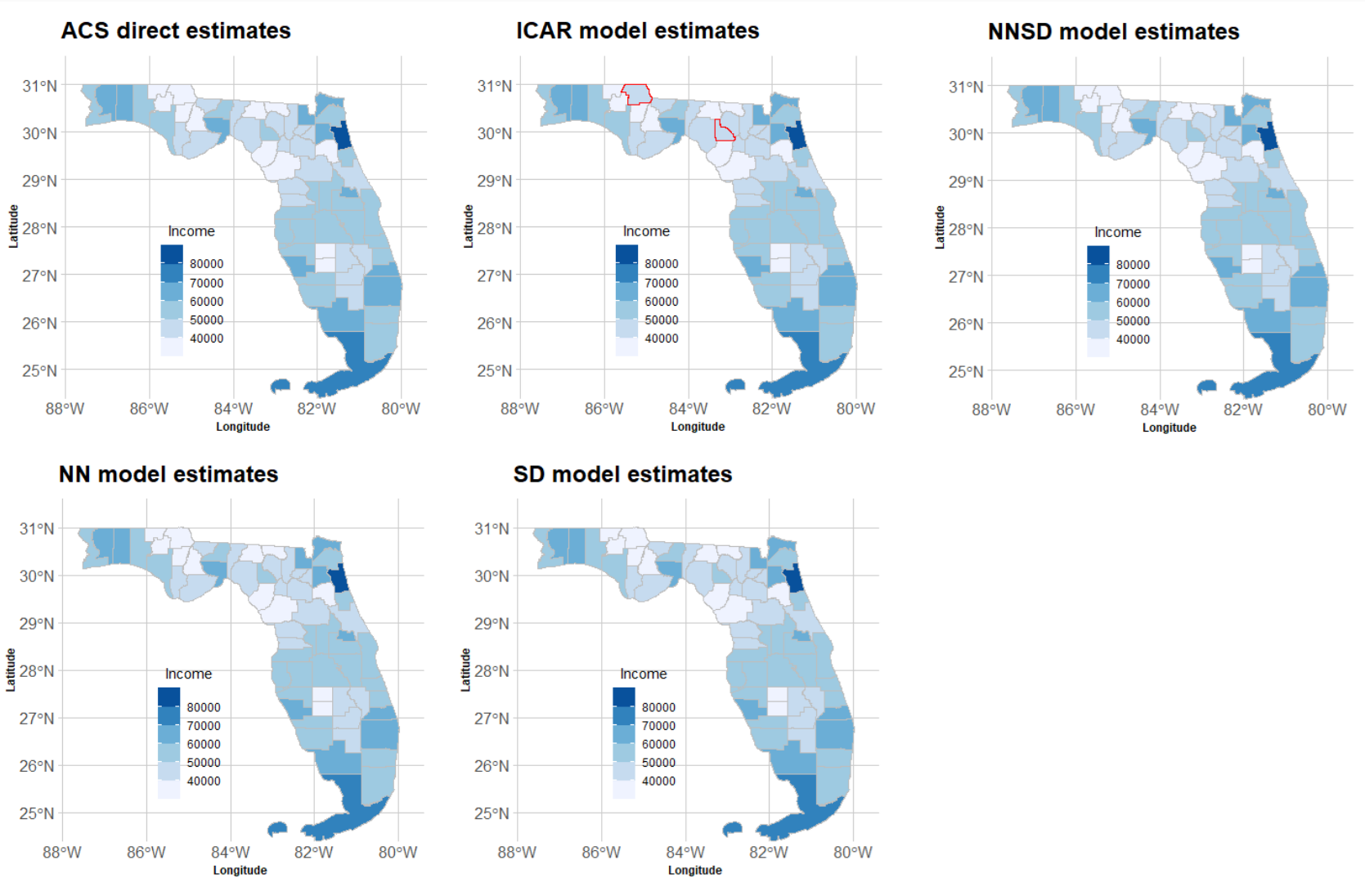


Figure 4: Maps of Florida counties showing (a) the ACS 2019 5-year period estimates of median household income of counties in Florida, on the log scale; (b) the ICAR model estimates; (c) the NNSD model estimates; (d) the NN model estimates; and (e) the SD model estimates.


ACS direct estimates
ICAR model estimates
NNSD model estimates
NN model estimates
SD model estimates
Income
80000
70000
60000
50000
40000
Latitude
Longitude
31°N
30°N
29°N
28°N
27°N
26°N
25°N
88°W
86°W
84°W
82°W
80°W

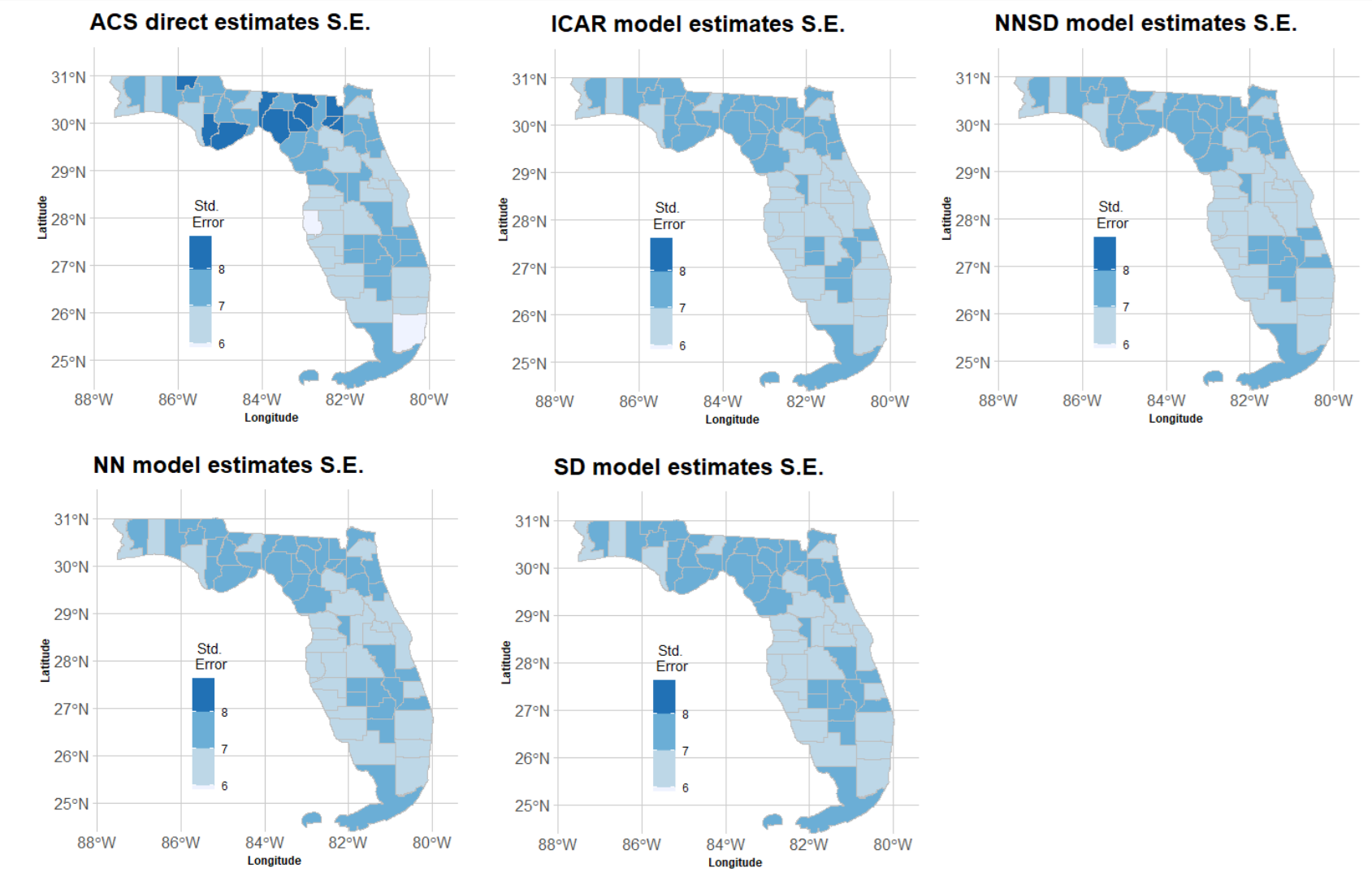


Figure 5: 2019 ACS data application: maps of Florida showing the standard errors (SE) of (a) the 2019 ACS design-based estimates, (b) the ICAR estimates, (c) the NN network model estimates, (d) the NNSD network model estimates, and (e) the SD network model estimates. The SE of the design- and model-based estimates are presented on different scales.

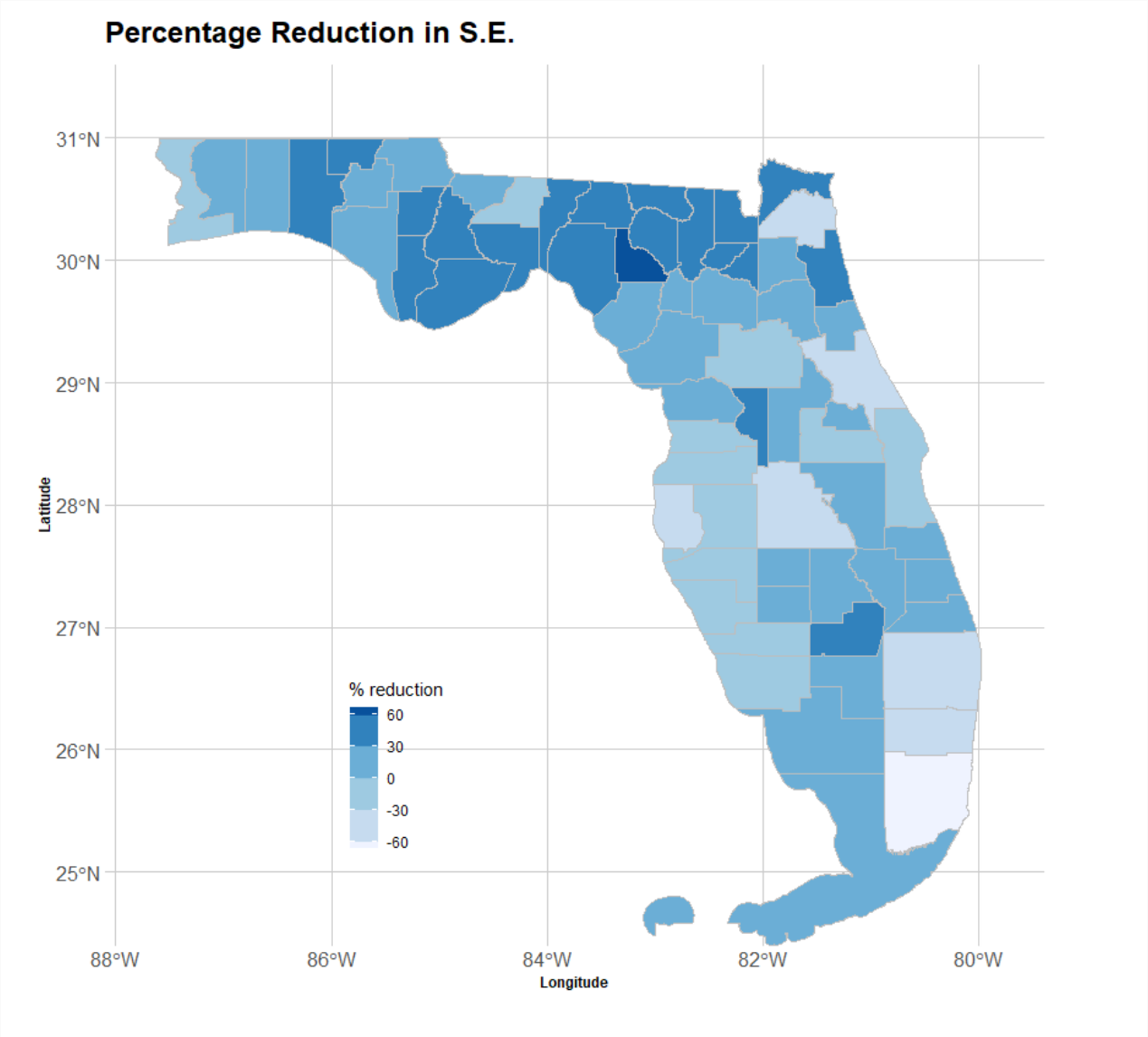


Figure 6: 2019 ACS data application: map of Florida showing the percentage reduction in standard error in the estimates using the NNSD model over the ACS direct estimates per county.

ACS Data Application

- We use the NNSD network to model the 2019 ACS design-based estimates of the median household income of counties in North and South Carolina.
- We use the 2019 ACS design-based estimates of the median housing cost for counties in North and South Carolina on the log scale as a covariate $\mathbf{X}$.
- The latent positions $\mathbf{Z}_i$ of county i in the socio-demographic space are estimated using the ACS 2019 5-year period estimates of the percentage of county population below the poverty line as a covariate $\mathbf{S}_i$.
- We assume that the sampling error variance σ_y^2 of the response variable $\mathbf{Y}$ is known, based on the sampling error variance reported by the ACS.

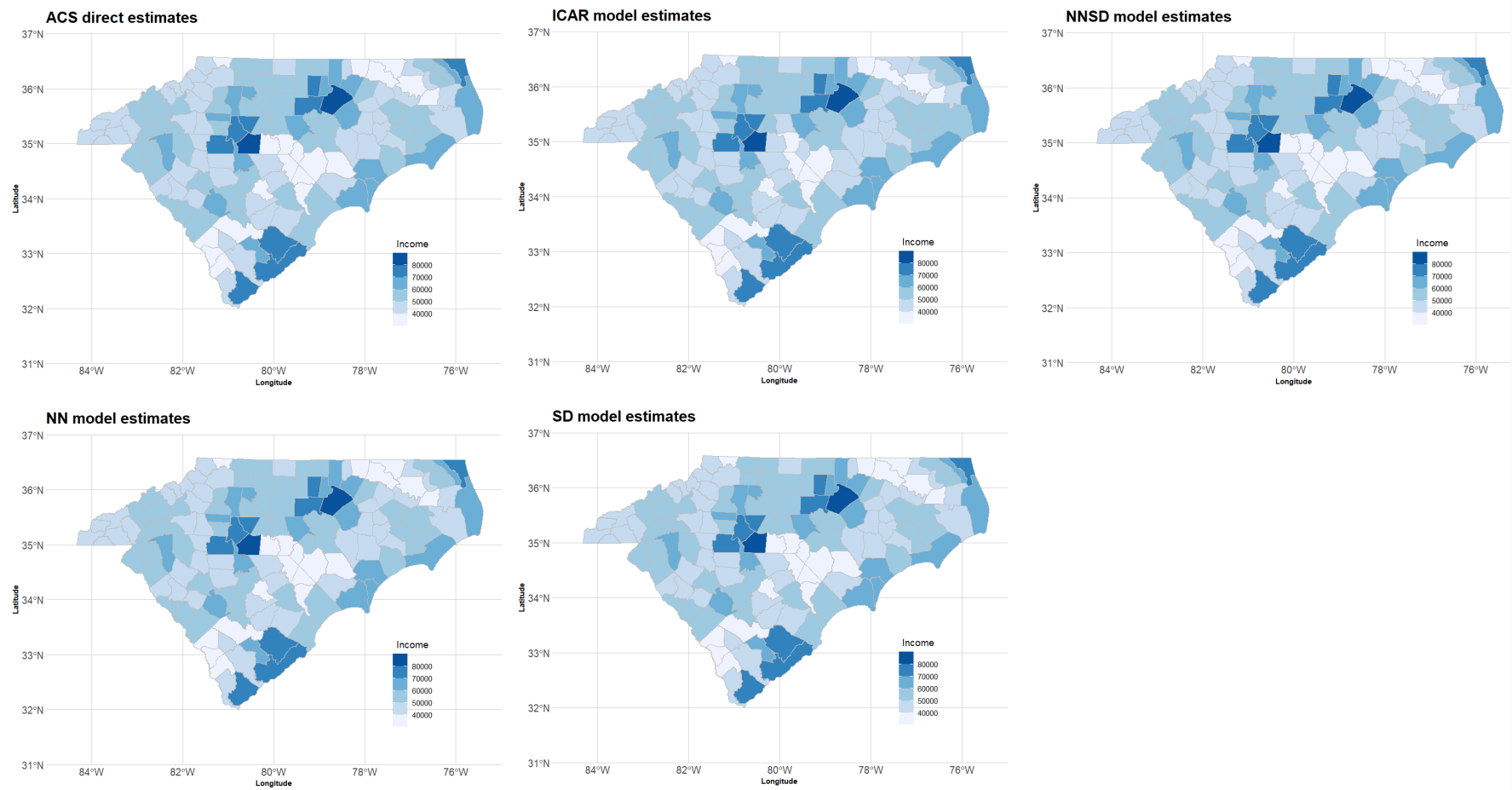
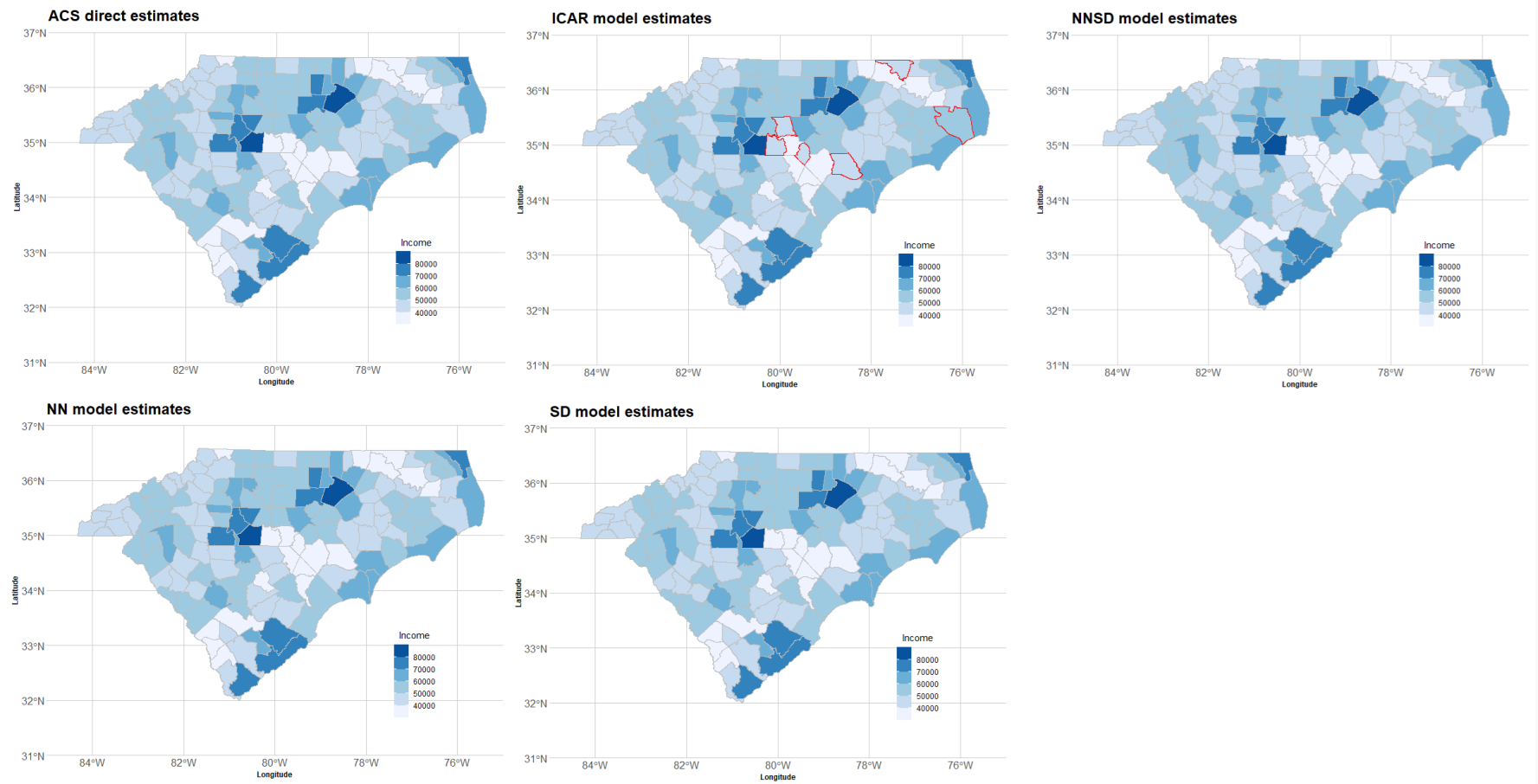


Figure 7: 2019 ACS data application: maps of North and South Carolina showing the estimates of median household income per county.


ACS direct estimates
ICAR model estimates
NNSD model estimates
NN model estimates
SD model estimates
Latitude
Longitude
37°N
36°N
35°N
34°N
33°N
32°N
31°N
84°W
82°W
80°W
78°W
76°W
Income
80000
70000
60000
50000
40000

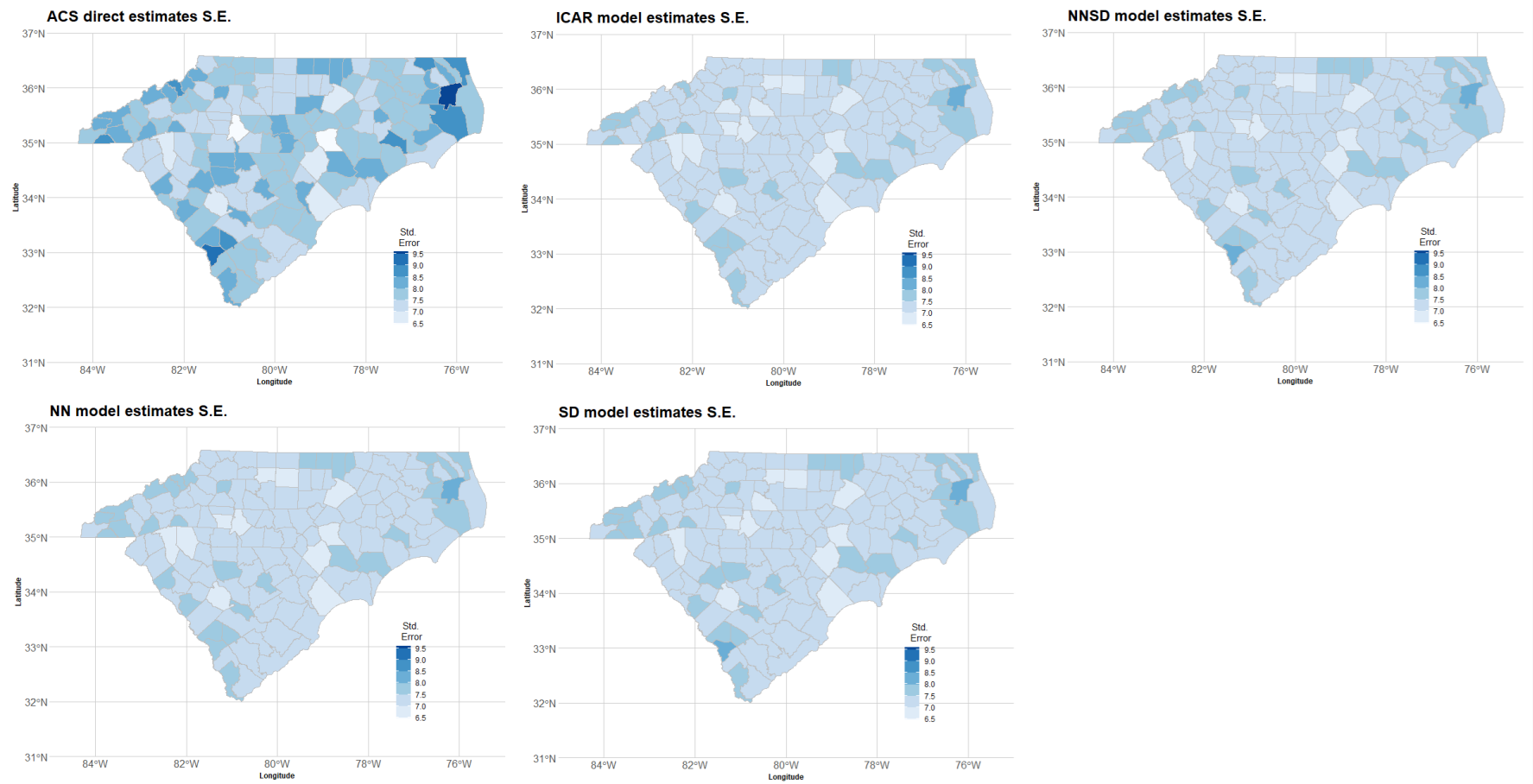


Figure 8: 2019 ACS data application: maps of North and South Carolina showing the standard error in the estimates of median household income per county.

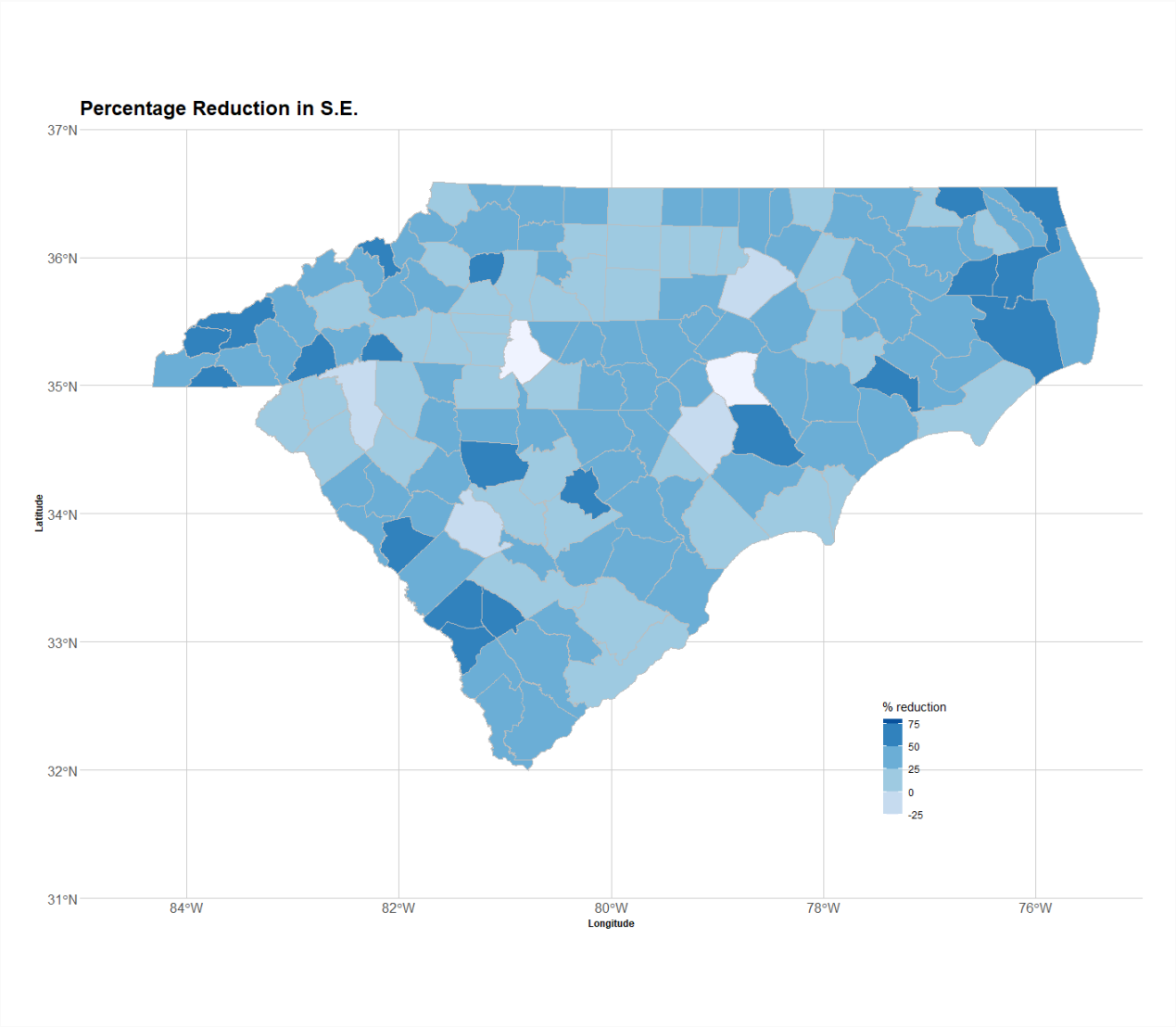


Figure 9: 2019 ACS data application: maps of North and South Carolina showing the percentage reduction in standard error in the estimates using the NNSD model over the ACS direct estimates per county.

Summary and Future Work

Summary

- Proposed a novel neighborhood network that defines the connections between units based not only on their physical proximity but also their socio-demographic similarities.
- Produced estimates with lower uncertainty that beat the industry standard ACS direct estimates.
- Provided a stochastic definition to neighborhood.

Challenges

- Dimensional issues: Run time increases significantly with number of nodes in the network.
- Sensitive to the choice of node covariants.

Extension

Varying coefficient Fay-Herriot model for dependent data based on a socio-spatial neighborhood network.

Thank you!

saikat.nandy@stjude.org

References

Besag, Julian (1974), "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 192–225.

Besag, Julian, Jeremy York, and Annie Mollié (1991), "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

Cressie, Noel and Christopher K Wikle (2011), *Statistics for spatio-temporal data*. John Wiley & Sons.

Tobler, W. R. (1970), "A computer movie simulating urban growth in the Detroit region." *Economic Geography*, 46, 234–240.