SPATIOTEMPORAL MODELING OF OPIOID ABUSE AND DEPENDENCE OUTCOMES USING BAYESIAN HIERARCHICAL METHODS

by

Natalie Sumetsky

BS, BA, University of Pittsburgh, 2007, 2007

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Natalie Sumetsky

It was defended on

July 27th 2017

and approved by

Stewart Anderson, PhD, Professor of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Christina Mair, PhD, Assistant Professor of Behavioral and Community Health Sciences, Graduate School of Public Health, University of Pittsburgh

Jeanine Buchanich, PhD, Research Associate Professor of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Joyce Chang, PhD, Professor of Medicine, Biostatistics, and Clinical and Translational Science, School of Medicine, Institute for Clinical Research Education, and Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Stewart Anderson, PhD, Professor of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Copyright © by Natalie Sumetsky
2017

SPATIOTEMPORAL MODELING OF OPIOID ABUSE AND DEPENDENCE OUTCOMES USING BAYESIAN HIERARCHICAL METHODS

Natalie Sumetsky, MS

University of Pittsburgh, 2017

Abstract Opioid addiction is a major public health concern that presents a significant disease burden. In the past decade, drug overdose rates have soared. More research is necessary to inform policy and to ensure provision of proper care to individuals and communities in need. This thesis explores spatiotemporal models to assess ecological and demographic factors associated with opioid addiction risk on a ZIP-code level in Pennsylvania.

Bayesian hierarchical models are commonly used to explore complex spatiotemporal disease trends. Markov chain Monte Carlo (MCMC) simulations are a valuable albeit computationally costly tool in fitting models of this class. A newer method, integrated nested Laplace approximation (INLA), offers improved computational efficiency with comparable results for models with latent Gaussian fields. For example, a 2014 cross-sectional model discussed in this thesis took 5581 seconds to run using MCMC simulations, while INLA offered comparable results in seven seconds. Cross-sectional and longitudinal misalignment models with opioid abuse and dependence outcomes are compared using both methods.

Higher outcome risk is associated with areas with greater proportions of 45- to 64-year-olds, higher density, more retail clutter and manual labor establishments per square mile, higher unemployment, lower median income, and greater proportion of residents below the 150%poverty line. As regional needs differ, identifying high-risk community-level factors and locations carries great public health significance. Interventions and preventive efforts could then be tailored specifically to areas where the disease burden is greatest.

Keywords: Bayesian hierarchical models, spatiotemporal models, conditional autoregressive models, Markov chain Monte Carlo simulation, integrated nested Laplace approximation, opioid abuse, opioid dependence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

According to a recent Centers for Disease Control and Prevention (CDC) report [1], opioid-related death rates in Pennsylvania have been steadily climbing the charts. Among all U.S. states, Pennsylvania ranked ninth in 2013 (19.4 deaths per 100,000 population), eighth in 2014 (21.9 per 100,000 population), and sixth in 2015 (26.3 per 100,000 population). Most of the higher-ranking states were in neighboring Appalachian regions; in 2015, West Virginia topped the list with 41.5 deaths per 100,000 population [1]. This is not merely an individual-level problem—increasing drug-related crime and health care burden can devastate communities and present a major public health concern.

Nationwide, over 27 million people have reported illicit or prescription drug misuse in 2015 [2]. Impact of opioid abuse and dependence has been staggering, with 60.9% of overdose deaths in 2014 involving opioids, while overall drug overdose rates tripled from 1999 to 2014 [1, 2]. This has resulted in drug overdose becoming the primary cause of accidental death in the United States.

The latest Surgeon General's report [2] states that systems-level approaches are effective, but more research on opioid misuse is needed to inform policy and prevention program focus [3]. In a time of policy transitions, scientific backing is particularly valuable in ensuring sustained provision and expansion of adequate health care related to substance abuse. Communities have different risks and needs. Studying ecological factors associated with opioid abuse and dependence is crucial for developing tailored solutions.

This thesis explores spatiotemporal patterns associated with opioid abuse and dependence risk across Pennsylvania during the years 2004 to 2014 on a ZIP-code level. Several methods are used to tackle these models, which tend to be quite complex. Particularly, space and time, whose omission can introduce biases and lead to misleading interpretations, offer

valuable information. The effect of spatial proximity must be considered since spatial units cannot be treated independently—neighboring regions tend to share certain attributes. As in space, units closer in time are often more similar than those further apart.

In-depth analyses of small spatial units are needed to better understand the spread of risk patterns over space and time. However, such subdivisions often create small-population regions that are difficult to model when outcomes are rare. Most approaches cannot adequately capture small-area counts of rare diseases, and unweighted incidence in regions with minute populations can tremendously distort results.

Traditional frequentist models are not adept at handling many of the complexities of spatiotemporal data. First, it is difficult to account for spatial dependence of neighboring regions. Small-area issues, too, present obstacles. Furthermore, confounding factors and latent effects are easily missed in such models—it is difficult to account for all contributing variables to complex outcomes, and many characteristics of interest are unavailable. Section 2.2 addresses modeling issues in greater detail.

Bayesian hierarchical models have been widely used to address issues in spatiotemporal modeling. Besag-York-Mollié (BYM) models with conditional autoregressive (CAR) random effects are particularly popular in addressing spatial autocorrelation problems. This thesis largely focuses on such models and their application to spatiotemporal modeling of opioid abuse and dependence data.

To give context to the models and methods subsequently discussed, section 2.3 introduces some terminology, including latent and Gaussian Markov random fields (GMRFs), exchangeability, and stationarity. Bayesian hierarchical models that address some of the aforementioned modeling complexities are explained in section 2.4. BYM models are discussed in conjunction with conditional autoregressive (CAR) approaches in section 2.5. In section 2.6, stochastic partial differential equations (SPDEs) are addressed. SPDE models are typically used for point-level (rather than aggregated) data. Here, they are offered as models of smoothed average trends and compared to BYM models.

Next, simulation and approximation methods are described. Formerly confined to theory, Markov chain Monte Carlo (MCMC) methods (section 2.7) have become feasible using software such as WinBUGS [4]. Growing computational power has made MCMC application

a viable approach; this, in turn, has been met by a parallel growth in extent and availability of data, resulting in a need for even faster computational methods.

A newer approach, integrated nested Laplace approximation (INLA), has demonstrated improved computational efficiency [5]. Though MCMC methods are monumentally valuable, they can still take days to run when model dimensions are large—this is notoriously termed the "big N problem." In section 2.8, INLA methods are discussed.

Section 3 describes the opioid abuse and dependence data used in inspired this thesis. Modeling these data is discussed in sections 3.2 and 3.3. Finally, in section 4, results are presented and compared across models and methods, followed by a discussion and conclusion in section 5.

2.0 BAYESIAN SPATIOTEMPORAL METHODS AND MODELING

Classical statistical models commonly assume independence and identical distribution of random variables. However, incorporating time and spatial components challenges this assumption. Modeling rare outcomes in small areas, spatial autocorrelation, unit misalignment (section 2.2.4), and confounding are some of the obstacles encountered in spatiotemporal modeling. Fortunately, there are methods that tend to these issues. This section describes some basic terminology, several types of Bayesian spatial modeling approaches, and methods to implement these models.

2.1 BAYESIAN BASICS

Bayesian inference is rooted in Bayes' Rule, $p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})}$. $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are the likelihood and prior probabilities, respectively. $p(\boldsymbol{x}|\boldsymbol{\theta})$ denotes how likely $\boldsymbol{x}$ would be observed conditional on $\boldsymbol{\theta}$, the unknown parameter(s) of interest. $\boldsymbol{x}$ is the observed data matrix, and $p(\boldsymbol{x})$ represents the fixed marginal probabilities, unaffected by differing values of $\boldsymbol{\theta}$. Bayesian approaches seek to estimate posterior distributions, $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Prior distributions are specified before computing posterior distributions and reflect preexisting beliefs about the distribution of $\boldsymbol{\theta}$. It is possible to select uninformative priors that do not rely on subjective judgment or previous knowledge; such priors tend to yield posteriors comparable to maximum likelihood estimators.

Though computational advances have generally narrowed the contrast between Bayesian and frequentist approaches, their foundations are fundamentally different. Essentially, frequentists state that true values for parameters exist, and observed data are assessed in terms of confidence about observing such data if the parameters were equal to certain values. Bayesians argue that we cannot know the values of unknown parameters, but we can estimate which values are more likely based on observed data and prior beliefs or observations. In Bayesian inference, true values of parameters (conditional on observed data) are treated as random variables.

Another difference in these two approaches lies in treatment of intervals. Though both ambiguously abbreviated as "CI," frequentist confidence intervals are interpreted quite differently from Bayesian credible intervals. Confidence intervals are used to state that the true value of the parameter is found in $(1 - \alpha)\%$ repetitions of an experiment (or sample draws), where α is the predetermined level of significance. Conversely, Bayesian analysis relies on credible intervals, which give the probability of the parameter falling within the interval given the observed data.

2.2 SPATIOTEMPORAL MODELING ISSUES

2.2.1 Autocorrelation

Spatial autocorrelation refers to the similarity of an object to its neighbors. Widely accepted as the first law of geography, Tobler's statement, "Everything is related to everything else, but near things are more related than distant things" [6], speaks to this. Indeed, spatial dependence is a common phenomenon that complicates statistical inference. Furthermore, in the case of areal [7] data (i.e., data that are aggregated over a region rather than reported point-by-point), spatial boundaries such as ZIP codes discretize spatial fields that are, in reality, continuous. The effect of a variable likely does not end where one ZIP codes ends and another begins—rather, these boundaries are somewhat artificial, and effects have a continuous influence that carry over to nearby regions.

2.2.2 Small-area modeling

Particularly for rare outcomes, regions with low population densities or counts may exhibit misleading risk levels; appropriate weights are required to account for such cases. Otherwise, even a single observation in a small-population area occurring merely due to chance could lead to inference of high risk in this area. There is a degree of borrowing strength [8] among nearby regions that can be used to temper such effects.

2.2.3 Overdispersion

The issue of overdispersion is not particular to spatiotemporal models, but it is discussed here as it does arise. Overdispersion refers to excessive variability given a particular statistical model. For example, this happens in Poisson models when the count of zeros is disproportionately high. In a standard Poisson distribution, the variance cannot be modeled separately from the mean—in fact, they are equal. A common solution is instead to fit a negative binomial model, which allows separate parameters for the variance and mean. Zero-inflated Poisson and hurdle models are other options. In Bayesian hierarchical models, autoregressive spatial effects can be used to successfully account for overdispersion [9, 10].

2.2.4 Misalignment

Another issue at times encountered in spatiotemporal data is misalignment. In point-level (unaggregated) data, misalignment can refer to observing a value at a removed location. For example, an outcome can be observed at different coordinates than a covariate when the outcome moves (e.g., an animal sighting) [11].

In our model, misalignment is discussed as regional shifting through time. Spatial data are frequently available at ZIP-code, rather than at Census-designated, levels. Unfortunately, this complicates spatial analyses as these regions reflect mail delivery convenience rather than research pragmatism. ZIP codes can be deleted, added, shrunk, or augmented to reflect population changes, resulting in spatial misalignment with time [12].

2.3 RANDOM FIELDS

A random field is an extension of a stochastic process to multidimensional space. Formally, for a parameter set $S \in D$, where D is the domain, all values of a random field, $x(\boldsymbol{s})$, are random variables for any $S \in D, D \subseteq \mathbb{R}^d$. An example of a random field algorithm applied to a Pennsylvania (PA) triangulation mesh (for simplicity, the same 2014 mesh is used for all years) for $t = 11$ years is given in Figure 1. The mesh is based on ZIP-code centroid locations. The simulations are associated with an example autoregressive factor of $\rho = 0.7$ and are based on the conditional distribution of a stationary Gaussian latent field (i.e., with a constant mean and a covariance function depending on $||s_i - s_j||$).

In data analysis, the goal is often to find the best balance between amount of data explained and efficiency. If the underlying domain, D, is continuous, observations—as they are finite—are only a partial realization of D. In areal data, observations intrinsically lack precision as such data are, by definition, aggregated or averaged over regions. However, the underlying latent field nonetheless tends to be continuous. In point-level (geostatistical) data, though observations tend to be more precise, it is generally impossible to obtain all values of parameters of interest. This section describes several aspects of random fields and their applications in Bayesian modeling.

2.3.1 Latent Gaussian fields

Latent Gaussian fields are "a subset of all Bayesian additive models with a structured additive predictor" [5]. This predictor (e.g., $\log(\boldsymbol{\lambda})$) is an additive combination of the intercept, observed fixed effects, and "$f(\cdot)$" terms. The $f(\cdot)$ terms either model random effects (e.g., spatial) or introduce non-linear components [5]. Neither the outcome nor the hyperparameters need to be Gaussian—in fact, in our model, they are not. This is further described in Chapter 3.

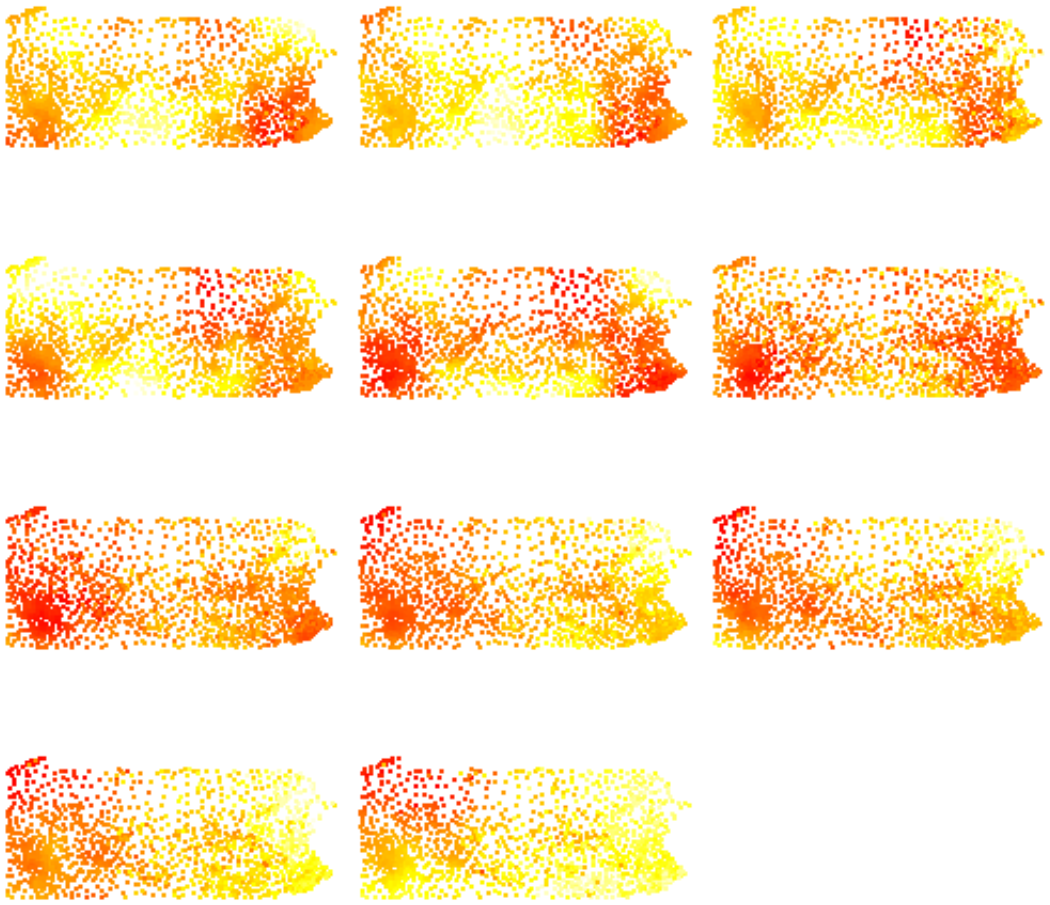

Figure 1: Example of a random field algorithm applied to a 2014 Pennsylvania triangulation mesh over a period of 11 years

2.3.2 Gaussian Markov random fields (GMRFs)

As the name implies, a Gaussian random field (GRF) is a random field whose variable distributions are multivariate normal. GRFs have continuous spatial domains with many convenient analytical properties that are unfortunately inefficient when N is large. Discretizing such fields, when possible, attenuates the costly computation of dense matrices to which GRFs are mapped.

GRFs that satisfy the Markov property are accordingly called Gaussian Markov random fields (GMRFs). The Markov property applies when each subsequent step of a stochastic process depends only on its current state. A Markov field, then, is a multidimensional version of the Markov property—each observation depends only on its neighbors. Conveniently, this property enables such fields to be mapped to sparse matrices, whose properties result in better computational efficiency [13]. GMRFs are associated with precision matrices $\boldsymbol{Q}$ in lieu of standard covariance matrices. Precision matrices are sparse and nonzero only for neighbor observations. They are also known as inverse covariance matrices.

2.3.3 Exchangeability

A random field is exchangeable if, for permutation π,

$$\forall (X_i, X_j) \in S,\ \forall \pi \in S_n, X_{\pi(i)}, X_{\pi(i+1)}, ..., X_{\pi(n)} \stackrel{d}{=} X_i, X_{i+1}, ..., X_n.$$

Exchangeability is essentially a conditional "independently and identically distributed" (iid) assumption. In contrast to spatially autocorrelated variables, the locations (or ordering) of exchangeable parameters are assumed to be independent, $\theta_{i,t} \overset{indep}{\sim} N(0, \tau_\theta)$. This essentially states that there is nothing tying X_i to its arbitrary subscript, i. This concept is important in Bayesian hierarchical models since the first layer is often conveniently exchangeable.

2.3.4 Stationarity

Weak stationarity refers to the stability of the mean, variance, and autocorrelation structure of a variable or field over time. That is, $\forall i \in \mathbb{N},\ \ E(x_i) = \mu$, and the autocorrelation is only a function of the lag between indices i and $i+k$, $\forall i, k \in \mathbb{N}$, but not of the indices themselves. A field $x(\boldsymbol{s})$ with this property is said to be second-order stationary [12].

2.4 BAYESIAN HIERARCHICAL MODELING

Bayesian hierarchical modeling involves nested parameters. That is, lower-level effects of hierarchical data are nested within (and are thereby dependent on) higher-level effects. We are typically interested in estimating the marginal posterior distribution of the first level of the hierarchy [12, 14]. Lower-level parameters are referred to as hyperparameters, and their priors are termed hyperpriors. A more detailed explanation and example is given in chapter 3.

2.4.1 Besag-York-Mollié (BYM) models

Besag-York-Mollié (BYM) [15] models are commonly used to model spatiotemporal disease data. The components of these models include a fixed intercept, a structured spatial autocorrelation effect, and an unstructured random noise effect. Fixed covariates are optionally included as well. These models include conditional autoregressive (CAR; see section 2.4.1.2) effects, which rely on neighborhood structures, to account for spatial autocorrelation and overdispersion issues.

2.4.1.1 Neighborhood structures and adjacency matrices

Developed by Esri [16], a shapefile geospatial vector format is a combination of files that store certain spatial information such as location coordinates and other attributes. Such files have many applications in spatial analyses. By manipulating neighborhood structure

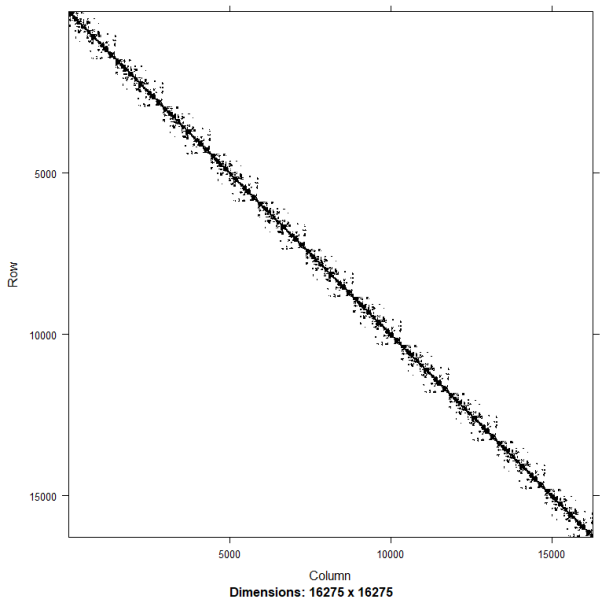

Figure 2: A sparse block matrix of 11 adjacency matrices for PA from year 2004 to 2014

information, R packages such as maptools and R-INLA [5, 13] can convert these files to neighborhood adjacency matrices through a series of simple commands.

There are several ways of specifying a neighborhood structure, and sparse adjacency matrices are among the most convenient. Figure 5 shows such a block matrix $\boldsymbol{R}$ that combines 11 yearly neighborhood structures. As ZIP codes from distinct years cannot be neighbors, all values $R_{ij} \neq 0$ are along the diagonal, and most values R_{ij} are, in fact, 0, making this a sparse matrix.

Neighborhood systems can be represented as $N_{s_i} : s_i = s_1, ..., s_n$, where s_n is the number of neighborhoods, and $s_i = s_1, ..., s_n$ are sites in these neighborhoods. N_{s_i} corresponds to the list of neighbors of s_i. If s_i and s_j are such sites, $s_j \in N_{s_i}$ iff $s_i \in N_{s_j}$ and $s_i \notin N_{s_i}$. That is, if s_i is a neighbor of s_j, s_j is likewise a neighbor of s_i, and no site is a neighbor to itself. As previously described, this is called the Markov property. For precision matrix $\boldsymbol{Q}$, as $Q_{ij} \neq 0$ iff $s_j \in N_{s_i}$, this neighborhood structure essentially corresponds to a GMRF.

2.4.1.2 Conditional autoregressive (CAR) effects

We assign the random effect ϕ_i the following "CAR Normal" distribution [11, 15, 17, 18]:

$$\phi_i|\boldsymbol{\phi}_{-i} \sim N\Big(\frac{1}{N_i}\sum_{j=1}^{n} a_{ij}\phi_j,\ \tau_i^{\phi}\Big),$$

where ϕ_i is the spatial autocorrelation associated with each s_i, $\phi_i|\boldsymbol{\phi}_{-i}$ are all sites except s_i, the mean is the neighborhood structure, a_{ij} is the neighbor indicator, and τ_i^{ϕ} is the associated prior distribution. It is generally assigned an uninformative distribution (e.g., $log(\tau_i^{\phi}) \sim logGamma(1, 0.001)$).

Due to misalignment, BYM models cannot simply be applied to the same region over different time periods. One solution to this is to stack years, whereby yearly adjacency matrices are combined into a single block-diagonal matrix [19]. The outcome matrix $Y_{i,t}$ is then transformed to vector $Y_{i+J[t]}$, where i is a unique ZIP code ID (consistent over time), and t is the year [19]. $J[t]$ is the cumulative count of non-unique ZIP codes for $t-1$ years. For example, if year 1 has 1469 ZIP codes, $J[2] = 1469$, and ZIP code number 340 is $Y_{340,2} = Y_{340+1469} = Y_{1809}$.

Temporal autocorrelation issues arise as well since observations are measured at the same space during different time points. However, as ZIP codes change with time, temporal autocorrelation is difficult to monitor. Instead, spatial variation is measured separately for each year while the variance is fixed to a common value across years [19].

2.5 STOCHASTIC PARTIAL DIFFERENTIATION EQUATION (SPDE) MODELS

SPDE models are a computationally efficient way to model point-level data with an underlying continuous domain. These models begin with the triangulation of a spatial domain. That is, a triangular mesh (e.g., Figure 4) is created as the basis for an SPDE/GMRF representation. The model is then built on the vertices of the mesh. Subsequently, GFs of the Matérn class can be linked to GMRFs by way of SPDEs. The mesh specification is crucial and must be a careful balance between desired precision of the results and computational

efficiency. The continuous domain is discretized by creating a weighted sum of finitely many basis functions. As previously discussed, GMRFs, as sparse matrices, have many convenient properties in linear algebra that increase computational efficiency.

Reviewing the notation, $x(\boldsymbol{s})$ is a latent spatial field, where $\boldsymbol{s}$ is a vector of the locations, $x(\boldsymbol{s}) \in D, D \in \mathbb{R}^d$. D is the d-dimensional Euclidean domain of the data [13] (e.g., the state of PA).

Matérn fields are actually stationary solutions to SPDEs of the form $(\kappa^2 - \Delta)^{\alpha/2}(\tau x(\boldsymbol{s}))$. Here, κ affects spatial scaling, Δ is the Laplace operator (see [20] for an in-depth discussion), α is the smoothing parameter (larger values result in precision matrix values over larger neighborhoods), and τ is the variance [13, 21]. When α is an integer, such a field is also Markov. Upon obtaining the Matérn solution to the SPDE, a GMRF can be constructed,

$$x(\boldsymbol{s}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow x(\boldsymbol{s}) \sim N(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}),$$

where $\boldsymbol{Q}$ is the precision matrix [13]. Due to favorable computational properties, precision matrices are frequently used in Bayesian analysis in lieu of standard covariance matrices.

The triangulated mesh is then linked to observation locations (an example is shown in Figure 4) using the observation matrix $\boldsymbol{A}$, while the covariates are linked to the identity matrix. $\boldsymbol{A}$ is applied to the spatial effects and linked to the log-expectation of the observations, $\eta(\boldsymbol{s})$, while the covariates, already assessed at the locations, are linked directly;

$$\eta(\boldsymbol{s}) \sim \boldsymbol{A} \cdot x(\boldsymbol{s}) + \boldsymbol{A} \ \cdot \ \text{Intercept} \ + \ \text{covariates}.$$

Our data are areal with a finite number of irregular regions with well-defined boundaries; in contrast to point-level data, areal observations are aggregated counts or rates. However, the underlying latent field of our data is conceptually continuous. Applying the SPDE approach to our data, we are creating a model of spatially-smoothed average trends. We treat the sums and averages by ZIP code as point-level central locations (centroids). The fit of this model to our data and a comparison to the BYM model is discussed in chapters 3 and 4.

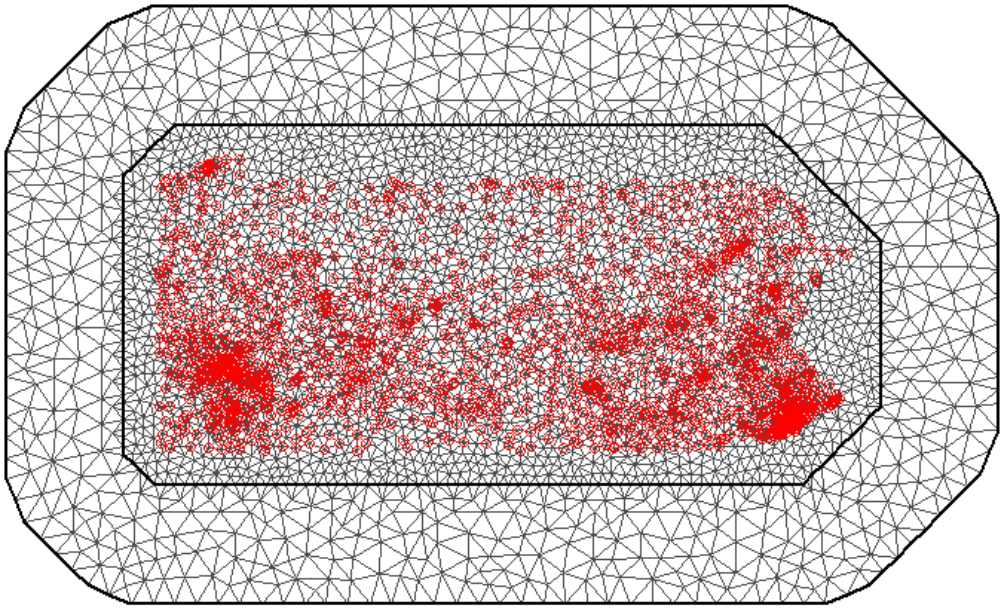

Figure 3: Pennsylvania triangulation mesh with centroids of 1490 ZIP codes for year 2014

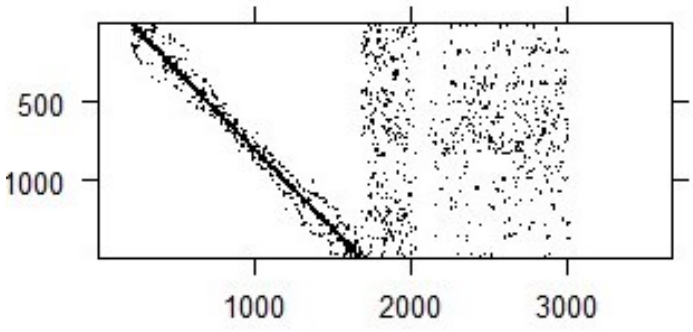

Figure 4: Example of an A matrix for the 2014 Pennsylvania triangulated mesh in Figure 3

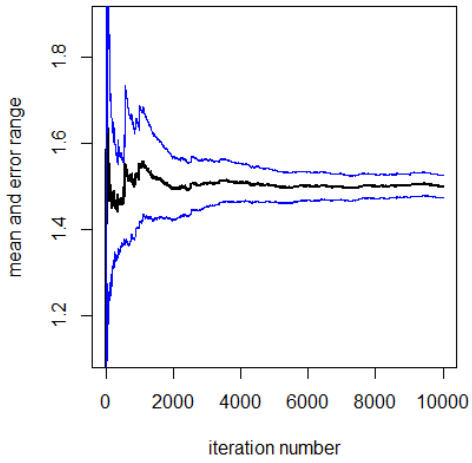

Figure 5: Simulation of the golden ratio value using Monte Carlo methods

2.6 MARKOV CHAIN MONTE CARLO (MCMC) METHODS

When the prior is not a conjugate form of the likelihood distribution, a closed form solution for the posterior distribution is typically unavailable. Historically, open-form integrals precluded using Bayesian hierarchical techniques for complex spatiotemporal modeling. With increased computational efficiency, the use of Markov chain Monte Carlo (MCMC) simulation methods has become more practical, allowing for the assessment of more complex models.

Monte Carlo (MC) simulation methods entail repeated random sampling from a given distribution; this is useful in solving complex integration problems. By the Law of Large Numbers, the limit of the random samples should approach the desired posterior distribution (in the case of Bayesian inference). A simple example for estimating the value of the golden ratio with the integral

$$\int_0^{20} \frac{x^{\pi/5-1}}{1+x^{2\pi}}\,dx$$

by using MC simulation methods is presented in Figure 5.

However, as our data are spatially autocorrelated and not independent, we need further constraints to achieve accurate approximation. Markov chains are random processes that are useful in this regard. Guided by the Markov property, such chains are memoryless—that is, each successive distribution of a random variable depends only on its current state. An example of such a chain is a "random walk," where successive movements are made in any viable direction with equal probability. Some chains, when run for enough iterations, converge to a distribution that is fairly stable. When MC simulations of this type are carried out, Markov chains converge to the joint posterior distribution, if it exists.

There are several types of MCMC algorithms, including the Gibbs sampler, which is the approach used in our model. This algorithm is more amenable to high-dimensional data because it samples from conditional densities rather than the often much larger joint densities. Bayesian inference Using Gibbs Sampling (BUGS) is a project that began in 1989 and has since made significant advances in MC simulations. Several programs, such as WinBUGS [4] (used for our analyses), use the BUGS language to carry out such computations.

Gibbs sampling entails sampling from conditional distributions of each parameter while keeping the other parameters fixed. This approach is iterative—a realization from the approximating distribution is drawn at each turn and is accordingly improved for each successive step. With sufficient iterations, the approximating distribution hopefully converges to that of the posterior.

A more mathematical explanation of Gibbs sampling is as follows. Each of the $j = 1, ..., k$ parameters is set to an initial value (it can be chosen from the prior distribution of the respective parameter), $x_j = x_j^{(0)}$. Subsequently, the Gibbs sampler process continues as follows for each iteration i (for k=3 parameters) [12, 22]:

$$x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}),$$

$$x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}),$$

$$x_3^{(i)} \sim p(X_3 = x_3 | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}).$$

As the sampler iterates through parameter distributions, the values are updated accordingly. So, after $x_1^{(i)}$ is sampled based on $i - 1$, $x_2^{(i)}$ runs with updated $X_1 = x_1^{(i)}$ values.

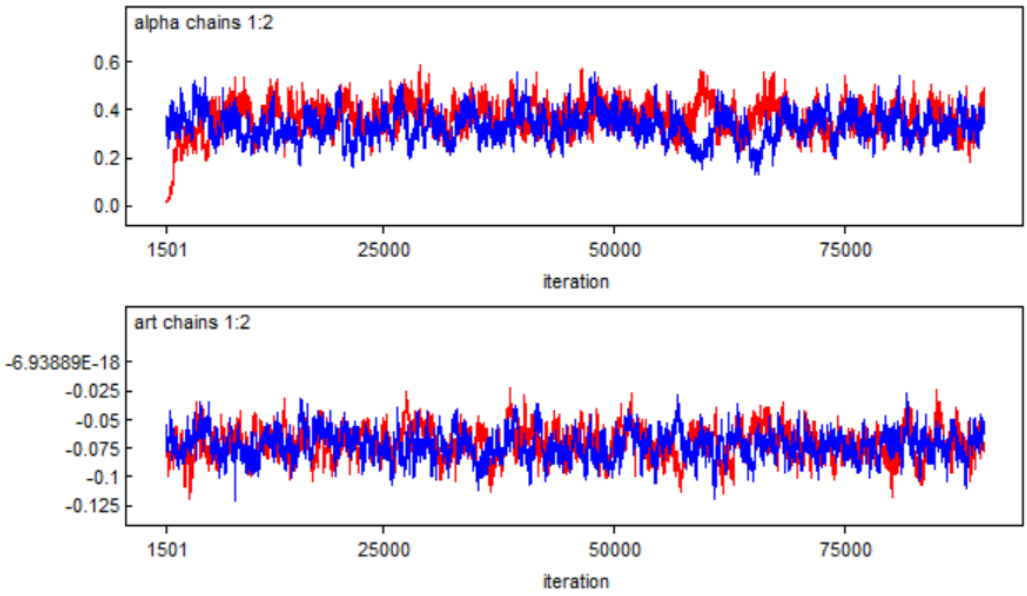

Figure 6: Example of an MCMC trace plot that has converged

Early iterations often have wildly oscillating and unrealistic values and are usually discarded. A certain "burn-in" period (e.g., 50,000 iterations is typical for complex models) is allotted. The process continues until convergence, which happens for properly-defined models when an invariant distribution that mimics the true joint posterior distribution is achieved—this occurs when the distribution remains (weakly) stationary. Weak stationarity (previously described) refers to the stability of the first two moments of the random field so that they no longer change with time (or iterations).

An example of a trace plot that has converged is shown in Figure 3. Here, two chains (in red and blue) are shown for two different parameters, alpha (proportion of spatial autocorrelation relative to all noise) and art (arthritis). Iterations 1501 to 90,000 are shown on this plot (the first 1500 are omitted due to high instability).

2.7 INTEGRATED NESTED LAPLACE APPROXIMATION (INLA) METHODS

Integrated nested Laplace approximation (INLA) is a fairly new deterministic algorithm used for tackling complex models [5]. INLA evokes an old technique (i.e., Laplace methods dating back to the 18th century) for solving complex integration problems. The Laplace method involves using Taylor series expansions to approximate integrals. For example, $Gamma(a, b)$ is approximated by $N(\frac{a-1}{b}, \frac{a-1}{b^2})$ using this method [11].

The INLA algorithm offers significant improvement in computational efficiency over MCMC simulation methods. In fact, a cross-sectional model of our data for the year 2014 took over 93 minutes to simulate using WinBUGS (Table 2), while using INLA yielded nearly identical results in seven seconds. Both approaches used the BYM model [15]. For an additional comparison (Table 4), an SPDE model was fit to the cross-sectional data as well.

3.0 APPLICATION OF SPATIOTEMPORAL METHODS TO OPIOID DATA

This chapter describes the data sources and variables included in five models that examine associations between opioid abuse/dependence and relevant ecological conditions. Three of the models are cross-sectional (year 2014), and two are longitudinal (years 2004 - 2014) misalignment models. Four are BYM models, and one is an SPDE model. Results using MCMC and INLA methods are reported in the next chapter.

3.1 DATA SOURCES AND VARIABLE DESCRIPTIONS

A total of 16,275 space-time units (11 years; ZIP code range per year: 1469 to 1490) were used in our analyses (Table 1). Pennsylvania Health Care Cost Containment Council (PHC4) [24] data from the year 2004 to 2014 were used to obtain case diagnoses from over 18 million observations by ZIP code. As approximately 5% of the original PHC4 data fell outside our domain of interest (i.e., non-PA ZIP codes), these observations were dropped. Diagnoses were all recorded using the ICD-9-CM [23] coding system.

Opioid abuse or dependence was the primary outcome for these analyses. The term opioid has become a catch-all term for substances that interact with opioid receptors in the brain. This includes opiates (an older classification of natural derivatives of the opium poppy), synthetic and semi-synthetic prescription painkillers (e.g., oxycodone, fentanyl, morphine), as well as street drugs, such as heroin. In our models, ICD-9-CM [23] codes 304.00-03, 304.70-73, and 305.50-53 were used to aggregate counts of opioid abuse or dependence diagnoses. Counts of opioid abuse and dependence diagnoses were aggregated by ZIP code and used

as the outcome measure. Other diagnoses (e.g., cancer, arthritis) were sourced from PHC4 data as well to represent medical need for prescription opioids.

To represent economic conditions and demographic and environmental covariates, estimates (age, median income, percent of population below the 150% poverty level, percent male, ethnicity/race) were obtained from GeoLytics [25] demographic data. In 2014, poverty levels for the 48 contiguous states were defined as incomes of $11,670, $15,730, and $19,790 for 1-, 2-, and 3-person households, respectively, with an increase of about $4,000 for each additional person [26]. Finally, North American Industry Classification System (NAICS) [27] data were used to determine retail clutter and manual labor industry counts. These were incorporated into our model as rates per square mile.

The GeoDa software [28] was used to obtain adjacency matrices for 2004-2014 shapefiles [16]. Stata 14 [29] was used to aggregate counts and averages of hospitalization data. WinBUGS [4, 22] and R [30] were employed for statistical analyses of the data. For INLA BYM and SPDE analyses, the R-INLA package [5, 13] was used.

3.2 MODEL FITTING

It is generally accepted that a model with an outcome of a rare disease count follows a Poisson distribution with a log link. Our model is exchangeable:

$$y_{i,t}|\lambda_{i,t} \overset{indep}{\sim} Poisson(\lambda_{i,t}).$$

$y_{i,t}$ is the unique opioid abuse or dependence diagnosis count for ZIP codes $i_t, ..., n_t$, and years $t = 1, ..., 11$, where n_t is the count of ZIP codes (range of 1469-1490) in PA for the given year, from 2004 to 2014. The total number of space-time units is 16275.

The longitudinal BYM model for our data is specified as:

$$log(\lambda_{i,t}) = log(\rho_{i,t}E_{i,t}),$$

$$\text{where } \log(\rho_{i,t}) = \alpha_t + \nu + X_{i,t}\boldsymbol{\beta} + \phi_{i,t} + \theta_{i,t}$$

is the risk, and $E_{i,t}$(used as an offset) is the expected outcome count of the given year and ZIP code. There are two intercepts: α_t is year-specific, and ν is the overall intercept. $X_{i,t}\boldsymbol{\beta}$ is the fixed effects matrix by ZIP code and time. The fixed effects included in our model are: age distribution (% 0-19, 20-24, 25-44, 45-64), median household income, percent population below the 150% poverty line level, percent male, population density, race and ethnicity proportions (% African American, Hispanic, White), and rates of overall hospitalization, cancer and arthritis hospitalizations, unemployment rate, and number of retail and manual labor establishments per square mile. Population density (population per square mile) was modeled as a categorical variable with the lowest density as the reference group. ZIP codes were split approximately equally among five density groups.

We also included two random effects. $\phi_{i,t} \sim \text{CAR}N(\text{adjacency matrix}_t,\ \tau^{\phi}_{i,t})$ captures spatial autocorrelation and accounts for overdispersion and misalignment based on year-specific adjacency matrices, while $\theta_{i,t} \sim N(\theta, \tau^{\theta}_{i,t})$ is unstructured random noise. The respective priors of the two hyperparameters, $\tau^{\phi}_{i,t}$ and $\tau^{\theta}_{i,t}$, are uninformative and are each assigned a Gamma distribution. $\boldsymbol{\beta}$ is a vector of fixed effects for their respective observed values, $x_{i,t}$, for each of the n observations of $1, ..., t$ time points.

The parameters of $\log(\rho_{i,t})$ are viewed collectively as a latent field, $\boldsymbol{\gamma}$. Conditional independence (exchangeability) is assumed in that each point y_i is connected only to its respective latent element, γ_i. $\boldsymbol{\gamma}$ is assumed to be Gaussian with mean 0 and a sparse precision matrix $\boldsymbol{Q}$, effectively making $\boldsymbol{\gamma}$ a GMRF [11].

For comparison, an SPDE model is fit as well. However, the interpretation of this model is somewhat different as the ZIP-code-level counts and averages are modeled as centroids. The intention is to model smoothed average trends, simultaneously performing a sensitivity analysis.

3.3 RESULTS

Five models are presented in this section. The first is a cross-sectional BYM model for the year 2014 using MCMC methods. This model it compared to its INLA BYM counterpart in Table 2. Next, longitudinal (years 2004-2014) BYM models are compared using MCMC simulations and INLA (Table 3). Table 4 compares two qualitatively different cross-sectional models (BYM and SPDE), both conducted using INLA methods. The INLA BYM model in Table 4 is the same as that presented in Table 2. For all models, the outcome measure is a unique count of opioid abuse or dependence cases (primary or secondary ICD-9-CM codes).

While the MCMC model took 5581 seconds to complete its simulations for the cross-sectional model, the INLA version yielded results in seven seconds. Effects that were well supported (i.e., those not containing 1.000 in the credible interval, shown in Table 4), did not differ between the two approaches. The BYM model is then extended to years 2004 to 2014. The longitudinal MCMC BYM model took 74544 seconds (40,000 iterations after a 50,000-iteration burn-in period, which is standard for such models); the INLA BYM model approximated results in 452 seconds.

In the longitudinal BYM model, greater risk for opioid abuse or dependence was observed in regions with greater proportions of residents ages 45-64, while regions with higher proportions of 20- to 44-year-olds were associated with lower risk. Increasing population density was positively associated with the outcome using both MCMC and INLA methods. A greater percentage of white and Hispanic ethnicity/race was associated with greater risk of opioid abuse/dependence; while the MCMC method found an association between greater proportion of African American residents in areas of greater risk, the INLA method did not—this was the biggest qualitative difference. Using INLA also resulted in a well-supported positive relationship between the unemployment rate and opioid abuse or dependence rate, while MCMC simulations did not. Interestingly, the 2014 INLA BYM and SPDE models (Table 4) both found a well-supported positive association between unemployment rate and the outcome.

Lower median household income, higher percentage of residents living below the 150% line, more manual labor establishments per square mile, and higher overall hospitalization rate had positive associations with greater opioid abuse and dependence hospitalization risk. Conversely, higher overall proportions of arthritis and cancer hospitalizations were negatively associated with outcome risk.

3.4 TABLES

Table 1: Descriptive statistics, ZIP codes in Pennsylvania, 2004-2014 (n=16,275 ZIP codes)

Covariate	Mean	SD	Minimum	Maximum	Mean change, 2014 vs. 2004
Population	8490.78	11258.67	5.00	73131.91	221.08
Population density	1278.49	3111.11	0.10	36466.28	21.50
Age, %					
0-19	24.02	3.71	0.00	50.0	-1.69
20-24	6.60	1.34	0.00	35.5	0.19
25-44	24.81	3.56	0.00	77.4	-1.83
45-64	27.41	3.15	0.43	68.8	0.80
Below poverty level, %	21.33	10.03	0.00	86.65	3.50
Med. HH income	44812.44	16700.06	0.00	184588.10	8591.90
Unemployment rate	7.42	5.24	0.00	100.00	2.30
Race, %					
African American	4.28	11.77	0.00	98.01	-0.07
Hispanic	2.24	5.50	0.00	71.45	0.37
White	89.12	15.27	0.40	100.00	-6.27
Male, %	49.42	2.73	5.26	100.00	0.01
Opioid abuse/ dependence hospitalizations, no.	18.76	45.86	0	898	11.63
Opioid poisoning (any) hospitalizations, no.	2.02	3.90	0	79	1.02
Heroin	0.47	1.37	0	41	0.42
Non-heroin	1.56	2.87	0	38	0.61
Overall hospitalization rate, per 100 people	19.57	114.87	0.00	12969.50	-11.59
Arthritis, rate per 100 hospitalizations	3.08	2.20	0.00	100.00	1.19
Cancer, rate per 100 hospitalizations	7.96	4.01	0.00	100.00	0.77
Manual labor, no. establishments per mi^2	6.52	17.09	0.00	269.74	-4.91
Retail clutter, no. establishments per mi^2	10.42	58.46	0.00	1856.55	-0.65

Table 2: Cross-sectional model (2014). Median relative rates (RRs) and ln(median RR), opioid use or abuse hospitalizations, Bayesian spatial BYM models using MCMC vs INLA methods (n=1,490 ZIP codes)

	MCMC	INLA		MCMC	INLA	
Node	ln med RR	ln med RR	ln RR diff	med RR	med RR	RR diff
Const	-1.931	-2.067	0.136	0.145	0.127	0.018
Age 0-19	0.023	0.025	-0.001	1.024	1.025	-0.001
Age 20-24	-0.056	-0.047	-0.010	0.945	0.955	-0.009
Age 25-44	-0.025	-0.024	-0.001	0.975	0.977	-0.001
Age 45-64	0.058	0.056	0.002	1.059	1.057	0.002
Hosp. rate	0.009	0.009	0.000	1.009	1.009	0.000
Density 1	0.210	0.202	0.009	1.234	1.223	0.011
Density 2	0.425	0.416	0.009	1.529	1.515	0.014
Density 3	0.686	0.678	0.008	1.986	1.970	0.017
Density 4	1.047	1.041	0.006	2.849	2.832	0.017
Retail	-0.002	-0.002	0.000	0.998	0.998	0.000
Manual	0.000	0.000	0.000	1.000	1.000	0.000
Unemp. rate	0.012	0.012	0.000	1.012	1.012	0.000
Hispanic	0.015	0.015	0.000	1.015	1.015	0.000
White	0.011	0.012	-0.001	1.011	1.012	-0.001
Black	0.005	0.006	-0.001	1.005	1.006	-0.001
Arthritis	-0.072	-0.072	0.000	0.931	0.931	0.000
Cancer	-0.018	-0.017	-0.001	0.982	0.983	-0.001
Med. income	-0.151	-0.151	0.000	0.860	0.860	0.000
Male	-0.004	-0.002	-0.002	0.996	0.998	-0.002
Runtime	MCMC BYM: 5581 seconds			INLA BYM: 7 seconds		

Table 3: Relative rates (RRs)[95% credible intervals] and ln(RR), opioid abuse or dependence hospitalizations, Bayesian spatial misalignment models using MCMC vs INLA methods (n=16,275 ZIP codes)

	MCMC Model	INLA Model	
Covariate	RR [95% CI]	RR [95% CI]	|Difference| in median RR
Demographic characteristics			
Age, %			
0-19	0.996 [0.991, 1.000]	0.996 [0.992, 1.001]	<0.001
20-24	0.959 [0.948, 0.970]	0.939 [0.928, 0.950]	0.020
25-44	0.980 [0.977, 0.983]	0.978 [0.973, 0.983]	0.002
45-64	1.017 [1.012, 1.024]	1.008 [1.001, 1.015]	0.009
Below poverty level, %	1.017 [1.015, 1.019]	1.017 [1.015, 1.020]	<0.001
Median HH income	0.841 [0.831, 0.851]	0.850 [0.839, 0.861]	0.009
Unemployment rate	1.002 [0.999, 1.005]	1.002 [1.000, 1.005]	<0.001
Density			
	1.236 [1.181, 1.296]	1.248 [1.192, 1.308]	0.012
	1.458 [1.388, 1.535]	1.473 [1.400, 1.551]	0.015
	1.836 [1.735, 1.939]	1.821 [1.721, 1.926]	0.015
	2.133 [1.993, 2.263]	2.087 [1.954, 2.228]	0.046
Overall hospitalization rate, per 100 people	1.001 [1.001, 1.001]	1.001 [1.001, 1.001]	<0.001
Male, %	0.996 [0.993, 1.000]	0.998 [0.992, 1.003]	0.002
Race/ethnicity, %			
African American	1.013 [1.010, 1.016]	1.002 [0.999, 1.005]	0.011
Hispanic	1.022 [1.018, 1.025]	1.011 [1.008, 1.014]	0.011
White	1.021 [1.019, 1.024]	1.009 [1.007, 1.012]	0.012
Arthritis	0.943 [0.933, 0.953]	0.942 [0.933, 0.952]	0.001
Cancer	0.977 [0.972, 0.982]	0.977 [0.972, 0.982]	<0.001
Manual labor establish. per mi^2	1.001 [1.001, 1.001]	1.001 [1.001, 1.001]	<0.001
Retail establishments per mi^2	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	<0.001
Misalignment effects			
ZIP code instability	1.001 [0.998, 1.005]	1.002 [0.999, 1.006]	0.001
Runtime	74544 seconds	452 seconds	

Table 4: Relative rates (RRs) and 95% credible intervals, opioid use or abuse hospitalizations, Bayesian spatial BYM and SPDE models and differences in median RRs (n=1,490 ZIP codes)

node	median	2.50%	97.50%	median	2.50%	97.50%	med RR diff
Intercept	0.145	0.092	0.214	0.637	0.421	0.958	-0.492
Age 0-19	1.024	1.005	1.035	1.020	1.003	1.038	0.003
Age 20-24	0.945	0.896	1.083	0.998	0.931	1.071	-0.053
Age 25-44	0.975	0.963	0.988	0.974	0.963	0.986	0.001
Age 45-64	1.059	1.036	1.075	1.058	1.037	1.079	0.002
Hosp. rate	1.009	1.007	1.010	1.009	1.007	1.010	0.000
Density 1	1.234	1.084	1.407	1.306	1.149	1.485	-0.072
Density 2	1.529	1.334	1.763	1.646	1.443	1.880	-0.117
Density 3	1.986	1.727	2.331	2.099	1.826	2.415	-0.113
Density 4	2.849	2.445	3.504	3.039	2.605	3.547	-0.190
Retail	0.998	0.994	1.000	0.998	0.995	1.002	0.000
Manual	1.000	0.999	1.002	1.000	0.999	1.002	0.000
Unemp. rate	1.012	1.005	1.020	1.012	1.004	1.019	0.001
Hispanic	1.015	1.008	1.025	1.007	0.997	1.016	0.009
White	1.011	1.006	1.020	1.004	0.996	1.012	0.007
Black	1.005	1.000	1.014	0.998	0.990	1.005	0.008
Arthritis	0.931	0.907	0.955	0.928	0.905	0.952	0.003
Cancer	0.982	0.970	0.994	0.976	0.963	0.989	0.006
Med. income	0.860	0.835	0.884	0.837	0.818	0.858	0.023
Male	0.996	0.982	1.013	0.997	0.981	1.014	-0.001
Runtime	INLA BYM: 7 seconds			INLA SPDE: 31 seconds			

4.0 DISCUSSION AND CONCLUSION

Much remains to be discovered to understand and abate opioid misuse trends, and the continuation of developments in spatiotemporal modeling can assist in these discoveries. Our analyses indicate important ZIP-code-level associations with opioid addiction risk. To avoid the ecological fallacy, individual-level inferences cannot be drawn based on the models described in this thesis. On the community level, however, these findings highlight important considerations for the implementation of intervention and preventive programs.

Both INLA and MCMC methods are not without fault. MCMC simulations can take a long time to reach convergence. Furthermore, there are no foolproof tests guaranteeing attainment of stationary distributions. It is difficult to achieve perfect convergence, and a careful compromise between MC error and runtime must be negotiated. INLA typically offers better precision but less flexibility than MCMC simulation. However, as a newer method, it has not been tested as extensively. In both methods (and particularly INLA), correct model specification is crucial—slight misspecification can result in reasonable posterior distributions that are nonetheless incorrect.

Further sensitivity analyses need to be conducted. INLA and MCMC methods have distinct strengths and weaknesses; corroborating findings using both methods is advantageous. Aside from BYM models, other approaches could be considered. A future step is to fine-tune the SPDE model and to extrapolate it to the full dataset. Modifications to the BYM models such as those suggested by Leroux [31] and Dean [32] could also be explored.

The longitudinal misalignment INLA model and the cross-sectional SPDE model in tables 2, 3, and 4 need to be assessed more carefully. Due to time constraints, there is room for improvement in these models. If the presented results persist even with further modification, it is of interest to inspect the inconsistencies between the MCMC and INLA analyses.

This thesis is not an exhaustive presentation of our findings. The results chosen to be discussed in here are somewhat simplified to accentuate differences in modeling and methodology. Outside of this work, we have explored opioid overdose as the outcome as well as other covariates. Moreover, we have identified areas with greatest relative risk based on model posteriors, which are good starting points for intervention efforts.

There is much opportunity for investigation of additional outcomes and covariates. Future models could explore effect modification, particularly between population density and other variables. Locations of treatment centers, buprenorphine providers, and pharmacies would be interesting to consider as well. Unfortunately, it is difficult to obtain these locations by year—these data are generally unavailable on the historical level. Assessing spatial lag is another consideration for future models.

Comorbidity of other mental health diagnoses, both as outcomes and covariates, also warrant consideration. Close to half of individuals with substance abuse disorders [2] have comorbid mental health diagnoses. In fact, based on the 2004-2014 PHC4 hospitalization data, approximately 25% of those with opioid abuse or dependence diagnoses had a unipolar depression or dysthymia diagnosis as well; this is more than twice that of the overall PA hospitalization population for this time period. It would be interesting to explore associated trends and population differences in future models.

We would also like to extend our model to years 2015 and 2016 as these hospitalization data have now become available. However, these observations are complicated by an ICD-9-CM to ICD-10-CM transition in the middle of 2015. It is particularly difficult to build a longitudinal model considering this since the change occurred in the middle of the year, while our time divisions are years. One possibility is look at smaller time units, which would also yield greater temporal precision and accentuate potential seasonality trends.

The ultimate goal of these studies is to provide practical information to abet opioid misuse prevention and treatment efforts. Innovative methods yielding improved precision and computational efficiency are a great asset. Further analyses will be beneficial for specific recommendations on the policy and community level, contributing to the public health impact of this work.

APPENDIX

R CODE

```
1
2     # load libraries
3     library(readr); library(INLA); library(lattice)
4     library(gridExtra); library(plyr); library(ggplot2)
5     library(maptools); library(spdep); library(rgdal)
6
7     # set working directory and store its location
8     setwd(".")
9     my.dir <- paste(getwd(), "/", sep="")
10
11
12    ##########################################
13    ### MONTE CARLO GOLDEN RATIO SIMULATION ###
14    ##########################################
15
16    n <- 10000
17    h <- function(x){(x^(pi/5-1))/(1+x^(2*pi))}
18    integrate(h,0,20)
19    x <- h(runif(n))
20    integral <- cumsum(x)/(1:n)
21    # standard error of Monte Carlo integral
22    stderr <- sqrt(cumsum((x-integral)^2))/(1:n)
23    plot(integral, xlab="iteration number", ylab = "mean and error range",
          type="l",
24    lwd=2, ylim = mean(x) + 30*c(-stderr[n], stderr[n]))
25    lines(integral + 2 * stderr, col="blue")
26    lines(integral - 2 * stderr, col="blue")
27    ##########################################
28
29    ###################################
30    ### PA RANDOM FIELDS SIMULATION ###
31    ###################################
32    # modified from: http://www.math.ntnu.no/inla/r-inla.org/tutorials/spde
          /html/
33    # read shape file
```

```
34    shape.loc <- paste(my.dir, "shape files", sep="")
35    pashape <- readOGR(shape.loc, layer="PA_Zips_2014")
36
37    # create mesh
38    coords <- coordinates(pashape)
39    k <- 7
40    mesh <- inla.mesh.2d(
41    coords, max.edge=c(1/k, 2/k), cutoff=0.1/k)
42    plot(mesh, asp=1)
43    points(coords, col='blue')
44
45    k <- 9
46    params <- c(variance=1, kappa=1)
47    set.seed(1)
48    x.k <- rspde(coords, kappa=params[2], variance=params[1], n=k,
49    mesh=mesh, return.attributes=TRUE)
50    dim(x.k)
51
52    rho <- 0.7   # auto-regressive parameter
53
54    x <- x.k
55    for (j in 2:k) {  # correlated sample over time
56      x[,j] <- rho*x[,j-1] + sqrt(1-rho^2)*x.k[,j]
57    }
58
59    # visualization
60    c100 <- rainbow(101)
61    par(mfrow=c(3,3), mar=c(0,0,0,0))
62    for (j in 1:k) {
63      plot(coords, col=c100[round(100*(x[,j]-min(x[,j]))/diff(range(x[,j]))
          )],
64      axes=FALSE, asp=1, pch=19, cex=0.5)
65    }
66    ##########################################
67
68    # load model variables from the file vars.csv in the working directory
69    vars <- read_csv(paste(my.dir, "varsModel.csv", sep = ""))
70    attach(vars)
71
72    # read shape file
73    # readOGR command: layer is the name of the shape file (no extension)
74    shape.loc <- paste(my.dir, "shape files", sep="")
75    pashape <- readOGR(shape.loc, layer="PA_Zips_2014")
76
77    # extract adjacency matrix from shape file
78    temp <- poly2nb(pashape)
79    nb2INLA("PAgraph", temp)
80    PAadj <- paste(getwd(), "/PAgraph", sep="")
81    H <- inla.read.graph(filename="PAgraph")
82    image(inla.graph2matrix(H), xlab="", ylab="")
83    # convert to sparse matrix of class dgTMatrix
84    inla.graph2matrix(H)
85
86  ########################################
```

```
### RUN CROSS-SECTIONAL BYM/CAR MODEL ###
########################################
# f(ID, ...): BYM (Besag-York-Mollie) model with (ID = zipcode)
# http://www.math.ntnu.no/inla/r-inla.org/doc/latent/bym.pdf
# graph option includes PAadj, the neighborhood structure
# y_i ~ Poisson(E_i rho_i)
# log(rho_i) = intercept + phi_i + theta_i
# phi_i is the spatially structured residual
# theta_i is the unstructured residual using exchangeability of the zip
    codes,
# theta_i~ Normal(0,sigma^2_theta)
# scale.model = TRUE scales the generalized variance to equal 1

# define formula
formula <- y ~ 1 + age1 + age2 + age3 + age4 + hosprate + dens1 +
dens2 +  dens3 + dens4 + retail + manual + unemprate + hispanic +
white + black + arth + cancer + medinc + male +
f(ID, model="bym", graph=PAadj, scale.model=TRUE,
adjust.for.con.comp = TRUE,
# hyperpriors theta and phi
hyper = list(prec.unstruct=list(prior="loggamma",  param = c(1, 0.001)),
prec.spatial=list(prior="loggamma", param = c(1, 0.001))))

# create data frame with ZIP codes and observation matrix of covariates
data.pa <- data.frame(NAME=zipcode, y=vars$opioid, E=vars$E, hosprate,
age1, age2, age3, age4, dens1, dens2, dens3, dens4,
retail, manual, unemprate, hispanic, white, black,
arth, cancer, medinc, male)

# re-order the data so it's the same
# in the covariate & adjacency matrices
data.pa.2 <- attr(pashape, "data")
data.pa.2$NAME <- zipcode
order <- match(data.pa2$NAME,data.pa$NAME)
data.pa <- data.pa[order,]
nzip <- length(data.pa[,1]) # determine number of ZIP codes
data.pa$ID <- seq(1, nzip) # create sequential IDs 1, ..., nzip
attr(pashape, "data") <- merge(data.pa2,data.pa,by="NAME")

# run the model
model.PA <- inla(formula14, family = "poisson", data = data.pa,
E = E, control.compute = list(dic = TRUE))

# run another (much faster) version of the mdoel
# (empirical Bayes integration strategy)
model.PA.2 <- inla(formula, family = "poisson",
data = data.pa, E = E, control.compute = list(dic = TRUE),
num.threads = 2, control.inla = list(int.strategy = "eb"),
control.mode = list(restart = TRUE), verbose = TRUE)

# review results
summary(model.PA)
summary(model.PA.2)

```

```
140
141  ##########################
142  # SPDE MODEL USING INLA #
143  ##########################
144  # below is a cross-sectional model
145  # the shape file (i.e., pashape) was previously imported
146  # if it was not, it needs to be imported
147  #   shape.loc <- paste(my.dir, "shape files", sep="")
148  #   pashape <- readOGR(shape.loc, layer="PA_Zips_2014")
149
150  # create coordinates matrix
151  coords <- coordinates(pashape)
152
153  # create mesh
154  k=7
155  mesh <- inla.mesh.2d(
156  coords, max.edge=c(1/k, 2/k), cutoff=0.1/k)
157  plot(mesh, asp=1)
158  points(coords, col='blue')
159
160  # define SPDE
161  spde <- inla.spde2.matern(mesh = mesh, alpha = 2)
162
163  # prepare some INLA stack elements
164  A.est <- inla.spde.make.A(mesh = mesh,loc = coords)
165  A.est2 <- inla.spde.make.A(mesh = mesh,loc = coords)
166  s.index <- inla.spde.make.index(name = "space", n.spde = spde$n.spde)
167  s.index2 <- inla.spde.make.index(name = "space2", n.spde = spde$n.spde)
168
169  # create stack
170  stack.est <- inla.stack(data=list(y=opdepab2), A=list(A.est,A.est2,1),
171  effects=list(c(s.index, intercept=1), s.index2,
172  list(retail=retail, manual=manual, unemprate=unemprate,
173  hosprate=hosprate, age1=age1, age2=age2, age3=age3,
174  age4=age4, dens1=dens1, dens2=dens2, dens3=dens3,
175  dens4=dens4,hisp=hisp, white=white, black=black,
176  arth=arth,cancer=cancer, medinc=medinc, males=males)),
177  tag="est")
178
179  # define formula
180  formula <- y ~ -1 + age1 + age2 + age3 + age4 + hosprate + dens1 +
181  dens2 + dens3 + dens4 + retail + manual + unemprate + hisp + white +
182  black + arth + cancer + medinc + males +
183  # ar1 is autoregressive
184  f(space, model=spde) + f(space2,model="ar1",
185  hyper=list(prec=list(prior="loggamma",
186  param = c(1,0.001))))
187
188  # run the model
189  model.PA.3 <- inla(formula, family = "poisson", E = E,
190  data = inla.stack.data(stack.est, spde = spde),
191  control.predictor = list(A = inla.stack.A(stack.est), compute=TRUE),
192  control.inla = list(int.strategy = "eb"),
193  control.mode = list(restart = TRUE), verbose = TRUE)
```

```
194
195   # review results
196   summary(model.PA.3)
197
198   ### MAPPING FITTED VALUES
199   # may need to scale as outliers dominate map
200   pashape14$RR<-model.PA.2$summary.fitted.values[,1]
201   spplot(pashape14, "RR")
202
203
204   ######################################
205   ### RUN LONGITUDINAL BYM/CAR MODEL ###
206   ######################################
207   # much of this code was adapted from the R-INLA project, r-inla.org
208   varsALL <- read_csv(paste(my.dir, "vars/varsALL.csv", sep = ""))
209   attach(varsALL)
210
211   # REPEAT FOR ALL YEARS:
212   # read shape file and compute adjacency matrix
213   pashape14 <-  readShapePoly(paste(my.dir,"shape files/2014/PA_Zips_2014"
          ,sep=""))
214   # transform to appropriate projection
215   pashape14 <- spTransform(pashape14, CRS("+proj=longlat +datum=WGS84"))
216   # convert to adjacency matrix
217   temp14 <- poly2nb(pashape14)
218   PAadj14 <- paste(getwd(), "/PAgraph14", sep="")
219   H14 <- inla.read.graph(filename="PAgraph14")   ## adj matrix
220   image(H14) # (optional) view adjacency matrix
221   g14 <- inla.graph2matrix(H14)
222   # combine yearly adjacency matrices into block-matrix
223   PAadj <- bdiag(g14, ...)
224
225   # define formula
226   formulaALL <- y ~ 1 + age1 + age2 + age3 + age4 + hosprate + dens1 +
          dens2 +
227   dens3 + dens4 + retail + manual + unemprate + hispanic + white + black +
228   arth + cancer + medinc + male + bl150 + factor(year) +
229   f(ID, model = "bym", graph = PAadj, scale.model = TRUE, adjust.for.con.
          comp = TRUE,
230   # hyperpriors theta and phi
231   hyper = list(prec.unstruct = list(prior = "loggamma",  param = c(1,
          0.001)),
232   prec.spatial = list(prior = "loggamma", param = c(1, 0.001))))
233
234   # create data frame with ZIP codes and observation matrix of covariates
235   data.pa.ALL <- data.frame(NAME = zipcode, y = y, E = E, hosprate, bl150,
236   age1, age2, age3, age4, dens1, dens2, dens3, dens4, year,
237   retail, manual, unemprate, hispanic, white, black,
238   arth, cancer, medinc, male)
239
240   # re-order the data so it's the same in the covariate & adjacency
          matrices
241   data.pa.ALL.2 <- attr(pashape, "data")
242   data.pa.ALL.2$NAME <- zipcode
```

```
243  order <- match(data.pa.2.ALL$NAME,data.pa.ALL$NAME)
244  data.pa.ALL <- data.pa.ALL[order, ]
245  nzip <- length(data.pa.ALL[ ,1]) # determine number of ZIP codes
246  data.pa.ALL$ID <- seq(1, nzip) # create sequential IDs 1, ..., nzip
247  attr(pashape, "data") <- merge(data.pa.ALL.2, data.pa.ALL, by = "NAME")
248
249  # run the model
250  model.PA.ALL <- inla(formulaALL, family = "poisson", data = data.pa.ALL,
         E = E,
251  control.compute = list(dic = TRUE),
252  control.mode = list(restart = T), verbose = T)
253
254  # review results
255  summary(model.PA.ALL)
```

BIBLIOGRAPHY

[1] Rudd RA, Seth P, David F, Scholl L. Increases in Drug and Opioid-Involved Overdose Deaths United States, 20102015. *Morbidity and Mortality Weekly Report*. 65:14451452, 2016.

[2] U.S. Department of Health and Human Services (HHS), Office of the Surgeon General, Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health. Washington, DC: HHS, 2016.

[3] Strang J, Babor T, Caulkins J, Fischer B, Foxcroft D, Humphreys K. Drug policy and the public good: evidence for effective interventions. *The Lancet*. 9810(379):71-83, 2012.

[4] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Journal of Statistics and Computing*. 10(4):325-337, 2000.

[5] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society*, 71(2):319-392, 2009.

[6] Tobler WR. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(Supplement): 234-240, 1970.

[7] Cressie NC. Statistics for Spatial Data, in Statistics for Spatial Data. John Wiley & Sons, Inc., Hoboken, NJ, 1993.

[8] Tukey JW. 1915-2000. The Collected Works of John W. Tukey. Belmont, Calif.: Wadsworth Advanced Books & Software, 1994.

[9] Mohebbi M, Wolfe R, Forbes A. Disease Mapping and Regression with Count Data in the Presence of Overdispersion and Spatial Autocorrelation: A Bayesian Model Averaging Approach. *International Journal of Environmental Research and Public Health*. 11(1):883-902, 2014.

[10] Congdon PD. Applied Bayesian Hierarchical Models. Chapman and Hall/CRC, 2010.

[11] Blangiardo M and Cameletti M. Spatial and Spatio-Temporal Bayesian Models with R-INLA. John Wiley & Sons, Inc., Chichester, UK, 2015.

[12] Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. John Wiley & Sons, Inc., Hoboken, NJ, 2004.

[13] Lindgren F, Rue H, and Lindstrom J. An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society*, 73(4):423-498, 2011.

[14] Carlin BP and Louis TA. Bayesian methods for data analysis. Boca Raton: CRC Press, 2009.

[15] Besag J, York J, Mollie A. Bayesian image restorarion, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1991.

[16] ESRI (2004-2014). Community Sourcebook America with ArcReader ZIP code County [cd] 2004-2014 edition.

[17] Banerjee S, Carlin BP, Gelfand AE. Hierarchical Modeling and Analysis for Spatial Data. Chapman and Hall/CRC, 2014.

[18] Cressie NC, Chan NH. Spatioal modeling of regional variables. *Journal of the American Statistical Association*, 84(406):393-401, 1989.

[19] Zhu L, Waller LA, Ma J. Spatial-temporal disease mapping of illicit drug abuse or dependence in the presence of misaligned ZIP codes. *GeoJournal*, 78(3):463-474, 2013.

[20] Arfken GB, Weber HJ, and Harris FE. *Mathematical Methods for Physicists*. Orlando, FL: Academic Press, 1985.

[21] Bolin D, Lindgren F. Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*. 5(1):523-550, 2011.

[22] Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS Book. Boca Raton: CRC Press, 2013.

[23] Medicode (Firm). *ICD-9-CM: International Classification of Diseases, 9th Revision, Clinical Modification*. Salt Lake City, Utah: Medicode, 1996.

[24] Pennsylvania Health Care Cost Containment Council (PHC4). PHC4 Inpatient Discharge Dataset, 2004-2014. Harrisburg, PA, 2014.

[25] GeoLytics Inc. Estimates Premium 2004-2014. East Brunswick, NJ, 2004-2014. [DVD]

[26] Sebelius K. Annual Update of the HHS Poverty Guidelines. Department of Health and Human Services, 2014.

[27] North American Industry Classification System (NAICS) (online). U.S. Census Bureau. County Business Patterns: 2004-2014, 2004-2014. Retrieved

from https://www.census.gov/data/datasets/2004/econ/cbp/2004-cbp.html to https://www.census.gov/data/datasets/2014/econ/cbp/2014-cbp.html

[28] Anselin L, Ibnu S, Youngihn K. GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis*. 38(1), 5-22, 2006.

[29] StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP, 2015.

[30] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[31] Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: A new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. Springer, 2000.

[32] Dean C, Ugarte M, Militino A. Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics*. 57(1): 197-202, 2001.