Bayesian Hierarchical Models For Spatially Correlated Multitype Survey Data Using Covariates Measured With Error

Saikat Nandy [1], **Scott H. Holan** [2,3], **Jonathan R. Bradley** [4], and **Christopher K. Wikle** [1]

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

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.

Outline

Introduction

Area-level models:

- Treat direct estimate as the response variable and typically incorporates smoothing through the model.
- Uses auxiliary data (covariates) that are considered fixed and known.

Relevant literature:

- Fay and Herriot (1979) Small area model uses auxiliary information but auxiliary information is measured with error.
- Ybarra and Lohr (2008) Area-level model used auxiliary estimators with sampling (measurement) error.
- Arima et al. (2015) Hierarchical Bayesian measurement error area-level model.
- Li et al. (2009) LMMs for spatial data in the presence of covariate measurement errors.
- Huques et al. (2014) Measurement error model with homoscedastic error variance in geostatistical data.
- Tadayon and Torabi (2019) Spatial models with covariate measurement error for non-Gaussian geostatistical data.

Motivation – Multi-type Data and Spatially Correlated Measurement Error

Our Goal

Extend the measurement error modeling paradigm to area level non-Gaussian data, that can be distributed from one or multiple classes of distributions (e.g., Gaussian, Poisson, Binomial etc.) and when the underlying true covariates are spatially correlated.

Our Solution

A multivariate *multi-type* Bayesian hierarchical mixed effect model for area-level data which can be distributed from multiple classes of distributions both Gaussian and non-Gaussian, using auxiliary data which are measured with error and are spatially dependent, under Hierarchical Generalized Transformation model specification (Bradley, 2022).

Methodology

Non-Gaussian Data Modeling Challenges

- The latent Gaussian process (LGP) model is a standard tool for modeling dependent non-Gaussian data.
- Their use typically leads to full-conditional distributions that are not conjugate.
- Often requires difficult-to-tune Metropolis-Hastings algorithms.
- Modern high-dimensional datasets poses a challenge to these modeling strategies.

Classical Approach

Assume a *multiple response-type data* Z_{ij} (e.g., Gaussian, Poisson or Binomial) for areal unit i ($j = 1, 2, 3$ indexes the response type). Then, if we assume conditional independence of Z_{ij} given an unobserved latent process Y_{ij}, we have the following hierarchical structure

$$
\begin{aligned}
Z_{i1}|Y_{i1} &\overset{ind.}{\sim} Normal(Y_{i1}, v), \\
Z_{i2}|Y_{i2} &\overset{ind.}{\sim} Poisson(\exp(Y_{i2})), \\
Z_{i3}|Y_{i3} &\overset{ind.}{\sim} Binomial\left\{b_i, \frac{\exp(Y_{i3})}{1+\exp(Y_{i3})}\right\}; \qquad i = 1, \ldots, N.
\end{aligned}
$$

A traditional approach to model such data would be to impose a transformation such that

$$
h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta} \overset{ind.}{\sim} f(h_j(Z_{ij})|Y_{ij}, \boldsymbol{\theta}), \qquad i = 1, \ldots, N, j = 1, 2, 3. \tag{1}
$$

where $f(\cdot)$ is some *"preferred model."*

Hierarchical Generalized Transformation (HGT) Model

Bradley (2022) introduced a Bayesian implementation of the Hierarchical Generalized Transformation (HGT) model, using the following three components:

- The distribution of the data given a transformation, $f(Z_{ij}|h_{ij})$ is referred to as the *"data model."*

- The prior distribution of the transformation, $f(h_{ij}|\gamma)$ is referred to as *"transformation prior,"* where γ is a vector valued hyperparameter and $f(\gamma)$ is referred to as the *"transformation hyperprior."*

- The density $f(\mathbf{h}|\mathbf{y},\theta) = \prod_i \prod_j f(h_j(Z_{ij})|Y_{ij},\theta)$ denotes the *"transformed data model"* and $f(\mathbf{y}|\theta)$ denotes the *"process model,"* where $\mathbf{h} = (h_1(Z_{11}),\ldots,h_3(Z_{N3}))'$ and $\mathbf{y}$ are the $N^*(=3N)-$dim transformed data and latent process respectively.

Hierarchical Generalized Transformation Model Hierarchy

Assume multiple response-type survey data Z_{ij} (e.g., Gaussian, Poisson or Binomial) for areal unit i ($j = 1, 2, 3$ indexes the response type).

The HGT model for the multi-type responses takes the form,

$$
\begin{aligned}
&\textit{Data Model } 1: \; Z_{i1}|h_{i1} \overset{ind}{\sim} \text{Normal}(h_{i1}, v) \\
&\textit{Data Model } 2: \; Z_{i2}|h_{i2} \overset{ind}{\sim} \text{Poisson}(\exp(h_{i2})) \\
&\textit{Data Model } 3: \; Z_{i3}|h_{i3} \overset{ind}{\sim} \text{Binomial}\left\{ b_i, \frac{\exp(h_{i3})}{1+\exp(h_{i3})} \right\} \qquad (2) \\
&\textit{Transformed Data Model}: \; \mathbf{h}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\gamma} \propto f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) m(\mathbf{h}|\boldsymbol{\gamma}), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, N.
\end{aligned}
$$

The $m(\mathbf{h}|\boldsymbol{\gamma})$ in the "transformed data model" is a proportionality constant defined as

$$
m(\mathbf{h}|\boldsymbol{\gamma}) = f(\mathbf{h}|\boldsymbol{\gamma}) / \int \int f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\mathbf{y} d\boldsymbol{\theta}.
$$

HGT - Spatial mixed effect Measurement Error Model (HGT-SME)

- Utilizing the HGT model, we have successfully transformed the Gaussian or non-Gaussian response(s) to continuous response data $\mathbf{h}$.

- We now define the following mixed effect model

$$\mathbf{h}|\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\xi}, \mathbf{W}, \mathbf{M}, \tau^2 \propto N(\mathbf{W}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\delta} + \mathbf{M}\boldsymbol{\eta} + \boldsymbol{\xi}, \tau^2\mathbf{I}) m(\mathbf{h}|\boldsymbol{\gamma}). \quad (3)$$

- Suppose, $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_{N^*})'$ is a $N^* \times p$ design matrix consisting of the $p \times 1$ vector of "true" unobserved covariates $\mathbf{w}_i$ that are measured with error.

- We do not observe $\mathbf{W}$, but rather an estimate $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_{N^*})'$.

The Measurement Error Component and Spatial Correlation

- We assume this estimate, $\mathbf{X}$, is prone to additive measurement error.

- Thus, we adopt a *classical measurement error* model as

$$\mathbf{x}_i = \mathbf{w}_i + \mathbf{u}_i, \quad i = 1, \ldots, N^*, \tag{4}$$

  where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{N^*})'$ is the measurement error component and we assume $\mathbf{u}$ to have mean $\mathbf{0}$ and variance covariance matrix $\mathbf{\Sigma_u}$.

- The true covariates, are also spatially correlated, and so we have assume a Gaussian prior on $\mathbf{W}$ with mean $\boldsymbol{\mu_W}$ and covariance matrix $\mathbf{\Sigma_W}$.

- For $p = 1$, we assume a univariate Conditional Autoregressive (CAR) prior on $\mathbf{W}$ and define $\mathbf{\Sigma_W} = \sigma^2_{\mathbf{W}}(\mathbf{D} - \rho\mathbf{A})^{-1}$ where $\mathbf{A}$ is the adjacency matrix and $\mathbf{D}$ a diagonal matrix with entries equal to the number of neighbors for location l.

The Other Components

$$\mathbf{h}|\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\xi}, \mathbf{W}, \mathbf{M}, \tau^2 \propto N(\mathbf{W}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\delta} + \mathbf{M}\boldsymbol{\eta} + \boldsymbol{\xi}, \tau^2\mathbf{I}) m(\mathbf{h}|\boldsymbol{\gamma})$$

- $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_{N^*})'$: $N^* \times q$ matrix of fixed covariates $\mathbf{s}_i$ that are measured without error.
- $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\delta} \in \mathbb{R}^q$: Unknown coefficient parameters to be estimated.
- $\boldsymbol{\eta}$: r-dimensional random coefficient vector,
- M : The real-valued spatial basis function matrix of rank r, with $r \ll N$. We use Moran's I basis functions.
- $\boldsymbol{\xi}$: Independent fine-scale variation component.

HGT Component

$$\textit{Data Model 1}: \; Z_{i1}|h_{i1} \overset{ind}{\sim} \text{Normal}(h_{i1}, v)$$

$$\textit{Data Model 2}: \; Z_{i2}|h_{i2} \overset{ind}{\sim} \text{Poisson}(\exp\{h_{i2}\})$$

$$\textit{Data Model 3}: \; Z_{i3}|h_{i3} \overset{ind}{\sim} \text{Binomial}\left\{b_i, \frac{\exp\{h_{i3}\}}{1+\exp\{h_{i3}\}}\right\}$$

$$\textit{Transformed Data Model}: \; \mathbf{h}|\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\gamma} \propto f(\mathbf{h}|\mathbf{y}, \boldsymbol{\theta}) m(\mathbf{h}, \boldsymbol{\gamma}),$$

$$i = 1, \ldots, N.$$

SME Component

$$\mathbf{h}|\mathbf{W}, \beta, \delta, \eta, \gamma, \xi, \mathbf{M}, \tau^2 \propto N(\mathbf{W}\beta + \mathbf{S}\delta + \mathbf{M}\eta + \xi, \tau^2\mathbf{I}) m(\mathbf{h}|\gamma)$$

$$\mathbf{X}|\mathbf{W}, \mathbf{\Sigma_U} \sim \textit{Gaussian}(\mathbf{W}, \Sigma_U)$$

$$\mathbf{W}|\boldsymbol{\mu}_\mathbf{W}, \sigma^2_\mathbf{W}, \mathbf{\Sigma_W} \sim \textit{Gaussian}(\boldsymbol{\mu}_\mathbf{W}, \sigma^2_\mathbf{W}\mathbf{\Sigma}_W)$$

$$\eta|\boldsymbol{\sigma}^2_\eta \overset{ind.}{\sim} \textit{Gaussian}(\mathbf{0}_r, \sigma^2_\eta \mathbf{I}_r)$$

$$\xi|\boldsymbol{\sigma}^2_\xi \overset{ind.}{\sim} \textit{Gaussian}(\mathbf{0}_N, \sigma^2_\xi \mathbf{I}_N)$$

Naive Model

The *naive* model has the following hierarchical structure

$$
\begin{aligned}
\mathbf{h}|\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\xi}, \mathbf{M}, \tau^2 &\propto N(\mathbf{L}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\eta} + \boldsymbol{\xi}, \tau^2\mathbf{I}) m(\mathbf{h}|\boldsymbol{\gamma}), \\
\boldsymbol{\eta}|\boldsymbol{\sigma}^2_{\boldsymbol{\eta}} &\overset{ind.}{\sim} \mathit{Gaussian}(\mathbf{0}_r, \sigma^2_\eta \mathbf{I}_r) \\
\boldsymbol{\xi}|\boldsymbol{\sigma}^2_{\boldsymbol{\xi}} &\overset{ind.}{\sim} \mathit{Gaussian}(\mathbf{0}_N, \sigma^2_\xi \mathbf{I}_N) \\
\boldsymbol{\beta} &\sim \mathit{Gaussian}(\mathbf{0}_p, \sigma^2_\beta \mathbf{I}_p) \\
\boldsymbol{\tau}^2 &\sim \mathit{IG}(\alpha_\tau, \beta_\tau) \\
\boldsymbol{\sigma}^2_{\boldsymbol{\xi}} &\sim \mathit{IG}(\alpha_\xi, \beta_\xi) \\
\boldsymbol{\sigma}^2_{\boldsymbol{\eta}} &\sim \mathit{IG}(\alpha_\eta, \beta_\eta).
\end{aligned}
$$

where $\mathbf{L} = [\mathbf{X}, \mathbf{S}]$ is the $N^* \times (p + q)$ design matrix consisting of covariates which are assumed to have no measurement error, with $\mathbf{X}$ and $\mathbf{S}$ as defined earlier.

Analysis Results

Data for Simulation Study

- Multiple response types: Gaussian and Poisson responses
- Study Area: 175 counties in the states in Pacific NW, USA- Washington, Oregon, Idaho, and Montana.
- We use the HGT-SME method to model pseudo data calibrated towards the ACS 2019 5-year period estimates of total count of county population below the poverty level as the Poisson count response and log of ACS 2019 5-year period estimates of median housing cost per county as the Gaussian response.
- As the error prone covariate, we use the log of ACS 2019 5-year period estimates of median household income in all the counties in the study area.
- As the covariate that is not measured with error, we use the ACS 2019 5-year period estimates of percentage of county households that receive food stamps/SNAP.

Empirical Simulation Study

Let Z_{I1} be the 2019 ACS 5-year period estimates of median housing cost for county 'I' ($I = 1, \ldots, N$) in the study area, then continuous pseudo data was generated as

$$Z^*_{I1} \sim Normal(\log(Z_{I1}), 1), \qquad I = 1, \ldots, N.$$

Subsequently, let Z_{I2} be the ACS 2019 5-year period estimate of the count of population below poverty threshold for county 'I' ($I = 1, \ldots, N$) in the study area, then count-valued pseudo data was generated as

$$Z^*_{I2} \sim Poisson(Z_{I2} + 1), \qquad I = 1, \ldots, N.$$

50 independent replicates of this data are produced.

Simulation Results

	Gaussian Response			Poisson Response		
Model	RMSE	MSE	MAE	RMSE	MSE	MAE
Naive	0.4357	0.1898	0.3465	4997.59	2.50×10^7	1636.978
HGT-SME	0.3262 ($\downarrow$ 25%)	0.1064 ($\downarrow$ 44%)	0.2583 ($\downarrow$ 25%)	3895.155 ($\downarrow$ 22%)	1.52×10^7 ($\downarrow$ 39%)	1115.470 ($\downarrow$ 32%)

The estimates are produced based on the posterior mean of 5000 post burn-in samples from the Gibbs sampler. The root mean squared error (RMSE), and the mean absolute error (MAE) are used to compare the model performance with the naive model.

American Community Survey Data Application

- Jointly model the ACS 2019 5-year period estimates of counts of the population under 18 years of age that are below the poverty threshold and the log ACS 2019 5-year period estimates of median housing cost per county in the states of Washington, Oregon, Idaho, and Montana, USA.

- Error-prone covariates: (i). Log ACS 2019 5-year period estimates of median household income per county and, (ii). the ACS 2019 5-year period estimates of the proportion of the county population that receives SNAP.

- Error free fixed covariates: County 'child tax filer rate', defined as the number of child exemptions in the county claimed on tax returns divided by the county population age $0 - 17$ years. This is derived from public use data released by the IRS under its Statement of Income (SOI) program.

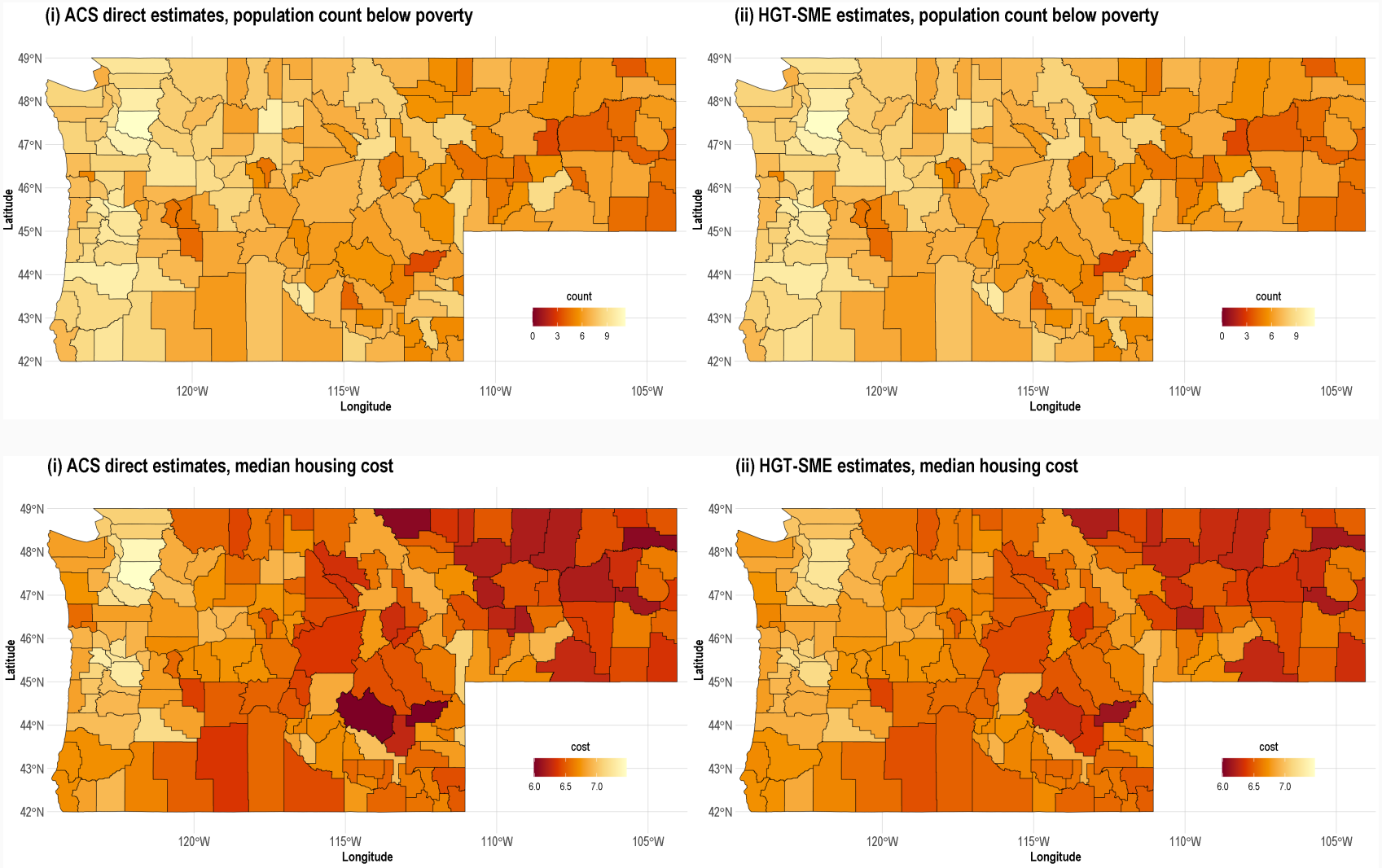

Figure 1: Maps showing (i) the ACS 2019 5-year period estimates of the count of county population below the poverty threshold in the 4 states and the log ACS 2019 5-year period estimates of median housing cost in the 4 states against (ii) the HGT-SME model estimates of the counts, both on the log scale and the HGT-SME model estimates of the log median housing costs.

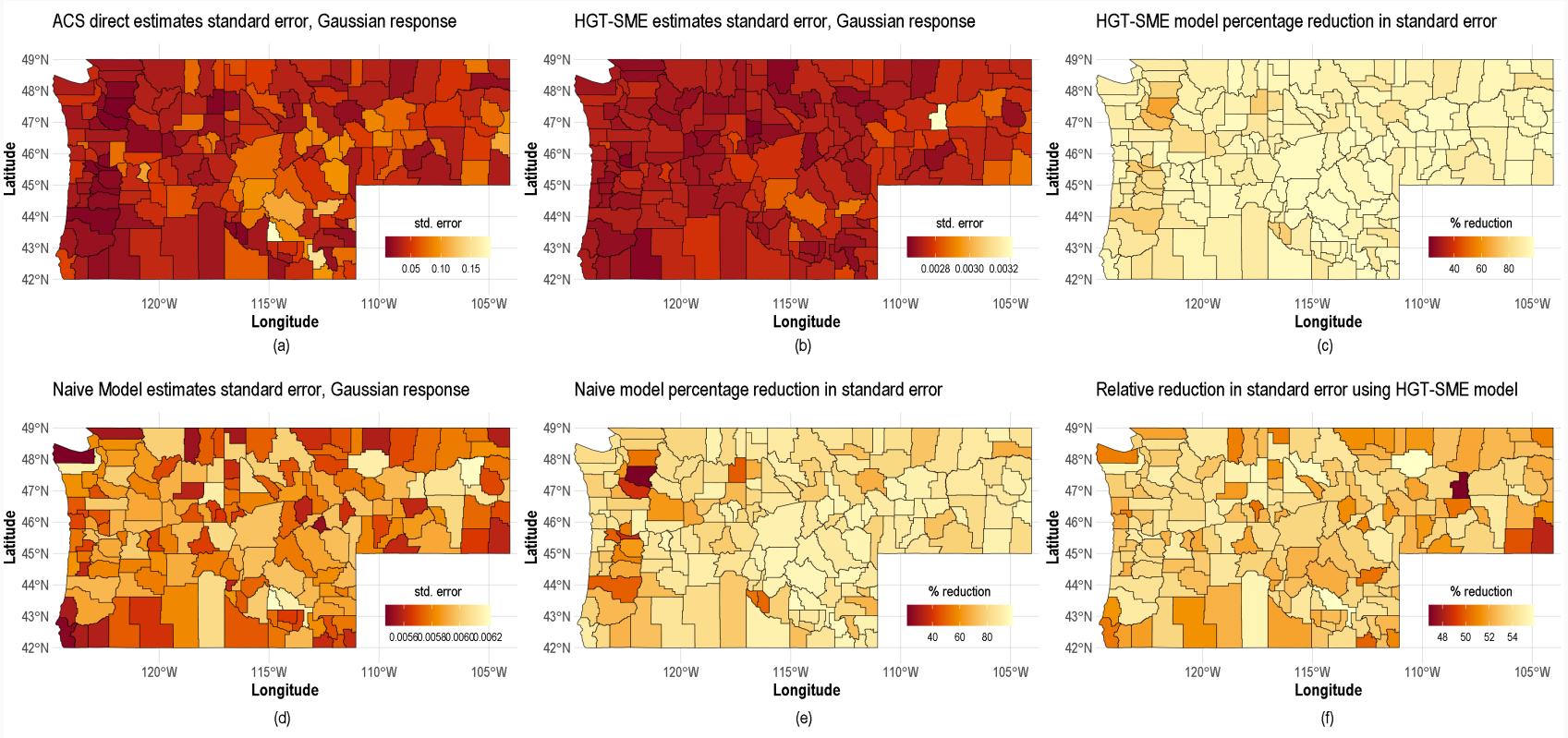

Figure 2: (a) Map of the study area showing the standard error in the ACS 2019 direct estimates of log median housing cost. (b-c) Map of the study area showing the standard error in the HGT-SME model estimates and the percentage reduction in standard error achieved by the HGT-SME model over the ACS direct estimates. (d-e) Map of the study area showing the standard error in the naive model estimates and the percentage reduction in standard error achieved by the naive model over the ACS direct estimates. (f) Map of the study area showing the relative percentage reduction in standard error achieved by the HGT-SME model estimates over the naive model. Maps (a), (b), and (d) do not share the same scale in legend due to the significant difference in the range of the corresponding values.

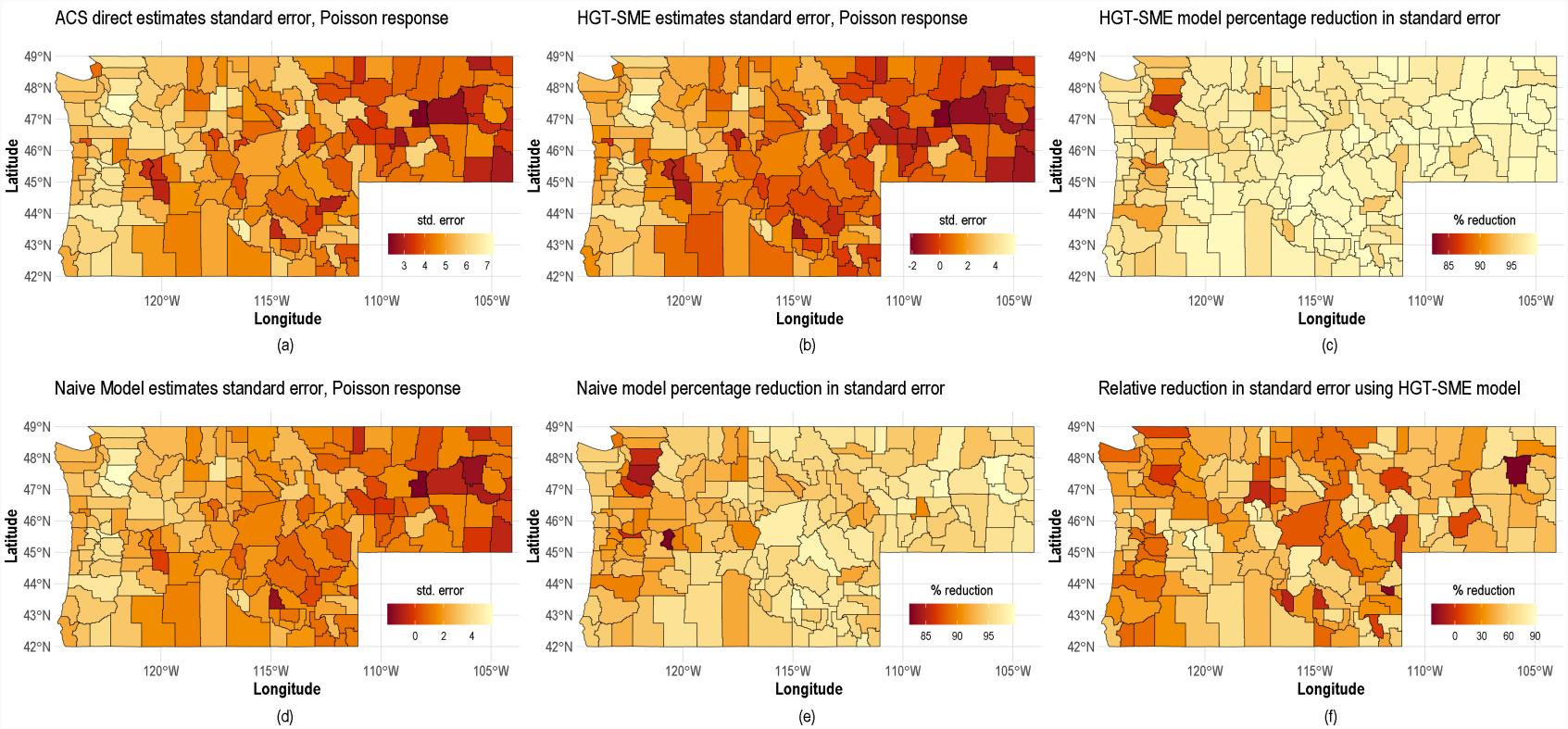

Figure 3: (a) Map of the study area showing the standard error in the ACS 2019 direct estimates of log median housing cost. (b-c) Map of the study area showing the standard error in the HGT-SME model estimates and the percentage reduction in standard error achieved by the HGT-SME model over the ACS direct estimates. (d-e) Map of the study area showing the standard error in the naive model estimates and the percentage reduction in standard error achieved by the naive model over the ACS direct estimates. (f) Map of the study area showing the relative percentage reduction in standard error achieved by the HGT-SME model estimates over the naive model. Maps (a), (b), and (d) do not share the same scale in legend due to the significant difference in the range of the corresponding values.

Conclusion

Conclusion

- We have proposed a fully Bayesian approach to model spatially distributed survey data, which can be Gaussian or non-Gaussian.
- We have used the HGT component to deal with the computational complexity associated with modeling non-Gaussian data.
- The HGT component of our model also enables us to model multi-type response datasets under a unified framework without significant changes to the model hierarchy.
- We have illustrated our model performance by designing empirical simulation studies based on data calibrated towards survey data produced by the ACS.
- Using the ACS, we have exhibited a multi-type response application, where we jointly estimated median housing costs and count of population below poverty, thus joint modeling of a Gaussian and Poisson response under a unified latent process framework.
- In both cases we illustrated superior performance.

Ongoing Work

- It is important to note that the sampling variances associated with the response variables have not been included in our proposed model and instead, this error has been treated as unknown.
- In the proposed model, the HGT model hierarchy transforms the multi-type data into a single response dataset with continuous support.
- However, unlike a delta method in a traditional FH model, where the transformations are derived based on known functions, the HGT framework transforms the data assuming a prior distribution on the transformation to allow for uncertainty.
- One might consider the direct survey variances as a response, and assume a mean-variance relationship between the direct survey estimates and the survey variances or appeal to definitions of effective sample size.

Thank you!

holans@missouri.edu

scott.holan@census.gov

References

Arima, Serena, Gauri S. Datta, and Brunero Liseo (2015), "Bayesian estimators for small area models when auxiliary information is measured with error." *Scandinavian Journal of Statistics*, 42, 518–529.

Bradley, Jonathan R (2022), "Joint Bayesian analysis of multiple response-types using the hierarchical generalized transformation model." *Bayesian Analysis*, 17, 127–164.

Fay, Robert E. and Roger A. Herriot (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.

Huques, Md Hamidul, Howard Bondell, and Louise Ryan (2014), "On the impact of covariate measurement error on spatial regression modelling." *Environmetrics*, 25, 560–570.

Li, Yi, Haicheng Tang, and Xihong Lin (2009), "Spatial linear mixed models with covariate measurement errors." *Statistica Sinica*, 19, 1077–1093.

Tadayon, Vahid and Mahmoud Torabi (2019), "Spatial models for non-Gaussian data with covariate measurement error." *Environmetrics*,