Spring 2023

Detecting Spatially Varying Coefficient Effects With Conditional Autoregressive Models: A Simulation Study Using Social Determinants of Health Screening Data

Reid J. DeMass

DETECTING SPATIALLY VARYING COEFFICIENT EFFECTS WITH CONDITIONAL AUTOREGRESSIVE MODELS: A SIMULATION STUDY USING SOCIAL DETERMINANTS OF HEALTH SCREENING DATA

by

Reid J. DeMass

Bachelor of Science
Clemson University, 2020

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2023

Accepted by:

Stella Self, Director of Thesis

Caroline Rudisill, Reader

Jiajia Zhang, Reader

Cheryl L. Addy, Interim Vice Provost and Dean of the Graduate School

Acknowledgements

I would like to thank Dr. Stella Self for her encouragement and guidance as an instructor, research mentor, and thesis advisor. I would also like to thank Dr. Caroline Rudisill for her support and training of me as a graduate research assistant and Dr. Jiajia Zhang for her insight both in the classroom and on my thesis. I would like to extend my gratitude to my parents for their love and support and to all my friends and family for bringing me continued joy.

Abstract

Generalized linear models which include spatially varying coefficient terms allow researchers to determine if the association between predictor and outcome variables vary across geographic space. Such models are particularly applicable to research with public health data where interventions and limited health care resources must be allocated carefully. The integrated nested Laplace approximation (INLA) methodology available in the R INLA package is a popular tool to estimate spatially varying coefficients. To assess the performance of the estimation procedure, patient emergency department (ED) visits were simulated from data sourced from a pilot study at Prisma Health. The INLA technique was used to assess whether the association between the patient response to a social determinant of health (SDoH) screening question and the number of ED visits varied across census block groups. The power of the estimation procedure for increasing numbers of positive screening rates of the SDoH question and for varying values of the variance parameter governing the distribution of the spatially varying coefficients was of interest. Furthermore, the type I error rate of the INLA estimation was also investigated. It was found that the power in detecting spatial variation increases as both the number of positive screens and variance parameter increases. The type I error rate remained below 0.1% for all simulations. The INLA estimation procedure was subsequently applied to the Prisma Health pilot study data, and no spatial variation for the association between screening positive for violence/abuse SDoH and ED visits was found.

Table of Contents

List of Tables

Chapter 1: Introduction

Modelling heath care data with spatial techniques allows researchers and practitioners to determine how populations across geographical areas may differ. Populations at different locations may possess different health needs, increased or decreased susceptibility to diseases, varied ability to access health resources, or differ in other important demographic features (McLafferty, 2003). Furthermore, there are often geographical differences in environmental factors and agents (e.g., pollution) as well as spatial differences in the quality, cost, and availability of health care systems (Bauer et al., 2020; Jerrett, Gale, & Kontgis, 2010). Statistics such as the Getis-Ord Gi^* and local Moran's I statistics can identify geographic locations that have significantly high or low values for a given variable of interest. However, these measures are relatively descriptive in nature. A variety of regression approaches for spatial modeling have also been developed. For example, in spatial random effects models, each geographic location has its own intercept called a spatial random effect (Krainski et. al, 2018). In spatially varying coefficient models, the effects of one or more covariates are allowed to change with geographic location. Multiple models exist specifying the structure of the random effect or spatially varying coefficients and each has its own unique assumptions. In this paper, we consider the intrinsic conditional autoregressive (CAR) model for Bayesian spatially varying coefficient models. Once the spatially varying coefficients are assumed to follow a CAR prior distribution, they can be integrated into generalized linear models and Bayesian statistical methods can be used to estimate model parameters.

1.1 Conditional Autoregressive (CAR) models

Consider some geographic region, G, which can be divided into n sub-regions, $s_i, i = 1, \dots, n$. There are many ways that G can be divided; some examples would be overlaying a simple lattice structure of evenly sized squares over G or using a predefined partition for G, such as states, counties, or census block groups. Furthermore, consider the neighborhood structure of the sub-regions within a given G. Define two sub-regions s_i and s_j as neighbors if, for $i = 1, \dots n; j = 1, \dots, n;\ i \neq j$, the s_i and s_j share a common boundary. Then, a binary adjacency matrix, $\boldsymbol{W}$, can be constructed where $w_{ij} = 1$ if i and j are neighbors, and 0 otherwise. It is apparent that $\boldsymbol{W}$ is symmetric since s_j and s_i must be neighbors if s_i and s_j are neighbors. Furthermore, the sum over a single row i equals the total number of neighbors for a given sub-region, s_i. Let all the row sums be denoted by $d_i, i = 1, \dots n$; then, we can define the diagonal matrix $\boldsymbol{D}$ where $\boldsymbol{D}_{ii} = d_i$.

Now, let us assume that each sub-region, $s_1, \dots, s_n$, possesses its own spatially varying coefficient, $\varphi_{s_1}, \dots, \varphi_{s_n}$. Under the CAR model, we assume that the joint distribution of the vector of spatially varying coefficients, $\boldsymbol{\varphi}$, follows a multivariate normal distribution

$$\boldsymbol{\varphi} = \left(\varphi_{s_1}, \dots, \varphi_{s_n}\right)^T \sim MVN(\mathbf{0}, \tau^2(\boldsymbol{D} - \rho\boldsymbol{W})^{-1})$$

as shown in Banerjee, Carlin, and Gelfand (2014) and originally proposed by Besag (1974). Here, $\boldsymbol{D}$ and $\boldsymbol{W}$ are the diagonal and binary adjacency matrices, respectively, defined above. The parameter ρ can be thought of as a spatial correlation parameter, typically defined on the range [0 1], and with values very near one suggesting a strong association between sub-regions that are neighbors (Banerjee et al., 2014). The parameter which controls the variance, τ^2, is typically unknown and assumed to possess some value

from a prior distribution (Gómez-Rubio, 2020). It can then be shown that the conditional distribution for a sub-region φ_{s_i} is

$$\varphi_{s_i}|\boldsymbol{\varphi}_{-\boldsymbol{s_i}} \sim N(\frac{\rho \sum_{j=1}^{n} w_{ij}\, \varphi_j}{\boldsymbol{D}_{ii}}, \frac{\tau^2}{\boldsymbol{D}_{ii}})$$

where $\boldsymbol{\varphi}_{-\boldsymbol{s_i}}$ indicates all spatial coefficients except φ_{s_i}; $\sum_{j=1}^{n} w_{ij}\, \varphi_j$ is the sum of all spatial coefficients for neighbors of s_i; $\boldsymbol{D}_{ii}$ the number of neighbors of s_i; ρ is the spatial correlation parameter; and τ^2 controls the magnitude of the variance (Banerjee et al., 2014).

Assuming a CAR model on $\boldsymbol{\varphi}$ is common practice as it allows the conditional distribution of φ_{s_i} to have desired and interpretable properties. The mean of the conditional distribution is simply the average value of all neighbors of a sub-region scaled by ρ, which is sensible if we think sub-regions near each other should be similar (Banerjee et al., 2014). Furthermore, it is apparent that the more neighbors a sub-region has, the smaller the estimate of the variance will be. Again, this appears reasonable as more neighbors would provide more information on the estimate of the mean, and we might then have less uncertainty (i.e., less variance) about the sub-region. Lastly, the precision matrix under CAR models is sparse making them computationally efficient, especially when a large number of regions is under consideration (Rue, Martino, & Chopin, 2009).

Certain values of ρ give rise to distinct interpretations of the CAR model. If ρ is zero, then it would follow that we are assuming no spatial correlation between sub-regions and the φ_{s_i}'s are independent each with mean zero. If ρ is 1, then we do not scale

the mean and it is exactly the average of its neighbors. This model is called the intrinsic CAR model and its conditional distribution can be shown to be

$$\varphi_{s_i}|\varphi_{-s_i} \sim N(\frac{\sum_{j=1}^{n} w_{ij}\,\varphi_j}{\boldsymbol{D}_{ii}}, \frac{\tau^2}{\boldsymbol{D}_{ii}})$$

(Banerjee et al., 2014). The intrinsic CAR model may be preferred for its interpretation of the mean being directly the average of its neighbors and it may be able to handle irregular spatial behavior better than ρ for values not equal to one (Banerjee et al., 2014; Gelfand & Vounatsou, 2003). However, the intrinsic CAR model is called an improper model because the variance matrix of the joint distribution is not invertible and, thus, the functional form of the joint distribution of the intrinsic CAR model cannot be determined (Gelfand & Vounatsou, 2003). This is an undesirable property, but it not a huge hindrance in practical applications where it is common practice to add the constraint that $\sum_{i=1}^{n} \varphi_{s_i} = 0$, allowing all conditional distributions to be proper (Banerjee et al., 2014). Even without such a constraint, an improper CAR prior on random effects or spatially varying coefficients in a regression setting generally results in a proper posterior distribution.

For $0 < \rho < 1$, the joint distribution of $\boldsymbol{\varphi}$ can be shown to be proper (i.e., for such values of ρ, the covariance matrix is invertible) and such models are referred to as proper CAR models (Banerjee et al., 2014). Proper CAR models have the theoretical advantage of a proper joint distribution and can accommodate spatial correlation that is not extremely strong – ρ is not near or equal to one – and this might be regarded as bolstering the CAR model. However, proper CAR models are criticized for viewing the mean of the conditional distribution as some proportion of the average of the neighbors rather than a true average, ρ misrepresenting the true strength of spatial correlation, and,

if the prior value of ρ is inaccurate, the range of $\boldsymbol{\varphi}$ may be overestimated (Banerjee et al., 2014; Gelfand & Vounatsou, 2003). Finally, Banerjee (2014) argues that if one is to assume a large degree of spatial correlation (i.e., a ρ relatively close to one), the use of an intrinsic CAR model is preferred to avoid adding an additional parameter ρ.

The proper CAR model and intrinsic CAR model are two commonly cited choices for spatial modelling. However, there are numerous alternatives to choose from depending on what assumptions are appropriate for the data and hypotheses of interest, such as the Besag York Mollié (BYM) model or Leroux mixture models (Duncan, Cramb, Baade, Mengersen, Saunders, & Aitken, 2020). For the simulations in this thesis, we will utilize the intrinsic CAR model for its relative simplicity, interpretability, and fit with our hypotheses of interest.

1.2 Estimating the Distribution of the Spatial Varying Coefficients

1.2.1 Bayesian Inference

Given some observed data, y, it may be of interest to make statements about the distribution of some parameter, θ. First, consider the joint probability distribution of y and θ,

$$p(\theta, y) = p(y \mid \theta)p(\theta)$$

then, applying Baye's rule, it can be shown that

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$

(Gelman, Carlin, Stern, & Rubin, 2014). The framework above is the general basis for Bayesian inference, and it is employed when distributions of θ are of interest (Gelman et al., 2014). Generally, $p(\theta)$ is referred to as the prior distribution of the parameter; $p(y \mid \theta)$ as the likelihood of observing the data given the parameter; $p(y)$ as the marginal

likelihood, which is commonly viewed as a normalizing constant; and $p(\theta \mid y)$ is the posterior distribution of θ given y, which is generally the distribution of interest (Gelman et al., 2014; Gómez-Rubio, 2020). Since the marginal likelihood can be viewed as a normalizing constant and is often difficult to evaluate numerically, the posterior distribution is sometimes presented in an unnormalized (kernel) form

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$

(Gelman et al., 2014; Gómez-Rubio, 2020).

A key component of Bayesian inference is specifying the prior distribution of the parameter. The prior distribution might reflect some pre-existing knowledge about the parameter, be a distribution that captures some intrinsic nature of the parameter (e.g., must be in the interval [0, 1]), or, if there is very little assumed knowledge about the parameter, be a non-informative prior distribution such as a uniform distribution (Gelman et al., 2014). In addition to the prior distribution, it is also important to carefully consider the structure of the data and how the likelihood should be calculated. After all, the unnormalized posterior distribution is simply the prior times the likelihood (Gelman et al., 2014).

It is important to note that this framework is not restrained to identifying probabilities or distributions of a single parameter, θ, and that Bayesian inference can be applied to find the posterior distribution of any number of parameters, $\boldsymbol{\theta}$, in the parameter space (Gómez-Rubio, 2020). Furthermore, the marginal distribution of a given parameter, $\theta_j, j = 1, \ldots, \dim(\boldsymbol{\theta})$, can be found by integrating out all other parameters, θ_{-j},

$$p(\theta_j \mid y) = \int p(\boldsymbol{\theta} \mid y) d\boldsymbol{\theta}_{-j}$$

(Gómez-Rubio, 2020).

1.2.2 Integrated Nested Laplace Approximation (INLA)

The Bayesian framework provides an excellent methodology to make inference about parameters and their distributions. However, complications can arise when computing components such as the marginal likelihood or the posterior distribution, especially if it is multivariate in nature or has no closed form (Gómez-Rubio, 2020). Estimation methods and sampling techniques, including popular Markov chain Monte Carlo methods, can be applied to assist with estimating the posterior distribution. However, many of these methods are computationally slow for high dimensional parameter spaces, which was the impetus for the development of INLA (Gómez-Rubio, 2020).

To model with INLA, three assumptions are needed: the total number of hyperparameters is relatively small, usually in the range of 2 to 5; the distribution of the latent field $\mathbf{x}|\boldsymbol{\theta}$ is, at least approximately, a Gaussian Markov random field (GMRF); and the data, $\boldsymbol{y}$, are mutually conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}$ (Rue et al., 2017). Here, $\boldsymbol{x}$ generally denotes the latent effects about the observed data, $\boldsymbol{\theta}$ represents hyperparameters governing the model, and $\boldsymbol{y}$ is the outcome being modelled. Giving such a model framework under these assumptions allows for the fast computational speed that INLA possesses over other techniques and for accurate estimation (Rue et al., 2017).

If the assumptions above hold, then the likelihood of $\boldsymbol{y} = \{y_1, \dots, y_n\}$ given $\boldsymbol{x}$ and $\boldsymbol{\theta}$ can be shown to be

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid x_i, \boldsymbol{\theta}) \text{ for } i = 1, \dots, n$$

and the desired posterior distributions are

$$p(x_i \mid \boldsymbol{y}) = \int p(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{\theta}$$

and

$$p(\theta_j \mid \boldsymbol{y}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-j}$$

(Banerjee, 2014; Gómez-Rubio, 2020). The INLA methodology as implemented by Rue et al. (2017) first computes an approximated $\tilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y})$ and then uses this approximation to find a subsequent nested Laplace approximation of $\tilde{p}(x_i \mid \boldsymbol{y})$ (Banerjee, 2014).

1.2.3 Deviance Information Criterion (DIC)

We have seen that spatial effects can be modeled with different prior distributions, for example, the intrinsic CAR model, the proper CAR model, the BYM model, or the Leroux model. Furthermore, a spatial effects parameter, $\boldsymbol{\varphi}$, can be integrated into a generalized linear model with any number of covariates, $\boldsymbol{\beta}$. Therefore, there is a need for some measure to aid in model selection for prior distributions and covariate pruning.

In the Bayesian framework, the DIC is defined as

$$DIC = -2 \log* p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}) + 2p_{DIC}$$

where $\widehat{\boldsymbol{\theta}}$ is the estimated posterior mean of $\boldsymbol{\theta}$ and p_{DIC} represents the number of effective parameters in the model (Gelman et al., 2014). The "better" the model fits the data, the larger the likelihood of observing $\boldsymbol{y}$ given $\widehat{\boldsymbol{\theta}}$ should be. However, over-fitted models with many parameters can exhibit "good fit," and so $2p_{DIC}$ acts as a penalization for unnecessary model complexity (e.g., unnecessary parameters). The DIC can then be used for model selection, with smaller DICs pointing towards the preferred model (Gómez-Rubio, 2020).

1.3 General Motivation

As discussed in the Introduction, modeling health care data with spatial techniques can provide insight into how patient health outcomes and health systems vary across geographic regions. For example, assessing the geographic distribution of social determinants of health (SDoH) and their effects across space might be of interest. SDoH are commonly cited as a major contributor to patient health outcomes ("Social determinants of health", n.d.). SDoH refer to a wide scope of forces shaping an individual's daily life conditions and include the contexts in which people are born, work, and live ("Social determinants of health", n.d.). An example of an SDoH would be lack of access to a sufficient amount of quality food (i.e., food insecurity) and individuals with such an SDoH need have been associated with higher health care expenditures, greater emergency department (ED) use, and longer hospitalizations (Berkowitz, Seligman, Meigs, & Basu, 2018).

Identifying patient SDoH needs and creating interventions may produce better patient outcomes and positively impact health systems (Garg, Homer, & Dworkin, 2019). An accurate and seamless method to identify patient needs and an effective intervention system to address them is needed (Thornton et al., 2016). One interesting example of a novel technique to identify SDoH was using natural language processing methods to automatically discern patient SDoH from digitized medical record notes (Hatef et al., 2021). The following section includes a description of an alternative SDoH identification and intervention method which was part of a pilot study at Prisma Health. The study aimed to assess the impact of SDoH on health care resource use and how these impacts vary geospatially. Data from the pilot work is the motivation for this thesis.

Chapter 2: Methods

In this chapter the Prisma Health pilot study and the processing of its data is discussed, and a general model setup is given. Additionally, the preparation of the data for analysis via INLA and implementation of INLA for inference is discussed. All data work was performed in R version 4.2.1 using the interface of RStudio version 2022.07.2 (R Core Team, 2022; R Studio Team, 2022). For parameter estimation, the inla(…) function from the INLA package version 22.05.07 was used (Rue, Martino, & Chopin, 2009; Martins, Simpson, Lindgren, & Rue, 2013).

2.1 Prisma Health Pilot Study and Data Processing

2.1.1 Prisma Health Pilot Study

A pilot study was conducted by Prisma Health from June 1, 2019-December 31, 2020. Over this study period, Prisma Health trialed a SDoH screening and referral software program, NowPow, which linked patients with community-based organizations (CBOs) depending on their SDoH needs and geographic location. Here, the intervention is the referral of patients to CBOs, with the idea that such organizations could aid in SDoH need reduction. The sample for the pilot study included Prisma Health patients 18 years or older in South Carolina's (SC) central and northwestern regions who engaged in inpatient case management, ambulatory care and condition management, or community health programs.

In addition to the data available from the NowPow system, electronic medical record (EMR) sources were leveraged and patient information was linked between the

two sources via medical record numbers. The primary data from NowPow included patient responses to a 13-item questionnaire designed to screen for SDoH needs commonly associated with health outcomes. The SDoH categories contained in the screener included food insecurity, housing instability and quality, financial instability, transportation needs, interpersonal violence/abuse, language and health literacy, and social connectedness. The EMR sources included general demographic information (e.g., race/ethnicity, gender), heath system use (e.g., the number of ED visits), and, importantly, the address for a given patient. Given the available data, the pilot study considered the question of how a patient's SDoH needs were associated with ED use and how this association varied across geographic regions. More specifically, it was asked: after adjusting for all other variables, does the association between ED visits after the screen and a positive screening for an SDoH need vary geospatially?

2.1.2 Address Geocoding

The study sample included 3,016 patient observations. Patients may have been administered the NowPow screening questionnaire multiple times over the study period, and only the earliest screening information available for a given patient, called an index screen (2,687, 89.1%), was used in analysis.

The geocode function from the tidygeocoder package in R was used to link patient addresses to a latitude-longitude coordinate pair (Cambon, Hernangómez, Belanger, & Possenriede, 2021). The geocode function can take addresses as a single line of information or by splitting the address information into the components of street number and name, city, state, and ZIP code.

For this study, if the single address input failed to geocode, the split option was subsequently attempted. There are multiple geocoding services compatible with the function via the 'method' input. The 'census' method was used for the data, which draws upon the United States census geocoding information. Patient addresses which successfully geocoded were selected (2,490, 92.7%) and further processed by removing obvious address miscodings, addresses outside of SC, and observations which resided in a SC county having less than 50% of its block groups containing observations (295, 11.8% removed; 2195, 88.2% used in analysis). The final set of observations used for analysis were contained in 10 neighboring counties and included 1,110 census block groups within those 10 counties.

2.1.3 Variables of Interest and Additional Data Processing

The response variable was chosen to be the number of ED visits after the index screen date, which was a zero-inflated count variable. Patients may have had different index screening dates and, thus, a different number of days after the index screen and before the end of the study period over which ED visits could have occurred. To account for this, an offset term was created and set to be the natural log of the number of days between the index screen and the end of the study period.

The primary independent variable of interest was the patient response to the SDoH need screening question. Some screening questions had only a 'Yes' or 'No' response choice while others included multiple response categories, such as 'Never,' 'Rarely,' 'Sometimes,' 'Often,' and Always.' Questions with multiple choices were re-categorized to only have two levels equivalent to a 'Yes' or 'No' response set, with the 'Rarely' and 'Sometimes' corresponding to the 'No' category.

Adjustment for nine additional covariates was included in the model. The other variables considered were race (White/Black/Other/Patient Refused), pregnant (Yes/No), Hispanic (Yes/No), primary payer (Medicaid/Medicare/Other), smoking status (Yes/No), weight category according to BMI (Underweight/Healthy/Overweight/Obese), female (Yes/No), total number of comorbidities (0-9), and age at the time of the screen.

As the response variable was a zero-inflated count variable, regression models that were deemed as sensible fits for the data were zero-inflated Poisson, Poisson hurdle, zero-inflated negative binomial, and negative binomial hurdle regression model, all using a natural log link function. For each of these models, we assumed the expected number of ED visits after the screen for an individual i was given by $\exp(\eta_i)$,

$$\eta_i = t_i + \boldsymbol{x_i}\boldsymbol{\beta} + r_i\varphi_{s_i}$$

where t_i is the offset term defined above; $\boldsymbol{x_i}$ is a vector of the nine additional covariates considered; $\boldsymbol{\beta}$ is the corresponding vector of covariate coefficients; r_i is the response to the SDoH screening question; and φ_{s_i} is the spatial coefficient for the block group to which the given patient address was associated. Furthermore, $r_i = 1$ if the patient response was 'Yes' to the SDoH screening question and $r_i = 0$ otherwise. In this way, the model allows patients in different block groups who responded 'Yes' to the SDoH screening question to have different predicted mean numbers of ED visits after the screen, after adjusting for all other variables. However, holding all other variables constant, individuals from different block groups who both responded 'No' to the SDoH screening question did not have the flexibility to possess different spatially varying coefficients.

2.2 Using INLA to Model the Prisma Health Data

It was decided an intrinsic CAR model would be used as the prior model for the spatially varying parameter. To conceptualize the intrinsic CAR model, the geographic region G was defined as the 1,110 block groups contained within the 10 counties for which there was patient data. Thus, the spatial parameter $\boldsymbol{\varphi} = \{\varphi_{s_1}, \ldots, \varphi_{s_{1110}}\}$. Next, the binary adjacency matrix $\boldsymbol{W}$ was needed. First, a polygon representation of the 1,110 block groups for the data was obtained via the block_groups(…) function from the R tigris package, which utilizes data from the U.S. Census Bureau, and the desired $\boldsymbol{W}$ matrix was created by subsequently using the poly2nb(…) and nb2mat(…) functions from the R spdep package (Bivand, 2022; Walker, 2022).

Next, each patient was linked to a block group. Recall that each patient observation contained a latitude and longitude coordinate pair. The patient data was linked to the block group spatial data frame by utilizing the st_join(…) function from the R sf package (Pebesma, 2018).

After obtaining the adjacency matrix and linking patient data to block group information, a model was created for parameter estimation using INLA. The number of ED visits after the index screen was the response variable and the model was fit with the nine covariates (from Section 2.1.3) as fixed effects and the response to the screening question as having spatially varying coefficients. As noted above, the spatially varying coefficients were assumed to follow an intrinsic CAR model. The sum to zero constraint was not imposed on the spatial coefficients.

2.3 Model Selection

Again, four different regression types, zero-inflated Poisson, Poisson hurdle, zero-inflated negative binomial, and negative binomial hurdle, were considered appropriate for the data and each could be applied to the model setup given in the previous section. It was decided that one model type should be chosen. To do this, a crude model formula was created for each screening question, $i = 1, \ldots 13$, which modeled the number of ED visits after the index screen solely by the response to the screening question, which had a spatially varying coefficient. The four different regression types were run for each of the crude screening question models and the DIC was calculated. It was found for all screening questions that the zero-inflated negative binomial produced the lowest DIC. Based on the DIC, the zero-inflated negative binomial was indicated to have the best fit for the data and was chosen as the regression type for the simulations.

Further model selection occurred by considering covariates for removal. Backwards stepwise selection was performed upon the model for each screening question using zero-inflated negative binomial regression. To start, the DIC was calculated for the full model which included all nine covariates and the response to screening question spatially varying coefficient term. Then, the DIC was calculated under the consideration that a single covariate was removed; this was done for all nine covariates. A covariate was officially removed if it lowered the model DIC by at least 2. If multiple covariates lowered the DIC past this threshold of two, then the variable that lowered the DIC to the greatest magnitude was removed. After the removal of a variable, the process was performed again for all remaining covariates and continued until no variable met the

removal criterion. Note, neither the spatially varying coefficient term nor the intercept was considered for removal.

Additionally, it was decided that a variable to capture the percentage of follow-up time that fell within the COVID-19 lockdown period in SC was needed in the model. The lockdown period was defined from March 15, 2020 to June 11, 2020. March 15th represented the date that public schools were initially closed in SC and June 11th was the date Governor McMaster signed executive order 2020-40, which lifted many of the gathering restriction in public spaces and stores in SC. The variable was calculated as the number of days of patient follow-up time that occurred between the two dates divided by the total number of days of patient follow-up time multiplied by 100%.

The idea to include such a variable in the model occurred after backwards stepwise selection was performed on the nine original covariates. Because the impact of the beginning of the COVID-19 pandemic and lockdown was such a pervasive and novel factor in SC, it was determined that this variable should be included in the model regardless of whether it was statistically important. Because the variable was primarily viewed as a needed adjustment within the model, it was simply appended to the model found by backwards stepwise selection on the nine original covariates.

2.4 Inference using the INLA Model Output

In Section 2.1.1, it was stated that a primary question of interest from the Prisma Health pilot data was: adjusting for all other variables, does the association between ED visits after the screen and a positive screening for an SDoH need vary geospatially? To answer this question, consider that the INLA output provides the 0.025 and 0.975 quantiles of the estimated spatially varying coefficients. From a Bayesian perspective, we

can construct a 95% credible interval from these two values as (estimated 0.025 quantile, estimated 0.975 quantile) for each estimated spatially varying coefficient.

Credible intervals considered in Bayesian inference are distinct from confidence intervals often used in frequentist statistics. From a conceptual statement, a 95% confidence interval can be interpreted as "if we collected every possible sample in this given manner and constructed 95% confidence intervals for each sample, then the true parameter would be contained in 95% of those confidence intervals." After observing the data, confidence intervals do not address any probabilities about whether the true parameter is contained in any single one of the confidence intervals. On the other hand, the Bayesian 95% credible interval allows for an interpretation of "given the observed data, there is 95% probability that the true parameter is contained within this credible interval."

Despite this distinction, Bayesian 95% credible intervals are used for inference very similarly to frequentist 95% confidence intervals. Again, consider constructing 95% credible intervals for all 1,110 spatially varying coefficients associated with the entire set of block groups. We can select any two of these credible intervals and assess whether their credible intervals overlap. If they do not overlap, then we can conclude that we find a statistically important difference between the two true spatially varying parameters associated with these two block groups. Moreover, if we wanted to assess that statistically important spatial variation is present across the *entire* geographic region of interest, we could compare all the 95% credible intervals for the 1,110 estimated spatially varying coefficients and determine if there are at least two intervals which do not overlap. To accomplish this parsimoniously, we can compare if the maximum 0.025 quantile

credible bound is greater than the minimum 0.975 quantile credible bound. If it is greater, then we can conclude that at least one pair of credible bounds do not overlap and that true spatial variation exists over the geographic region.

2.5 Motivation for Simulations

This chapter detailed how the INLA estimation method could be implemented in R specifically for the Prisma Health pilot data. Furthermore, a procedure to answer the question, "Does the association between ED visits after the screen and a positive screening for an SDoH need vary geospatially?", was provided. It is useful that the INLA methodology and relevant R packages allow for spatial analysis of large and sometimes complex datasets. However, creation of models that utilize an estimation technique such as INLA begs the question of: How well does the technique work for such a scenario as the Prisma Health pilot data?

Consider again the general model for the Prisma Health pilot data

$$\eta_i = t_i + \boldsymbol{x_i}\boldsymbol{\beta} + r_i\varphi_{s_i}.$$

For this model setup, we can consider 13 specific models: one for each of the 13 SDoH screening questions. A brief descriptive analysis of the Prisma Health pilot data reveals that, after dichotomizing the screening questions into "Yes" or "No" levels as discussed in Section 2.1.2., the percentage of positive (i.e., "Yes") screens varied across SDoH needs. For example, one question that targeted food insecurity had 252 (11.5% of total patients contained in the 1,110 block groups) positive screens and a different question for housing insecurity possessed 89 (4.1%) positive screens. At the very end of the spectrum, the question for domestic violence/abuse had only 14 (0.6%) positive screens.

One question that could be answered in a simulation study would be: assuming the association between a positive screen and ED visits after the screen does truly vary across geographic sub-regions, what is the power in picking up that difference using the INLA methodology outlined above? Furthermore, does this power change for different numbers of positive screens as seen in the pilot data? Another question is, conversely: assuming the association between a positive SDoH screen and ED visits after the screen does not vary across geographic sub-regions, what is the error in concluding that there is spatial variation? The following chapter on simulations considers these questions, among others, and describes the process in which they were answered.

Chapter 3: Simulation Study

In this chapter, a single model from the Prisma Health pilot data is selected. Next, a general setup for simulation work is provided and the details of three different simulations of interest are explained.

3.1 Screening Question Model of Interest

It was determined that restricting the context of simulations to a single model would aid in the simplicity and interpretability of the simulation studies. The previous chapter detailed that the zero-inflated negative binomial model was found to have the best fit, as evidenced by DIC, and that variable selection was performed for each of the screening question models. The model for the domestic violence screening question was selected for evaluation in the simulation study and this question had the lowest positive response rate out of all the study questions. The covariates included in the final model for this screening question were smoker (Yes/No), female (Yes/No), and primary payer (Medicare/Medicaid/Other), age at the time of the index screen, and the time of follow-up that fell in the defined COVID-19 lockdown period.

3.2 General Simulation Setup

The end goal of the general simulation setup was to simulate a vector of values representing ED visits after the screen and for this vector to closely represent the true distribution of ED visits after the screen from the Prisma Health pilot data. Again, consider the model

$$\eta_i = t_i + \boldsymbol{x_i}\boldsymbol{\beta} + r_i\varphi_{s_i}.$$

After the model selection procedure outlined in Section 3.1, the above model is now defined with t_i representing the natural log of the number of days after the index screen for a given patient; $\boldsymbol{x_i}$ representing the patient data for smoking status, age at the time of the index screen, female/not female, primary form of payment, and the COVID variable; $\boldsymbol{\beta}$ is the vector of coefficients associated with each of these variables; r_i indicates whether or not the patient screened positive for violence/abuse; and φ_{s_i} is the associated spatially varying coefficient of the block group in which the patient resides.

For the simulation study, the values of t_i and $\boldsymbol{x_i}$ were taken directly from the Prisma Health pilot data. To set values for $\boldsymbol{\beta}$, the inla(…) function was run for the observed patient ED visits after the screen data and the estimated $\widehat{\boldsymbol{\beta}}$ was recorded. Then, these $\widehat{\boldsymbol{\beta}}$ values were rounded for simplicity, giving the simulation $\boldsymbol{\beta}' = [-6.0, 0.4, 0.003, -0.4, 0.3, 0.3, 0.2, -0.1]$. The r_i vector of screening responses was not taken from the patient data but was instead simulated. This was done as some simulation questions required different rates of positive screens not found in the data. The desired number of positive screens were selected from among the entire set of individuals using simple random sampling without replacement.

Next, the vector of spatially varying coefficients was simulated. The intrinsic CAR model was assumed as the prior distribution and, thus, the vector of spatially varying coefficients was assumed to follow a multivariate normal (MVN) distribution, from Section 1.1,

$$\boldsymbol{\varphi} = \left(\varphi_{s_1}, \ldots, \varphi_{s_n}\right)^T \sim MVN(\mathbf{0}, \tau^2(\boldsymbol{D} - \rho\boldsymbol{W})^{-1})$$

where $\rho = 1$. However, it was noted that centering the MVN distribution on zero limited the strength of the signal of the spatially varying coefficients, and that this might limit the

practical ability of the INLA methodology to estimate any differences in the coefficients. Because of this, the MVN for the simulations was centered at three, giving

$$\boldsymbol{\varphi} = \left(\varphi_{s_1}, \ldots, \varphi_{s_n}\right)^T \sim MVN(\mathbf{3}, \tau^2(\boldsymbol{D} - \rho\boldsymbol{W})^{-1}).$$

The required covariance matrix, $\tau^2(\boldsymbol{D} - \rho\boldsymbol{W})^{-1}$, was calculated in the following manner. First, τ^2 was assumed to equal 1. Then, the $\boldsymbol{W}$ matrix was created as described in Section 2.2.1. The $\boldsymbol{D}$ matrix was then created from the $\boldsymbol{W}$ matrix. Next, ρ was set to equal 0.9999. In an "ideal" intrinsic CAR model, ρ should equal one; however, doing so would result in the invertible matrix complication described in Section 1.1. Therefore, the value of 0.9999 was chosen as it was extremely close to one and, for practical purposes, emulated the intrinsic CAR model. With all the necessary matrices and values set, the spatially varying coefficients were simulated with a MVN distribution.

Now, all the required components to simulate ED visits after the screen, $\exp(\eta_i)$, were at hand considering our model,

$$\eta_i = t_i + \boldsymbol{x_i}\boldsymbol{\beta} + r_i\varphi_{s_i}.$$

The regression model type chosen was a zero-inflated negative binomial model with a log link. Data generation from this model required two steps. First, a proportion, p, of the patients were assumed to not be able to experience an ED visit after their index screens and were assigned a zero for the variable. For each patient, a Bernoulli(0.1) random variable was generated. Patients for which this random variable was equal to one were assigned zero ED visits.

ED visits for the remaining patients were generated from a negative binomial distribution, with mean $\mu_i = \exp(\eta_i)$ for patient i and dispersion parameter $\phi = 0.1$.

The dispersion parameter value was found empirically by generating data using many different dispersion values and selecting one which produced data very similar to the observed ED visits after the screen data from the Prisma Health pilot study. Note that the randomly generated values here could also be equal to zero; that is, patients who were not automatically assigned a zero from the Bernoulli assignment step could be assigned a zero in this second step.

In conclusion, the general simulation procedure described above created a pathway to simulate the zero-inflated ED visits after the index screen variable in a manner such that it closely resembled the observed data from the Prisma Health pilot study. Furthermore, small changes can be made to the simulation methodology described above to pose and answer different questions regarding the performance of INLA in estimating the spatially varying coefficients. The following sections describe different simulation studies of interest and note any changes required in the procedure outlined in this section.

3.3 Simulations

3.3.1 No Spatial Variation: Type I Error

The general simulation procedure assumed that the vector of spatially varying coefficients followed an intrinsic CAR model and, because of this, that there was true spatial variation. However, consider if there was no true spatial variation. In the model,

$$\eta_i = t_i + \boldsymbol{x_i\beta} + r_i\varphi_{s_i},$$

this would be equivalent to each φ_{s_i} being equal to some number, say zero. However, assigning a zero to each φ_{s_i} not only assumes no spatial variation, but also assumes no association between the response to the screening question and number of ED visits.

Although the latter assumption is more directly related to the strength of association between variables and not spatial variation between block groups, it may affect the INLA estimation of spatial variation. Therefore, it was pertinent to assess type I error for another value other than zero.

To simulate type I error rate, the general simulation procedure was followed except that instead of assigning values to φ_{s_i} based off of a MVN distribution, each was assigned a zero for one set of simulations and a three for another set of simulations. The value three was chosen as it matched the centering of the MVN distribution when spatial variation was assumed to be present in subsequent power simulations.

These simulations were performed for the minimum and maximum number of positive screens found in the pilot data, 14 and 256, respectively. Additionally, simulations were performed for 550 and 1,100 positive screens, representing approximately a 25% and 50% positive screening rate in the entire sample, respectively.

3.3.2 Differing Number of Positive Screens: Power

Another question of interest was whether the power of our model to detect spatial variation across block groups differs depending on the number of positive screens in the data. To assess this, the number of positive screens was set to 14, 256, 550, and 1,100, as was done in the type I error simulation. Data was simulated for each of these different number of positive screens and the vector of spatially varying coefficients was assumed to follow a MVN distribution with mean equal to three and variance parameter of one.

3.3.3 Differing Magnitude of Spatially Varying Coefficients: Power

In the general simulation procedure, the values of each φ_{s_i} was simulated under the assumption that τ^2 was equal to one. Changing the value of τ^2 would be directly

changing the variance of the MVN distribution from which the φ_{s_i}'s are simulated. In this simulation study, the values of τ^2 were increased or decreased and included values of 0.25, 1, 2 and 3. Note that data for τ^2 equal to one is available from previous simulations. This simulation was performed for the number of positive screens being set to 14, 256, 550, and 1,100.

3.3.4 Assessing the Simulations

For each scenario described above, 250 datasets were generated for each sample size. During each of the 250 dataset generations, the assignment of positive screens and number of ED visits after the screen to each patient as well as the values of the random spatial effects were re-simulated.

The number of times comparisons of the 95% credible intervals for the spatially varying coefficients indicated statistically important spatial variation was divided by the total number of datasets. For the simulation on type I error where data was simulated such that there was no spatial variation, this value estimated the type I error. For all other simulations, this value provided an estimate of model power.

Chapter 4: Simulation Results and Discussion

4.1 Simulation Results

The results of the type I error simulations are shown in Table 4.1. The type I error remained very low when the entire vector of spatial coefficients was set to both zero and three and it was similar for the different numbers of positive screens. In fact, only one type I error occurred when the coefficients were set to zero and the number of positive screens was 256. The type 1 error for all other simulation sets was zero.

Table 4.1 Type I Error for setting the spatial coefficients to zero and three

	Proportion of simulations with type I error	
Number of positive screens	Coefficients set to zero	Coefficients set to three
14	0.00	0.00
256	0.004	0.00
550	0.00	0.00
1,100	0.00	0.00

The results of the power simulations for different numbers of positive screens and for differing values of the variance parameter, τ^2, are summarized in Table 4.2. Unlike the type I error, the power did change for different numbers of positive screens with larger numbers of positive screens exhibiting higher power. For the number of positive screens at 256, 550, and 1,100, greater statistical power in detecting spatial variation was found for larger values of τ^2. However, this was not the case for the number of positive screens being set to 14. Even for τ^2 set equal to three, the largest variance parameter value used for simulation, spatial variation was not detected for any simulation for which the number of positive screens was 14. The mean of the MVN distribution from which

the spatially varying coefficients were simulated was set to be three for all power simulations.

Table 4.2 Power across different number of positive screens and different values of τ^2

	Power: proportion of simulations exhibiting spatial variation			
Number of positive screens	$\tau^2 = 0.25$	$\tau^2 = 1.00$	$\tau^2 = 2.00$	$\tau^2 = 3.00$
14	0.00	0.00	0.00	0.00
256	0.004	0.06	0.2	0.296
550	0.032	0.312	0.44	0.744
1,100	0.14	0.724	0.964	0.984

4.2 Discussion of Simulation Results

The results of the type I error simulations are encouraging for the use of INLA to estimate spatially varying coefficients. All but one of the proportions of type I error for the different simulation sets was zero. Moreover, the simulation set that did produce a non-zero type I error proportion was extremely low at 0.004, with only one type I error produced over 250 simulations. From these simulation results, we can be confident that we have a very small probability of erroneously concluding there is spatial variation when there is actually no spatial variation present. It is especially encouraging that this is true for all the different numbers of positive screenings that were used for simulations.

Generally, maintaining a very low type I error rate might come at the expense of lower power. In our case, that would imply we would be less likely to detect spatial variation when there actually is spatial variation present. From the power simulations, it is apparent that power is lacking when the number of positive screens is both 14 and 256. For the number of positive screens set to 14, the INLA methodology did not detect spatial variation for any of the simulations for any value of τ^2. This finding is discouraging, although it is not entirely surprising as 14 positive screens represents a 0.64% positive

screening rate. It would also be surprising to consistently detect an association between a positive screen and number ED visits and, furthermore, to conclude the associations vary across block groups when only provided with 14 positive screens. However, such a low positive screening rate did occur as the lowest rate in the real Prisma Health data and, thus, we would like a methodology that can work with such situations.

The maximum number of positive screens for an SDoH need in the Prisma Health data was 256. The INLA methodology did have higher power for this number of simulated positive screens as compared to 14 positive screens, specifically for larger values of τ^2. However, even at the largest τ^2 value of 3, the simulation study only produced a power of 28% at 256 positive screens. Although this is a drastic improvement from a power of 0% found for 14 positive screens at $\tau^2 = 3$, it is low considering that 256 was the highest number of positive screens found in the Prisma Health data. We would have wanted to observe a much larger power for the highest positive SDoH screening rate we might realistically see in a study population.

The power from the simulations for both 550 and 1,100 simulated positive screens increased as the variance parameter increased in magnitude. In general, the power was noticeably greater than that of 14 and 256 positive screens. For 550 positive screens, the variance parameter needed to be relatively high at three to observe a desirable power of 76.5%. In contrast, the simulations for 1,100 positive screens achieved 70% power when the variance parameter was set to one. At both τ^2 equal to two and three with 1,100 simulated positive screens, a very large power of over 95% was achieved. In this way, the 1,100 positive screen scenario does not appear to benefit greatly from increases in the variance parameter past a value of two. In contrast, both 256 and 550 did show marked

improvement when τ^2 was increased from two to three. Combined with the fact that the type I error for 1,100 positive screens was found to be zero percent, it is apparent the INLA methodology works very well for a categorical variable with responses split relatively evenly for a total sample size of around 2,200.

Desirable levels of power were only found for 550 and 1,100 positive screens, neither of which were actually observed in the Prisma Health SDoH screening and both of which were well above the maximum 256 positive screens that was observed. Assuming that the rates for SDoH needs do not typically extend up to levels such as 25% and 50%, the findings of the power analysis indicate that detecting spatial variation for SDoH screenings variables may be difficult, or even practically impossible for very rare SDoHs, noting 0% power for 14 positive screens at all values of τ^2. While not explored in this simulation study, it is possible that a large sample size could produce enough positive screens to give sufficient power to detect spatially varying coefficients, even under relatively low positive screening rates typically observed in practice.

Although 550 and 1,100 positives screens were not observed in the Prisma Health data, the values and their respective power analysis should not be ignored. Conceptually, a "positive screen" is simply a binary indicator and the results of these numbers of "positive screens" can be viewed in the light of variables that might actually follow such a distribution of yes's and no's or 0's and 1's. For example, 1,043 patients were categorized as obese based on their BMI in the Prisma Health data. A spatial analysis of the impact of being categorized as "obese" or "not obese" would be based on an approximately 50% categorical variable split. In this way, the results of the 1,100 positive screens simulations would be encouraging that the INLA model would perform very well

in detecting spatial variation of the association between weight category and some highly zero-inflated outcome of interest. Similarly, the results of the 550 positive screens simulations would be encouraging for a binary variable that follows an approximately 25/75 split.

In summary, power was noted to increase for a given sample size as the magnitude of τ^2 increase. This seems reasonable because as τ^2 increases, the magnitude of the differences between the simulated spatially varying coefficients increases. To illustrate this, the vector of spatially varying coefficients was simulated 250 times setting τ^2 to each of 0.25, 1, 2, and 3. The incident rate ratio (IRR) of the maximum and minimum φ_{s_i} coefficient across all samples was calculated and the median of the distribution of IRR's for a given τ^2 value are shown in Table 4.3. The median was chosen instead of the mean as the distributions of the IRR's were positively skewed.

Table 4.3 Median minimum and maximum IRR for increasing τ^2 values

τ^2	Median minimum IRR	Median maximum IRR
0.25	5.081	69.905
1	1.376	239.265
2	0.636	860.067
3	0.275	2318.813

From the table, the median minimum IRR tends to zero and the median maximum IRR increases as the value of τ^2 increases. In this way, the spread of the spatially varying coefficients, and thus the respective IRR's, becomes greater in magnitude as τ^2 increases. The median maximum IRR is extremely large, even at its smallest when $\tau^2 = 0.25$. Considering that for each simulation the vector of spatially varying coefficients consists of 1,110 φ_{s_i}'s and we exponentiate the largest one to get the IRR, it is sensible that we would find enough large coefficients that the distribution of the outputted exponentiation

would consist of very large numbers. Furthermore, the MVN distribution was centered on three; in this sense, the assumed mean would suggest an IRR of eight. However, this does not imply that these IRR values are realistic.

Additionally, the power increased for a given value of the variance parameter as the number of simulated positive screens increased. This also is sensible because as more positive screens are observed, more block groups are associated with a spatially varying coefficient. In turn, the INLA model has more block group data from which to detect spatial variation.

4.3 Limitations of Simulations

The power simulations are limited by setting the mean of the simulated MVN vector of spatially coefficients to three. Conceptually, setting the mean to a positive number was sensible under the hypothesis that answering 'yes' to the violence/abuse SDoH screener would be associated with more ED visits. However, the level of power observed from the simulations, especially for higher values of τ^2, may have benefited from setting the mean to three; setting the mean to something smaller in magnitude such as one might have resulted in less simulated power.

The values of τ^2 were chosen such that changes in power could be observed across the simulations. However, the information from Table 4.3 indicates that the IRR's that would be implied by the larger values of τ^2 would not be realistically observed in real data applications for this given geographic setup. It may be possible that a different geographic region setup, and thus different $\boldsymbol{D}$ and $\boldsymbol{W}$ matrices, would produce more realistic simulated IRR's for the τ^2 values used in this simulation study.

A general limitation of the simulation is that the spatial distribution of the patients and the neighborhood system of the 1,100 block groups were taken directly from the Prisma Health data. Notably, the spatial distribution of the patients were centered around urban cities. Although using the pilot data setup allowed the simulation to be practical for what may be observed in real life, it also impacted the simulation as the neighborhood system directly influences the $\boldsymbol{W}$ and $\boldsymbol{D}$ matrices used to simulated the MVN vector of spatially varying coefficients.

Furthermore, since patients were clustered around cities, there were some block groups without patient observations. The INLA estimation procedure can handle empty regions, but we should be cautious about how accurate we view such estimates. Relatedly, it should be noted that the simulation study focused on the INLA methodology's ability to detect variation across the block groups. The simulation study did not assess how biased the estimates of the spatially varying coefficients were. For example, it is possible to correctly conclude that spatial variation exists while also having biased estimates for the spatially varying coefficients.

Chapter 5: Prisma Health Data Application

The INLA methodology was applied to the Prisma Health pilot data and the model of interest was the same as outlined in Chapter 3 and which served as the foundation of the simulation study. Specifically, the response to the violence and abuse SDoH screening question was assumed to have spatially varying coefficients and the reference level was 'No', which indicated a negative screen. From the total Prisma Health sample, 14 individuals screened positive for this SDoH need. The outcome variable of interest was the total number of ED visits after the index screen and the other covariates included in the model were smoker (Yes/No), female (Yes/No), primary payer (Medicare/Medicaid/Other), age at the time of the index screen, and the percent of follow-up time that fell between March 15th, 2020 and June 11th, 2020 to capture the effect of COVID-19 lockdown in SC. An offset term set to be the natural log of the number of days after a patient's index screen was included and a zero-inflated negative binomial regression with a log-link was used. The vector of spatially varying coefficients was estimated with INLA. As with the simulations, all calculations and analysis were performed in R.

5.1 Results

There were two major model estimates of interest: the fixed effects estimates and the estimated vector of spatially varying coefficients. The estimates of the five non-spatial covariates included in the model are summarized in Table 5.1 which includes the estimated mean and upper and lower bound of the 95% credible interval.

Table 5.1 Fixed effects

Effect	Mean	0.025 quantile	0.975 quantile
Constant (intercept)	-4.981	-6.920	-3.161
Tobacco use (yes)*	0.434	0.098	0.748
Age	0.003	-0.008	0.014
Female*	-0.361	-0.652	-0.069
Primary payer: Medicare	0.296	-0.527	1.117
Primary payer: Other	0.152	-0.723	1.000
Primary payer: Missing	0.305	-0.563	1.169
COVID-19 time %	-0.107	-0.248	0.043

*indicates a statistically important covariate

Covariates which had their entire 95% credible interval entirely above or below zero were categorized as statistically important in predicting the mean number of ED visits per day after the index screen. From the model, the variables ‘tobacco use’ and ‘female’ were statistically important and their estimated means can be interpreted in the context of an IRR of mean ED visits per day after the index screen. Specifically, we estimate that individuals who use tobacco experience 1.543 times the incidence of ED visits per day after the index screen as compared to individuals who do not use tobacco. We also estimate that individuals who are female experience 0.697 times the incidence of ED visits per day after the index screen as compared to non-female individuals.

The estimated vector of spatially varying coefficients was used to conclude whether the association between answering ‘yes’ to the violence/abuse SDoH and the mean number of ED visits per day after the index screen varied across block groups. The maximum 0.025 quantile estimated for all spatially varying coefficients was 0.558, and the minimum 0.975 quantile estimated for all spatially varying coefficients was 1.285. Since the maximum 0.025 quantile estimate was not above the 0.975 quantile estimate, we do not have enough evidence to conclude the association of interest varied across block groups.

Although the model did not suggest spatial variation, the estimates of the spatially varying coefficients for each block group may be important. If there is a specific block group or set of block groups of interest, then these estimates can be specifically drawn from the model output and viewed. It may also not be practical to view each individual block group coefficient alone, but the associations between answering 'yes' to the SDoH violence/abuse question and mean number of ED visits after the index screen for all block groups can be visualized with a map. As was done with the fixed effects coefficients, the IRR for each of the block groups was calculated and are shown in Figure 5.1.

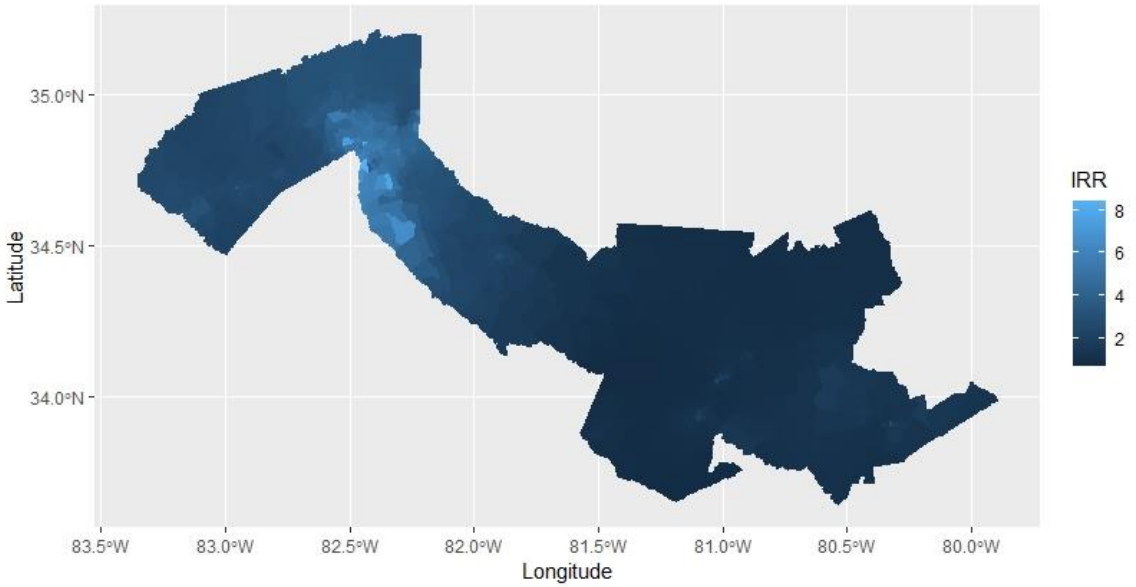

Figure 5.1 IRR for ED visits and positive violence/abuse screen across block groups

From the map, we can see that the larger IRR values tended to cluster in block groups in the more northern and west block groups.

5.2 Limitations

As noted in section 4.3, the patient observations were spatially distributed around larger towns and cities. In fact, there were 267 block groups with no patient data.

Although the INLA estimation procedure can handle empty sub-regions, we should not be as confident for estimates for block groups that do not have patient observations. Furthermore, only 14 patients screened positive for the violence/abuse SDoH variable for which spatially varying coefficients were estimated. Each of these patients resided in a different block group which allowed the model to utilize as much spatial information as possible. Nonetheless, we should be cautious in estimating spatial coefficients for 1,110 block groups when only 14 block groups contain observations from the covariate of interest. Relatedly, this would negatively impact whether these estimates capture the true spatial variation of the association between the SDoH question and mean number of ED visits per day after the index screen across block groups.

Finally, there is the limitation of generalizability to other populations outside of the study's geographic region of interest and individuals outside the types of health system care management on which the data was collected. Furthermore, since there was only data for 14 individuals with the violence/abuse SDoH need, there is a lack of generalizability of these results to other individuals who experience violence/abuse SDoH needs. Given the small number in the sample, we are not confident that the characteristics and demographics of these 14 individuals are representative of all individuals with such an SDoH need.

Chapter 6: Conclusions

Generalized linear models with spatially varying coefficients are flexible tools for analyzing health care data. Furthermore, such models allow one to determine if the association between a predictor and outcome variable of interest varies across geographic space. As health care systems become increasingly incentivized to study and intervene on SDoH, spatial models will provide an avenue to answer practical questions of interest regarding associations between specific SDoH needs. However, running such spatial models can be computationally expensive. The INLA estimation procedure offers a relatively efficient methodology to analyze spatial models. Given its potential for widespread use, an analysis of the INLA procedure under a variety of situations is warranted.

The simulation study in Chapter 4 was built upon data collected from a pilot study by Prisma Health which focused on SDoH needs. The goal of the simulation study was to assess the power of the INLA procedure in detecting true spatial variation in the association between responding ‘yes’ to an SDoH screener question and the mean number of ED visits per day after a screen, which was a highly zero-inflated variable. Furthermore, the type I error was also assessed. The simulation study found that the rate of type I error when using the INLA procedure was extremely low, practically zero, for both very small and large numbers of simulated positive SDoH screens.

Regarding power, it was found that as the variance parameter that governs the vector of spatially varying coefficients increases, the power of the INLA procedure also

increases. This was particularly true for larger values of simulated positive screens. However, for very low numbers of positive screens, such as the 14 positive screens observed in the pilot data, the INLA model was not able to detect spatial variation for any simulation. Furthermore, for a given value of the variance parameter, the power in detecting spatial variation increased as the number of simulated positive screens increased.

In the practical application of INLA in analyzing the Prisma Health pilot data, it was found that both tobacco use and being categorized as female were statistically important in predicting the mean number of ED visits per day after the index screen. By comparing the 95% credible intervals for the estimated spatially varying coefficients, it was found that the association between responding 'yes' to the violence/abuse SDoH and the mean number of ED visits per day after the index screen did not vary across block groups.

Both the simulation study and practical application were limited by the spatial distribution of the patients tending towards larger towns and cities. Furthermore, there are issues of generalizability of the study due to the specific geographic region of interest, forms of care management from which data was collected, and extremely low number of positive screens to the SDoH question of interest. Additionally, in the simulation study it should be noted that the multivariate normal vector of spatially varying coefficients was assumed to have a mean of three and that the IRR's implied by the values of τ^2 used for simulation were unrealistic. Centering the distribution on a value lower in magnitude and using smaller values of τ^2 would likely decrease power.

Future simulation studies should focus on regions with a more equal distribution of subjects across space and under the assumption of a weaker spatial signal for the vector of spatially varying coefficients. Despite these limitations, the current simulation study showcases the low type I error and high power, especially for larger numbers of positive screens, of the INLA procedure. And the practical application demonstrates the ease in which the INLA procedure can be applied to real life data.

References

Banerjee, S., B. P. Carlin, and A. E. Gelfand. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Bauer, J., Brüggmann, D., Klingelhöfer, D., Maier, W., Schwettmann, L., Weiss, D. J., & Groneberg, D. A. (2020). Access to intensive care in 14 European countries: a spatial analysis of intensive care need and capacity in the light of COVID-19. *Intensive care medicine*, *46*(11), 2026-2034.

Berkowitz, S. A., Seligman, H. K., Meigs, J. B., & Basu, S. (2018). Food insecurity, healthcare utilization, and high cost: a longitudinal cohort study. *The American journal of managed care*, *24*(9), 399.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192-225.

Bivand, R. (2022). R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data Geographical Analysis URL https://doi.org/10.1111/gean.12319.

Cambon, J., Hernangómez, D., Belanger, C., Possenriede, D. (2021). tidygeocoder: An R package for geocoding. *Journal of Open Source Software*, *6*(65), 3544, https://doi.org/10.21105/joss.03544 (R package version 1.0.5)

Duncan, E., Cramb, S., Baade, P., Mengersen, K., Saunders, T., & Aitken, J. (2020). Developing a Cancer Atlas using Bayesian Methods: A Practical Guide for Application and Interpretation. Vol. v1. 0.1.

Garg, A., Homer, C. J., & Dworkin, P. H. (2019). Addressing social determinants of health: challenges and opportunities in a value-based model. *Pediatrics*, *143*(4).

Gelfand, A. E., & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, *4*(1), 11-15.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2).

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Hatef, E., Rouhizadeh, M., Nau, C., Xie, F., Padilla, A., Lyons, L. J., ... & Roblin, D. (2021). A Pilot Study to Improve the Use of Electronic Health Records for Identification of Patients with Social Determinants of Health Challenges: A Collaboration of Johns Hopkins Health System and Kaiser Permanente. *Health Services Research*, *56*, 27-28.

Havard Rue, Sara Martino, and Nicholas Chopin (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion), *Journal of the Royal Statistical Society*. B, *71*, 319-392.

Jerrett, M., Gale, S., & Kontgis, C. (2010). Spatial modeling in environmental and public health research. *International journal of environmental research and public health*, *7*(4), 1302-1329.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., ... & Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC.

McLafferty, S. L. (2003). GIS and health care. *Annual review of public health*, *24*(1), 25-42.

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

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, *4*, 395-421.

Social determinants of health. (n.d.). *World Health Organization*. Retrieved from https://www.who.int/health-topics/social-determinants-of-health.

Thiago G. Martins, Daniel Simpson, Finn Lindgren and Havard Rue (2013). Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, *67*(2013) 68-83.

Thornton, R. L., Glover, C. M., Cené, C. W., Glik, D. C., Henderson, J. A., & Williams, D. R. (2016). Evaluating strategies for reducing health disparities by addressing the social determinants of health. *Health affairs*, *35*(8), 1416-1423.

Walker, K. (2022). _tigris: Load Census TIGER/Line Shapefiles_. R package version 1.6, <https://CRAN.R-project.org/package=tigris>.