Distribution Agreement

Signature:

Qiang Ling

Date

Bayesian Spatial-Temporal Models for Areal Count Data

By

Qiang Ling

Doctor of Philosophy

Biostatistics

Howard H. Chang, Ph.D.
Advisor

Brent A. Johnson, Ph.D.
Committee Member

Lance A. Waller, Ph.D.
Committee Member

Kevin C. Ward, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Bayesian Spatial-Temporal Models for Areal Count Data

By

Qiang Ling
M.S., University of Massachusetts Amherst, 2002

Advisor: Howard H. Chang, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2014

Abstract

Bayesian Spatial-Temporal Models for Areal Count Data

By

Qiang Ling

Analyses of spatial-temporally correlated areal data arise frequently in public health. In this dissertation, a series of hierarchical models are developed for spatial-temporal count data in the Bayesian framework, with a focus on addressing potential overdispersion and inflated zero counts in the data.

The ability to predict areal count data and quantify the associated prediction uncertainty is valuable for describing population health. We first consider recent developments in Bayesian hierarchical modeling approaches with flexible spatio-temporal interactions, and examine their use in projecting future annual county-level cancer incidence rates, with an application to the Colorado cancer incidence data reported to the National Program of Cancer Registries at the US Centers for Disease Control and Prevention for 1998 to 2007. By examining the 2-year ahead predictive performance of models with different random effect specifications, our results demonstrate the advantages of considering temporal trends in spatial associations when modeling cancer incidence rates.

Overdispersion due to zero-inflation is a common challenge in analyzing count data. To address this issue, we first develop spatial-temporal zero-inflated models, which has two parts: a Poisson count model and a logisic model for predicting excess zeros. We further consider a class of two-part hurdle models. The hurdle models also consist of two components: a binary component modeling the probability of any occurrence and a truncated count component modeling the counts given occurrence. The two components in zero-inflated and hurdle models address, respectively, the abundance of zeros and the skewness of the nonzero counts. Several distributions for the non-zero component are considered, including Poisson, negative binomial, and generalized Poisson. We also evaluate the spatial-temporal dependence between the two model components via multivariate conditionally autoregressive priors, which provide spatial and temporal smoothing.

The zero-inflated and hurdle models are applied to (1) Iowa cancer data reported to the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute during 1998-2007, and (2) emergency department visits data in the Duke University Health System. Results demonstrate that the two component models using negative binomial and generalized Poisson as the base distribution outperform the standard Poisson models.

Bayesian Spatial-Temporal Models for Areal Count Data

By

Qiang Ling
M.S., University of Massachusetts Amherst, 2002

Advisor: Howard H. Chang, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2014

Acknowledgement

Thank you to the faculty, staff, and students for providing me a caring and supportive environment while I studied in the Department of Biostatistics and Bioinformatics. In particular, I am grateful to my advisor, Howard Chang, for his consistent patience, guidance, and support. He has been a valuable mentor and teacher throughout the completion of my degree and has generously offered his advice in areas beyond the realm of this research. I would also like to acknowledge my committee members, Brent Johnson, Lance Waller and Kevin Ward, for their thoughtful comments and constructive feedback. In addition to those individuals directly involved in my research, I would like to thank Mary Abosi for always being there to listen and Melissa Sherrer for her course and career advice throughout my graduate career.

In addition, I would like to take this opportunity to thank my family and friends for their support. Special thanks to my wife, Yan Yuan, for offering me consistent patience and love while supporting and taking my side throughout this long term journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Health-related data are often aggregated over different geographic areas and collected at many time points, exhibiting complex spatio-temporal structure. Examples of spatial areal data include the percentage of a surveyed population with household income below the federal poverty limit for a collection of regions, or a choropleth map of land use classification. These different types of data are like pieces of a jigsaw puzzle, into which researchers have devoted themselves to modeling and interpreting the data, such that government funds can be efficiently allocated to public health programs, diseases can be effectively monitored and diagnosed at early stages, and a healthy population can be maintained.

Motivated by the extensive and diverse databases of areal information available in public health, in this dissertation, I develop Bayesian hierarchical models that utilize recent methodological advances for application in medical and public health research.

In the analysis of areal units, statistical research interests over spatio-temporal data in recent years have been focused on a few areas. A common inferential issue of areal data is to identify and quantify spatial patterns of disease. Health outcomes for areal units near each other tend to be more similar than units further apart. Intuitively, changes in disease rates should be indifferent from how the arbitrary, administrative boundaries were set. Second, there exist extreme values as in any sample data. To obtain a surface of estimated disease rates with better precision, especially for small areas and rare diseases, it may be beneficial to consider smoothing the data. Third, the *modifiable areal unit problem (MAUP)*[Banerjee et al., 2004] causes inferential challenges. It's very common that areal units are modified in geographic boundary sets. For example, U.S. Census tract boundaries often change to accommodate population shifts in those areas. Statisticians are challenged with using data from one set of geographic boundaries to make inference for another[Goovaerts and Xiao, 2011]. Finally, in studies with multiple outcomes, such as several potentially related cancer types, another research topic is to handle the dependence among the multivariate components, while accounting for spatial dependence between areal units.

1.2 Proximity Matrix W

A core concept in any spatial-temporal methodology for areal data is the proximity matrix, which describes how areal units are spatially-related to each other. Given a total of n areal units in a study, the proximity matrix W is defined as a $n \times n$ matrix where the ω_{ij} describes the closeness of areal unit j to areal unit i. The matrix W has diagonal elements $\omega_{ii} = 0$ for $i = 1, 2, \cdots, n$. There are different ways to set up values of ω_{ij}. One intuitive approach is to set ω_{ij} equal to the Euclidean distance between the two centroids of areal unit i and unit j. Another possibility is to use a binary indicator: if unit i and unit j share common boundary, then $\omega_{ij} = 1$; otherwise $\omega_{ij} = 0$. In the two choices above, $\omega_{ij} = \omega_{ji}$ and W represents a symmetric matrix. However, W does not necessarily need to be a symmetric matrix. For example for a given i, we could set $\omega_{ij} = 1$ if j is one of the K nearest neighbors of i. For unit j, $\omega_{ji} \neq 1$ if i is not one of the K nearest neighbors of j.

1.3 Spatial and Temporal Association

Spatial, temporal and spatial-temporal effects are often parameterized as random effects within the framework of generalized linear mixed models (GLMM). Standard linear mixed models often assume that individual contribution to the log-likelihood of different groups/clusters can be summed up due to the assumed independence of random effects. However, the independence assumption is often violated in spatial-temporal data, and statistical research has developed new flexible approaches to address this challenge.

Currently, to model autocorrelation for areal data, one of the most widely used frameworks is conditional independence, which assumes that data from different areal units are independent of each other, conditional on random effects at a higher level of the hierarchy. Often, inclusion of spatial correlation through conditional independence only partially accounts for spatial effects as there are residual correlation effects due to unobserved covariates or due to complete randomness.

Based on Brook's lemma[Gelman et al., 2013],

$$\begin{aligned} p(y_1, y_2, \cdots y_n) &= \frac{p(y_1|y_2, \cdots, y_n)}{p(y_{10}|y_2, \cdots, y_n)} . \frac{p(y_2|y_{10}, y_3 \cdots, y_n)}{p(y_{20}|y_{10}, y_3 \cdots, y_n)} \\ &\cdots \frac{p(y_n|y_{10}, \cdots, y_{n-1,0})}{p(y_{n0}|y_{10}, \cdots, y_{n-1,0})} . p(y_{10}, \cdots, y_{n0}) \end{aligned}$$

where $\mathbf{y} = (y_{10}, \cdots, y_{n0})'$ is a realization of the distribution of $p(y_1, \cdots, y_n)$. Brook's lemma states, given $p(y_1, y_2, \cdots y_n)$, if the full conditional distributions $p(y_i|y_j, j \neq i), i = 1, 2, \cdots n$, are uniquely determined, then the joint distribution exists. In spatial-temporal data, the use of a conditional distribution for each spatial unit forms the basic construct of *Markov random field* (MRF)[Rue and Held, 2004]. It also provides a convenient approach for sampling the joint posterior distribution in Bayesian Markov chain Monte Carlo (MCMC) computation. An explicit form of the joint distribution is not necessary in Bayesian inference and computation. Instead, posterior samples are realized via iterative simulation from the joint distribution. Spatial correlation is accounted for and appears inside a prior distribution instead of in the likelihood itself. For spatial data, we often assume that dependence only exists between areal units close to each other; so for each spatial unit i, the conditional distribution can be simplified as:

$$p(y_i|y_j, j \neq i) = p(y_i|y_j,\ j \in \partial_i) \tag{1.1}$$

where ∂_i denotes the neighborhood of unit i.

1.4 Conditional Autoregressive Prior Distributions

The Conditional Autoregressive (CAR) Distribution has been widely used as a prior in spatial-temporal modeling because of its convenience in computation under the Bayesian hierarchical framework. To account for spatial correlation, the general form of the conditional distribution in (1.1) is usually assumed to follow a Gaussian distribution. The conditional autoregressive distribution is specified as:

$$Y_i|y_j, j \in \partial_i \sim N\left(\sum_j b_{ij}y_j,\, \tau_i^2\right), \quad i = 1, 2, \cdots n.$$

If let $b_{ij} = \omega_{ij}/\omega_{i.}$ and $\tau_i^2 = \dfrac{\tau^2}{\omega_{i.}}$, through Brook's Lemma, the joint distribution is given by:

$$p(y_1, y_2, \cdots, y_n) \propto exp\left\{-\frac{1}{2\tau^2}\boldsymbol{y}'(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{y}\right\},$$

where W is the proximity matrix and $\boldsymbol{D}$ is diagonal with $D_{ii} = \omega_{i.}$. The joint distribution expression above suggests Y follows a joint multivariate normal distribution with mean $\mathbf{0}$ and variance matrix $\boldsymbol{\Sigma} = (\boldsymbol{D} - \boldsymbol{W})^{-1}\tau^2$.

1.5 Bayesian Model Comparison

There are several approaches to access model fit. In this dissertation, we used two methods.

The first one is a discrepancy measure, which is used to check data fit by comparing the expected number under the model to the observed, Here we use the Weighted Mean Squared Error (WMSE) given by:

$$T(y,\theta) = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - E(y_i|\theta)\right)^2 / var(y_i|\theta).$$

to measure the lack-of-fit specifically.

The other comparison, Deviance comparison, is a standard summary statistic to examine model fit. Deviance is denoted as -2 times the log-likelihood: $D(y,\theta) = -2\log p(y|\theta)$ [Gelman et al., 2004], which shows the discrepancy between data and model depends on model parameters θ as well as the data y.

To derive a summary statistic that depends only on y, $\bar{D}$ is defined as:

$$D_{\hat{\theta}}\left(y\right) = D\left(y, \hat{\theta}(y)\right)$$

where $\hat{\theta}$ is a point estimate for θ, for example the mean of the posterior sample.

Another Deviance statistics, the $\hat{D}$, is the average deviance over the posterior simulations of θ_n:

$$\hat{D}_{avg}\left(y\right) = \frac{1}{N}\sum_{n=1}^{N} D(y,\theta^n).$$

The estimated average discrepancy above is a better summary of model error than the discrepancy of the point estimate. The point estimate $\hat{\theta}(y)$ generally results in the model fitting better than it really does. In that sense, $\hat{D}$ is generally larger than $\bar{D}$.

The difference between the posterior mean deviance and the deviance at $\hat{\theta}$

$$pD = \hat{D}_{avg}(y) - D_{\hat{\theta}}(y), \tag{1.2}$$

describes the effect of model fitting and is used as a measure of the effective number of parameters in a Bayesian model. So pD is conveniently used in Bayesian statistics as a measure of the model complexity. For a normal linear model with unconstrained parameters pD is equal to the number of parameters in the model. So pD can be considered as the number of 'unconstrained parameters' in the model, where a parameter counts as 1 if it is estimated with no constraints or prior information; 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution; or an intermediate value if both the data and prior distribution are informative. From another point of view, in (1.2), we see that pD represents the decrease in the deviance or the expected improvement of model fit.

To estimate the expected error when applying the fitted model to out-of-sample (replicate) data y^{rep},

$$D^{pred}_{avg}(y) = E\left[\frac{1}{n}\sum_{i}^{n}(y_i^{rep} - E(y_i^{rep}|y))^2\right] \tag{1.3}$$

Similarly, the expected deviance for replicated data is computed as:

$$D^{pred}_{avg}(y) = E\left[D(y^{rep}, \hat{\theta}(y))\right] \tag{1.4}$$

The estimated predictive deviance has been suggested as a criterion of model fit and has been used in selecting the best fit model. In practice, Deviance Information

Criterion (DIC) is a approximation of the expected predictive deviance (1.4), which can be written as:

$$DIC = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y),$$

which is the criterion we use for Bayesian model selection in this dissertation work. Model with the minimum DIC will have the best short-term predictions. DIC can also be used to compare alternative prior distributions as well, if the priors are independent of the data [Spiegelhalter et al., 2007]. However, when the difference in DIC is small, such as $\Delta DIC < 5$, relaying on DIC only for model selection might be misleading [Spiegelhalter et al., 2007]. WinBUGS provides user the convenience of requesting deviance and DIC values for parameters in WinBUGS distribution syntax. However, user might have to calculate DIC themselves for parameters of other distributions.

1.6 Overall Goals and Organization

The overarching goal of this dissertation is to develop, apply and evaluate Bayesian spatio-temporal hierarchical models for mapping and projecting areal counts. Specifically, our interest is in providing areal incidence and rate estimates, as well as their associated uncertainties for small areas. We consider two motivating examples: U.S. Cancer Registries[NCI, 2010, CDC, 2010] and Emergency Department Visits in the Duke Health System[Neelon et al., 2013]. The Bayesian framework allows flexible spatial and/or temporal correlation structures in cancer incidence data across counties and years. We are particularly interested in the use of Gaussian Markov random fields (GMRFs) to model relative risks across discrete areal units, which has been employed extensively in disease mapping applications. Given the contiguous spatial units that arise naturally from administrative boundaries, the GMRF framework borrows information across spatial units. Similar to small-area estimation problems, spatial smoothing is particularly useful for areas with small at-risk populations where crude cancer rates are associated with high estimation uncertainty. Finally, another advantage of the Bayesian approach is that uncertainty in parameter estimation can be easily propagated in cancer incidence projections via posterior predictive distributions.

In study 1, we are interested in evaluating Bayesian spatio-temporal approaches for short-term **cancer incidence projection**. To serve the needs of epidemiology research and facilitate decision making for improving public health and reducing cancer burden, registries should have the capacity to provide estimates about newly diagnosed cancer cases before the actual cases are reported to National Program of Cancer Registries (NPCR). However, there is a standard two-year delay in reporting cancer cases to central registries after the end of a calendar year. It is a common practice in cancer surveillance to estimate the number of new cases diagnosed at the current year, which provides an up-to-date perspective on the occurrence of different types of cancers in different geographic regions.

In study 2, we will develop **spatio-temporal zero-inflated Poisson (ZIP) mixture models** to account for excess zeros in count data. Excess zeros are ubiquitous in cancer registry data, especially for rare cancer sites at the county level. Also, the data often become sparse when stratified into different demographic categories or a finer spatial/temporal scale. High occurrence of zeros can reach a level such that standard Poisson regression exhibits lack of fit even after adjusting for covariates or introducing random effects in the model.

In study 3, to address the excess zeros and the corresponding overdispersion, we will develop **spatio-temporal hurdle models**. We will compare model fit and prediction between hurdle Poisson models, hurdle generalized Poisson models and hurdle negative binomial models. In addition, we will examine the correlation between two random effect components: the zero component and the positive count component in hurdle models.

This dissertation is organized as follows:

Chapter 2 provides a review of Bayesian spatio-temporal areal models and presents a case study of their application in Colorado lung and bronchus cancer. Model performance is compared and 2-year ahead cancer incidence projections are obtained and compared with the actual reported cancer incidence at the county level.

Chapter 3 begins with a review of current methods to account for zero inflation in count data. We then describe the development of zero-inflated mixture models, including zero-inflated Poisson models, zero-inflated negative binomial models and zero-inflated generalized Poisson models, for modeling cancer incidence data. The methods are implemented in a case study using Iowa lung and bronchus cancer data. Preliminary results assessing the degree of zero-inflation are also given.

Chapter 4 considers spatial-temporal hurdle multivariate conditional autoregressive (CAR) models, including hurdle Poisson models, hurdle generalized Poisson models and hurdle negative binomial models, for large number of zero cases reported in Emergency Department Visit data from Duke Health System. Model fit will be compared.

Finally, future directions appear in Chapter 5.

Chapter 2

STUDY 1: BAYESIAN SPATIAL-TEMPORAL DISEASE MAPPING AND PROJECTION FOR COLORADO LUNG AND BRONCHUS CANCER DATA

2.1 US Cancer Surveillance

Cancer is the second-leading cause of death among Americans. Disease surveillance for public health, defined as "the ongoing systematic collection, analysis, and interpretation of health data"[Thacker and Berkelman, 1988], is essential to the planning, implementation and evaluation of public health practice.

In 1971, as a result of the National Cancer Act, the National Cancer Institute's (NCI) Surveillance Epidemiology and End Results (SEER) program was initialized and debuted as a system of population-based cancer registries in five states and four metropolitan areas. Since 1992, following the passage of the Cancer Registries Amendment Act, the National Program of Cancer Registries (NPCR) at the U.S. Centers for Disease Control and Prevention (CDC) has been funding and supporting cancer registries outside the SEER areas. As a result, additional state-wide cancer registries were established and existing registries began collecting cancer data of increasing quality in a timely manner during the 1990s. In 2002, jointly with SEER, NPCR progressed to become a national cancer surveillance which receives cancer incidence data from 49 states, the District of Columbia (DC) and three U.S. territories (Palau, Puerto Rico and the Virgin Islands), covering 96% of the U.S. Population[Hutton et al., 2001].

The ability to utilize information from population-based surveillance data is critical to the mission of the U.S. central cancer registries. Since 1999, CDC, NCI and the North American Association of Central Cancer Registries (NAACCR) have jointly published the *United States Cancer Statistics*, a complete annual report to the nation on US cancer incidence and cancer mortality[CDC, 2010]. Also, the SEER program publishes annually the *Cancer Statistics Review (CSR)* that reports on the most recent cancer incidence, mortality, and survival statistics. In addition, the American Cancer Society (ACS) publishes the *Cancer Facts & Figures* annually which provides estimates of the contemporary cancer burden every year[Pickle et al., 2007].

In the above annual reports, statistical methods have played an important role in describing the trends in cancer incidence and mortality, identifying disparities by population demographics, estimating public health burden, and projecting new cancer cases and deaths. With an ambition of recording every primary cancer in a timely and accurate manner, cancer central registry data are massive and complex. Different cancer sites also have diverse cancer profiles. For example, in 2008, the highest annual cancer incidence rate was estimated to be 144.8 per 100,000 population at risk for prostate, but ovarian cancer has a much smaller estimated annual rate of 9.2 per

100,000 population at risk[CDC, 2010]. Over the last 40 years, different cancer sites have demonstrated different cancer incidence trends.

2.2 Statistical Challenges and Objectives

Due to changes in cancer incidence and the growing availability of U.S. cancer data, there has been a constant interest in developing statistical methods for estimating timely cancer trends over the past decades[Pickle et al., 2005, 2007]. Earlier methods were developed based on the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program, a comprehensive cancer surveillance database dating from 1975. First, a piecewise joinpoint linear regression was used to model temporal trends and identify temporal changes in incidence data[Kim et al., 2000]. Then an autoregressive quadratic time-trend model was developed to forecast cancer incidence to be diagnosed and cancer mortality to be anticipated[Tiwari et al., 2004, Pickle et al., 2003]. This method is a multistep procedure. It first models the long-term trend of cancer incidence as a function of time and the square of time. An autoregressive model was then fit to the residuals to account for the short-term fluctuation of the incidence trend. Additional approaches to model temporal trends have also been proposed, including state-space models[Tiwari et al., 2004] and semi-parametric regressions[Ghosh and Tiwari, 2007].

To address geographic variability in incidence rates, the projection method has been further improved to a spatial-temporal model implemented in SAS/GLIMMX by ACS [Pickle et al., 2007]. County-level socioeconomic status (SES) and lifestyle profiles, including urban/rural status, household characteristics, income, education, occupation, medical facilities and the percentage distribution of the population by race and ethnicity, are taken from the Area Resource File (Bureau of Health Professions 1999) and Census data. By including spatially-varying SES and lifestyle profiles as additional cancer risk factors, this method provides both state- and county-level estimated cancer incidence for the first time. Residual spatial correlations are accounted for by including a non-parametric median-based smoother using the population as weights[Mungiole et al., 1999].

The above methods, however, have several limitations when applied to the NPCR data. First, to conduct piecewise joinpoint linear regression, it is necessary to have data covering a long time period. For the SEER data, previous methods have utilized cancer incidence data over 27 years [Kim et al., 2000] to successfully capture long time trends, whereas NPCR only began in 1994. Moreover, the joinpoint approach cannot effectively deal with recent changes in cancer rates. Specifically, as new data become available, the joinpoint procedure described in [Kim et al., 2000] will identify the best fitting set of joinpoints over the whole range of data. The sequential test procedure by Zhang [1995] fixes the joinpoint regression parameters and searches for additional joinpoints, which might lead to different parameter solutions for the same data.

A second challenge in cancer incidence estimation is the ability to estimate and project cancer incidence for locations lacking high quality incidence data, for example due to small population size. To address this issue, the autoregressive quadratic time-trend model made an assumption that the ratio of each state's incidence to mortality is the same as that for the combined SEER registries, which might not be justifiable in all cases[Frey et al., 1994]. The spatial-temporal model by Pickle et al. [2007] could provide estimates of cancer incidence rates at the county-level. Nevertheless, the authors have observed that counties with fewer residents may have a higher degree of uncertainty in the estimated number of expected cancer cases. Practice from using those methods in estimating US cancer incidence at the ACS has shown substantial geographic variation in projection performance for many cancer sites. Another challenge in relying on spatial projection using spatially-varying SES and lifestyle covariates is that the temporal resolution and availability of these variables are often limited. Uncertainty quantification for incidence projection is also difficult with the non-parametric spatial smoothing approach[Mungiole et al., 1999].

2.3 Overview

This chapter presents a case study utilizing a NPCR registry from Colorado between the period 1998 to 2007. We focus on Lung and Bronchus cancer which is the most common cause of cancer-related death worldwide, responsible for 1.37 million deaths annually[Organization, 2012]. We consider 4 spatial-temporal models for areal data with different specifications of the spatial and temporal components of baseline risk, and evaluate their performance in 2-year ahead projection. This is motivated by the standard 2-year delay in cancer reporting. We also describe an approach to account for geographical boundary change during the study period.

2.4 Methods

Cancer incidence data were obtained from CDC's NPCR program reported by state health departments or their designees as of January 2009. Primary cancer sites were coded according to the International Classification of Diseases for Oncology (ICD-O) 3^{rd} Edition (for International Classification of Disease codes for these sites, see http://seer.cancer.gov/siterecode/icdo3_d01272003/). Only malignant tumors were included, while in situ and other benign tumors were excluded. Individual level demographic variables were also available, including patient's diagnosis year, race, sex, age and county of residence. Only lung and bronchus cancer cases reported for diagnosis years between 1998 to 2007 from Colorado were analyzed in this study. Since minority race groups account for a very small proportion of the total incidence data for Colorado (0.73%), only white (including both Hispanic and non-Hispanic) and black subpopulations were included in this study. Population data for years 1998 to 2007 were obtained from the CDC National Center of Health Statistics (NCHS).

2.4.1 Bayesian spatio-temporal models for areal data

Given the 10 years of incidence data available, we used the first 8 years (1998-2005) for fitting the spatio-temporal models and the last 2 years (2006, 2007) for evaluating model projections. Let s index a patient's residency county, $s = 1, 2, \cdots, S$ ($S = 64$ for Colorado), and let t index reporting year. We categorized age into 6 groups ($\leq$ 44, 45-54, 55-64, 65-74, 75-84, and $\geq$ 85 years old). We aggregated the data by combinations of race, sex, and age group (24 strata in maximum). About 46% of the combinations of county, reporting year, sex, age group and race had no cases reported, which were treated as 0. For the s^{th} county, we observed O_{stk}, the number of observed cancer cases reported from county s during year t in stratum k. With P_{stk} denoting the corresponding at-risk population size, we assume the following hierarchical Poisson regression model:

$$O_{stk} \sim Poisson\left(P_{stk} e^{\mu_{stk}}\right). \tag{2.1}$$

The log of relative risk, μ_{stk}, is partitioned into the following different terms:

$$\mu_{stk} = \mu + \beta_0(t) + \beta_0(s) + \beta_0(s,t) + \mathbf{X}_{stk}\boldsymbol{\beta}, \tag{2.2}$$

where μ is the overall average log baseline relative risk; $\beta_0(t)$ is the purely temporal component that describes the temporal trend in cancer incidence; $\beta_0(s)$ is the purely spatial component that describes the spatial trend in cancer incidence; $\beta_0(s,t)$ represents space-time interaction or the residual error term; and lastly, $\mathbf{X}_{stk}$ denotes the vector of indicators for a particular age, sex and race group, and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients.

2.4.2 Modeling space-time random effects

We first consider a **static model** where the spatial random effects are constant in time without space-time interaction ($\beta_0(s,t) = 0$). A common approach in disease mapping is to decompose $\beta_0(s)$ into two independent random effects:

$$\beta_0(s) = \theta_s + \psi_s.$$

Here θ_s captures the unstructured heterogeneous random effects with $\theta_s \overset{iid}{\sim} N(0, \sigma_\theta^2)$, and ψ_s captures the spatial dependence. We assume $\boldsymbol{\psi} = (\psi_1, \psi_2, \cdots, \psi_S)$ follows a Gaussian Conditionally Autoregressive (CAR) distribution. The CAR(τ_ψ^2) distribution is often specified via a conditional probability density function with

$$\psi_s | \psi_{s' \neq s} \sim N(\mu_s, \sigma_s^2),$$

where $\mu_s = \frac{\sum_{s' \neq s} w_{ss'} \psi_{s'}}{\sum_{s' \neq s} w_{ss'}}$ and $\sigma_s^2 = \frac{1}{\tau_\psi^2 \sum_{s' \neq s} w_{ss'}}$. The weights $w_{ss'}$ are fixed constants that measure the proximity of counties s and s'. In this study, $w_{ss'}$ takes a value of 1 when county s and s' share boundaries, and 0 otherwise. Therefore the full conditional mean is the average of the spatial neighbors and σ_s^2 controls the degree of spatial dependence. Since ψ_s is unidentifiable jointly, we add the constraint $\sum_{s=1}^{S} \psi_s = 0$ [Xia and Carlin, 1998]. To facilitate projection for function years, we assume a linear temporal trend $\beta_0(t)$ during the last 4 years (2002-2005)[Fay et al., 2006]. Since the spatial trend is constant in time in this model, all years prior to 2002 have a year-specific effect.

We next consider a **nested model** described by Waller et al. [1997], where the static model above is applied to each time point separately. Therefore the spatial random effects and their spatial dependence are allowed to vary across time points. Following equation 2.2, we set $\beta_0(s) = \beta_0(t) = 0$, and the log of the space-time varying relative risk, $\beta_0(s,t)$, is parameterized as

$$\beta_0(s,t) = \theta_{s,t} + \psi_{s,t}.$$

Here $\theta_{s,t}$ and $\psi_{s,t}$ are the unstructured heterogeneity and spatial random effects that vary in time, with

$$\begin{aligned}
&\theta_{s,t} \overset{iid}{\sim} N(0, \sigma_t^2), \\
&\boldsymbol{\psi}_t \equiv (\psi_{1,t}, \psi_{2,t}, \cdots, \psi_{s,t}) \sim CAR(\tau_t^2).
\end{aligned}$$

The above model does not borrow information across time in estimating spatial association because σ_t^2 and τ_t^2 are year-specific. Therefore, we also considered a nested model with identical precisions (σ^2 and τ^2) across the time periods.

The third model we consider allows the spatial random effects to evolve across time. Following the **dynamic spatial model** by Chang et al. [2011] and Banerjee et al. [2004], we assume the following structure for log relative risk effects partitioned in 2.2 as follows. The purely spatial effect $\beta_0(s) = \psi_s$ is modeled jointly by a intrinsic CAR distribution,

$$\boldsymbol{\psi} \equiv (\psi_1, \psi_2, \cdots, \psi_S) \sim CAR(\tau_\psi^2)$$

The spatio-temporal random effects, $\beta_0(s,t)$, have a dynamic structure:

$$\begin{aligned}
&\beta_0(s,t) = \rho\beta_0(s,t-1) + \xi_{s,t} \\
&\boldsymbol{\xi}_t \equiv (\xi_{1t}, \xi_{2t}, \cdots, \xi_{St}) \sim CAR(\tau_\xi^2)
\end{aligned}$$

with $\rho \in [-1,1]$. For identifiability purpose, we set $\beta_0(s,1) = 0$. Parameter ρ describes the temporal dependence between the spatial random effect at each location.

Residual errors $\boldsymbol{\xi}_t$ are modeled as another spatial CAR process independent across time. The temporal trend in the dynamic model follows a flexible first-order random walk distribution with

$$\beta_0(t)|\boldsymbol{\beta}_0(-t) = \begin{cases} N\left(\phi\beta_0(t+1), \tau^2\right), & t = 1 \\ N\left(\frac{\phi}{2}\left(\beta_0(t-1) + \beta_0(t+1)\right), \frac{\tau^2}{2}\right), & t = 2, \cdots T-1 \\ N\left(\phi\beta_0(t-1), \tau^2\right), & t = T \end{cases}$$

Finally, the 4^{th} approach we assessed in this study is an **autoregressive** approach recently introduced by Martínez-Beneito et al. [2008]. Similar to the dynamic model, this model allows the spatial random effects to vary smoothly in time. However, it also constrains the spatial random effect to be stationary with an identical covariance matrix at each time point. Following equation 2.2, we set $\beta_0(s) = 0$, and

$$\begin{aligned} &\beta_0(s,t) = (1-\rho^2)^{-\frac{1}{2}}(\theta_{s,t} + \psi_{s,t}), t = 1 \\ &\beta_0(s,t) = \rho\beta_0(s,t-1) + \theta_{s,t} + \psi_{s,t}, t > 1 \\ &\theta_{s,t} \overset{iid}{\sim} N(0, \sigma_\theta^2) \\ &\boldsymbol{\psi}_t \equiv (\psi_{1t}, \psi_{2t}, \cdots, \psi_{St}) \overset{iid}{\sim} CAR(\tau_\psi^2). \end{aligned}$$

Again parameter $\rho \in [-1, 1]$ captures temporal dependence. For this model, we consider modeling the temporal trends, as (1) linear in last 4 years; (2) first-order random walk; and (3) setting $\beta_0(t) = 0$.

2.4.3 Accounting for geographic unit boundary changes

In November of 2001, a new county (Broomfield, FIPS code 08014) was created from portions of Adams, Boulder, Jefferson and Weld counties in Colorado. Consequently, some of the census tracts before 2000 in these four counties now wholly or partially belong to the newly created Broomfield County. Spatial misalignment is a common challenge in spatial epidemiology where the analysis is carried out at different spatial resolution or aggregation than the collected data. There exists considerable literature on this issue. For areal unit data, this problem was also known as the modifiable areal unit problem (MAUP), to describe the variable's distribution at a new level of spatial aggregation[Banerjee et al., 2004, Gotway and Young, 2002].

In order to allow the spatial random effects to vary temporally, we propose the following approach to account for changes in county boundaries during the study period (1998-2007). When fitting the spatial-temporal models, we use the most recent county geography (2007) that includes Broomfield. Then from 2002 till 2007,we observe count O_{stk} given the at-risk population size P_{stk} without boundary mismatch. Before 2002, we assume O_{stk} arises from the unobserved counts $\tilde{O}_{itk}$ for $t = 1998, \cdots 2001$ given by

$$O_{stk} = \sum_{i=1}^{I} \alpha_{ik}^{s} \tilde{O}_{itk},$$

where α_{ik}^{s} are known constants that represent the proportion of the at-risk population group k in region i that resides in the desired region s. By conditional independence, the observed counts before 2002 has the likelihood

$$O_{stk} \quad \sim \quad Poisson\left(\sum_{i=1}^{I} \alpha_{ik}^{s} P_{itk} e^{\mu_{itk}}\right)$$

If county i did not change boundary, $\alpha_{ik}^{s} = 1$ for a unique i, and $\alpha_{ik}^{s} = 0$ for all other i. Otherwise, the counts observed in county s consists of a weighted sum of counts as if the 2007 county geography was used before 2002. Therefore, our approach models the latent log relative risk μ_{itk} across the entire study period and treats observed values that do not conform to the spatial geography as coarsened data. We note that the choice of county geography in this approach is not important as long as we can identify sensible weights to relate the latent risks and the observed counts, usually through the at-risk population distribution. Here we chose to use the county boundaries in the most recent year (2007) to simplify posterior predictive calculations when projecting future cancer incidence.

The CDC National Center for Health Statistics (NCHS) does not provide population estimates of years before 2002 for Broomfield County. To estimate the proportions of population in the newly-created Broomfield who actually lived inside Adams, Boulder, Jefferson or Weld counties before 2002, we utilized the American Community Survey 5-year estimate data of census tract-level population counts.

2.4.4 Estimation and computation details

All model fitting and posterior distribution analysis were implemented in WinBUGS and R using MCMC algorithm. We chose prior distributions, $gamma(0.5,\ 0.005)$ and $gamma(0.5,\ 0.005)$, respectively for precision parameters τ_θ^2 and τ_ψ^2. Noninformative priors were chosen for fixed effects $\boldsymbol{\beta}$. In each model, we ran 2 independent sampling chains with 20,000 iterations, where the first 10,000 samples were discarded as burn-in.

Cancer incidence projections were obtained by sampling from the corresponding posterior predictive distributions. Specifically, from equation 2.2, the log relative risk is decomposed into the fixed effect $\mathbf{X}_{stk}\boldsymbol{\beta}$ and the county-level baseline risks: $\beta_0(s,t)$, $\beta_0(s)$ and $\beta_0(t)$. We assumed the fixed effects to be identical for the future period and their posterior samples are used directly for future year's projection.

For the nested models, spatial correlations were estimated independently across years. Therefore, parameters associated with $\beta_0(s,t)$ for the most recent year (2005) were carried forward as they should most closely reflect the spatial residual for the counties in the future years. However, for the dynamic and auto-regressive models, with a lag-1 autoregressive spatial-temporal effect, it's not necessary to carry over the model's last year heterogeneity and spatial random effects. Instead, we predicted the future values forward in time, by directly sampling the spatial effects and heterogeneity terms for year 2006 and year 2007 in sequence, given the parameter values in year 2005.

We followed Schmid and Held [2004] to obtain posterior predictive samples of the CAR spatial random effects. Specifically, we assume $\boldsymbol{\theta}$ follows a zero-mean CAR distribution with precision parameter τ^2. We wish to obtain a sample of $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_S)$ under the linear constraint $\sum \theta_s = 0$. The precision matrix of $\boldsymbol{\theta}$ is given by $\mathbf{Q} = \tau^2(\mathbf{D} - \mathbf{W})$ of rank $n-1$, where $\mathbf{W}$ is a symmetric adjacency matrix, with its element $W_{ss'}$=1 indicating that two units share boundaries, and $\mathbf{D}$ is a diagonal matrix with $D_{ss} = \sum_s W_{ss'}$.

The density of θ subject to the constraint $\mathbf{1}'\boldsymbol{\theta} = 0$ is proportional to

$$exp\left(-\frac{1}{2}\boldsymbol{\theta}'(\mathbf{Q} + \mathbf{1}\mathbf{1}')\boldsymbol{\theta}\right). \tag{2.3}$$

To obtain a sample $\boldsymbol{x} \sim [\boldsymbol{\theta}|\mathbf{1}'\boldsymbol{\theta} = 0]$, first sample $\boldsymbol{z} \sim N(\mathbf{0}, \hat{\mathbf{Q}}^{-1})$, where $\hat{\mathbf{Q}} = \mathbf{Q} + \mathbf{1}\mathbf{1}'$.

Then compute

$$\boldsymbol{x} = \mathbf{z} - \hat{\mathbf{Q}}^{-1}\mathbf{1}(\mathbf{1}'\hat{\mathbf{Q}}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\mathbf{z}).$$

With the log relative risk of each record calculated, the corresponding expected number of cases were obtained by using the population size with the relative risks stratified by age, sex and race at each county. Posterior samples of the expected cases were further used as the mean of a Poisson distribution, from which posterior samples of projected cases were drawn. We utilized the following 4 statistics to evaluate the predictive performance of the projections obtained from different spatio-temporal model specifications: the Average Absolute Relative Deviation (AARD), Root-Mean-Square Error (RMSE), Mean Absolute Error (MAE), 95% Posterior Interval (PI) length, and its empirical coverage probability.

2.5 Results

There were a total of 19,398 Lung and Bronchus cancer cases diagnosed between 1998 and 2007 reported to NPCR from Colorado by Dec 2009, within a range of 1744 to 2128 incidence per year. 96.6% of the cases were white, and 52.7% were male. Among the cases, the age distribution was as follows: 2.1% less than 45 years old, 8.2% between 45 and 54, 20.7% between 55 and 64, 33.6% between 65 and 74, 28.1% between 75 and 84, and 7.3% older than 84. The total population of Colorado increased from 3,966,442 in 1998 to 4,628,508 in 2007.

We use the deviance information criterion (DIC) to compare fit between models. Table 2.1 summarizes the 7 model specifications, their DIC and the effective number of parameters (pD). The nested models have the largest DIC and pD as they do not borrow information across years to estimate the spatial random effects. The static model with constant spatial effects has the smallest DIC and pD, hence the best fit.

Table 2.1: Model specification and deviance comparisons

Model	No.	Temporal effect	pD	DIC
Static model	1	linear trend of last 4 years	56.586	13806.3
Nested Model	2	constant precision parameters	162.175	13924.3
	3	year-specific precision parameters	163.398	13935.5
Dynamic model	4	auto-regressive	75.38	13818.4
Auto-regressive	5	auto-regressive	89.046	13814.4
	6	none	86.425	13814.8
	7	linear trend of last 4 years	88.451	13815.5

Tables 2.2 and 2.3 give the RMSE, AME, 95% PI coverage probabilities and their average lengths. Results are stratified by counties with less than or greater than 20 cases reported in 2006 and 2007. In counties with less than 20 cases reported, we have an average of 4.9 cases observed in 2006 and 5.2 in 2007; in counties with more than 20 cases, there were an average of 109.3 cases observed in 2006 and 117.3 in 2007. We observed that model 7, the autoregressive model with 4 years linear trend, has a better performance with a coverage closer to 95%, narrower 95% PI lengths, as well as smaller RMSE and AME for both 2006 and 2007. Across models, greater prediction errors were also associated for year 2007 compared to year 2006, as we move farther from the data.

Table 2.2: Average county-level aggregated root mean squared error (RMSE), mean absolute error (AME), 95% posterior interval (PI) average length and coverage probability, and average absolute relative deviation (AARD) by county Incidence size (< 20 vs. ≥ 20) , 2006

Model	95%Coverage		PI length		AARD		RMSE		AME	
	≥ 20	<20	≥ 20	<20	≥ 20	<20	≥ 20	<20	≥ 20	<20
1	97.40%	99.44%	5.84	1.04	8.18	0.24	1.38	0.24	0.63	0.81
2	97.14%	99.68%	5.64	1.19	8.61	0.26	1.41	0.26	0.64	0.80
3	96.35%	99.68%	5.61	1.18	8.67	0.27	1.42	0.26	0.64	0.81
4	96.61%	99.54%	5.81	1.09	8.85	0.24	1.41	0.25	0.63	0.81
5	97.14%	99.54%	5.87	1.09	8.58	0.24	1.39	0.25	0.63	0.81
6	97.14%	99.68%	5.83	1.13	8.83	0.25	1.41	0.25	0.64	0.81
7	96.88%	99.54%	5.62	1.06	8.09	0.24	1.37	0.25	0.63	0.81

Table 2.3: Average county-level aggregated root mean squared error (RMSE), mean absolute error (AME), 95% posterior interval (PI) average length and coverage probability, and average absolute relative deviation (AARD) by county incidence size ($<$ 20 vs. $\geq$ 20) , 2007

Model	95%Coverage		PI length		AARD		RMSE		AME	
	$\geq$ 20	<20	$\geq$ 20	<20	$\geq$ 20	<20	$\geq$ 20	<20	$\geq$ 20	<20
1	97.11%	99.32%	6.24	1.07	8.94	0.25	1.46	0.26	0.62	0.82
2	96.33%	99.56%	5.78	1.20	10.08	0.29	1.50	0.28	0.63	0.82
3	96.33%	99.43%	5.73	1.19	10.29	0.30	1.51	0.28	0.63	0.82
4	96.85%	99.45%	6.07	1.10	9.58	0.26	1.48	0.26	0.62	0.81
5	97.11%	99.45%	6.24	1.12	9.66	0.26	1.48	0.26	0.62	0.81
6	97.11%	99.45%	6.09	1.15	9.91	0.26	1.48	0.27	0.63	0.81
7	96.31%	99.45%	5.75	1.06	9.19	0.25	1.47	0.25	0.62	0.82

Although having a smaller DIC value, the static model (model 1) has a larger coverage probability, a wider PI length, and an elevated RMSE and AME. The spatial CAR precision (τ_{ψ}^2) was estimated to be 6.5 with a 95% PI [3.43, 25.8] in model 1, while the spatial CAR precision of the autoregressive model (model 7), was 175 with a 95% PI [64.8, 591]. This indicates that stronger spatial associations were estimated when the spatial random effects are stratified by year. Model 1 also has a smaller posterior precision (σ_{θ}^2) for the unstructured heterogeneous random effects compared to model 7. Therefore, for the static model, the constant spatial random effects could not effectively explain the spatial-temporal variation in the data, resulting in larger projection uncertainty through the unstructured random effect.

Table 2.4 gives the posterior means and 95% PIs of the parameters in model 7. In this study, the race effect was found to be not significant in the final model, which is consistent with findings from others studies[Stellman et al., 2003]. However, due to small proportion of black population living in Colorado, the estimated relative risk (RR) of white vs. black should not be extended to the US general population for scientific reference. We found that higher cancer incidence was also associated with males and the older age groups. There was also a recent decreasing temporal trend. The baseline spatial relative risks showed strong temporal correlation from year to year. In this autoregressive model, the parameter ρ was estimated to be 0.965 [95% P.I. 0.923, 0.985].

Table 2.4: Parameter estimates from model (7) with a last 4-years linear trend

Parameters	mean	2.5%	97.5%
intercept	-6.05	-6.20	-5.92
male	0.26	0.23	0.30
white	-0.03	-0.14	0.09
age			
< 45	-5.16	-5.28	-5.05
$45 \sim 54$	-0.79	-0.86	-0.72
$55 \sim 64$	-2.51	-2.57	-2.44
$65 \sim 74$	-1.24	-1.29	-1.19
$75 \sim 84$	-0.16	-0.20	-0.12
≥ 85		reference	
year			
1998	-0.02	-0.10	0.07
1999	-0.06	-0.13	0.02
2000	-0.01	-0.09	0.06
2001	0.01	-0.06	0.08
linear	-0.04	-0.06	-0.01
σ_θ^2	635	264	1294
σ_ψ^2	214	65	591
ρ	0.97	0.92	0.99

Figure 2.1 plots the posterior means of the county-level log baseline relative risks, $\beta_0(s,t) + \beta_0(s) + \beta_0(t)$, between 1998 and 2005 from model 7. It shows that the counties along the state borders had generally higher baseline relative risks. Finally, Figure 2.2 and 2.3 plot the expected $(P_{stk}e^{\mu_{stk}})$ and the projected lung and bronchus cancer cases for each county. The observed cancer cases are shown by the hollow circles. Counties with higher projected cancer incidence are associated with higher standard deviation. Overall, the autoregressive model with linear trend of last 4 years was able to provide adequate short-term projections.

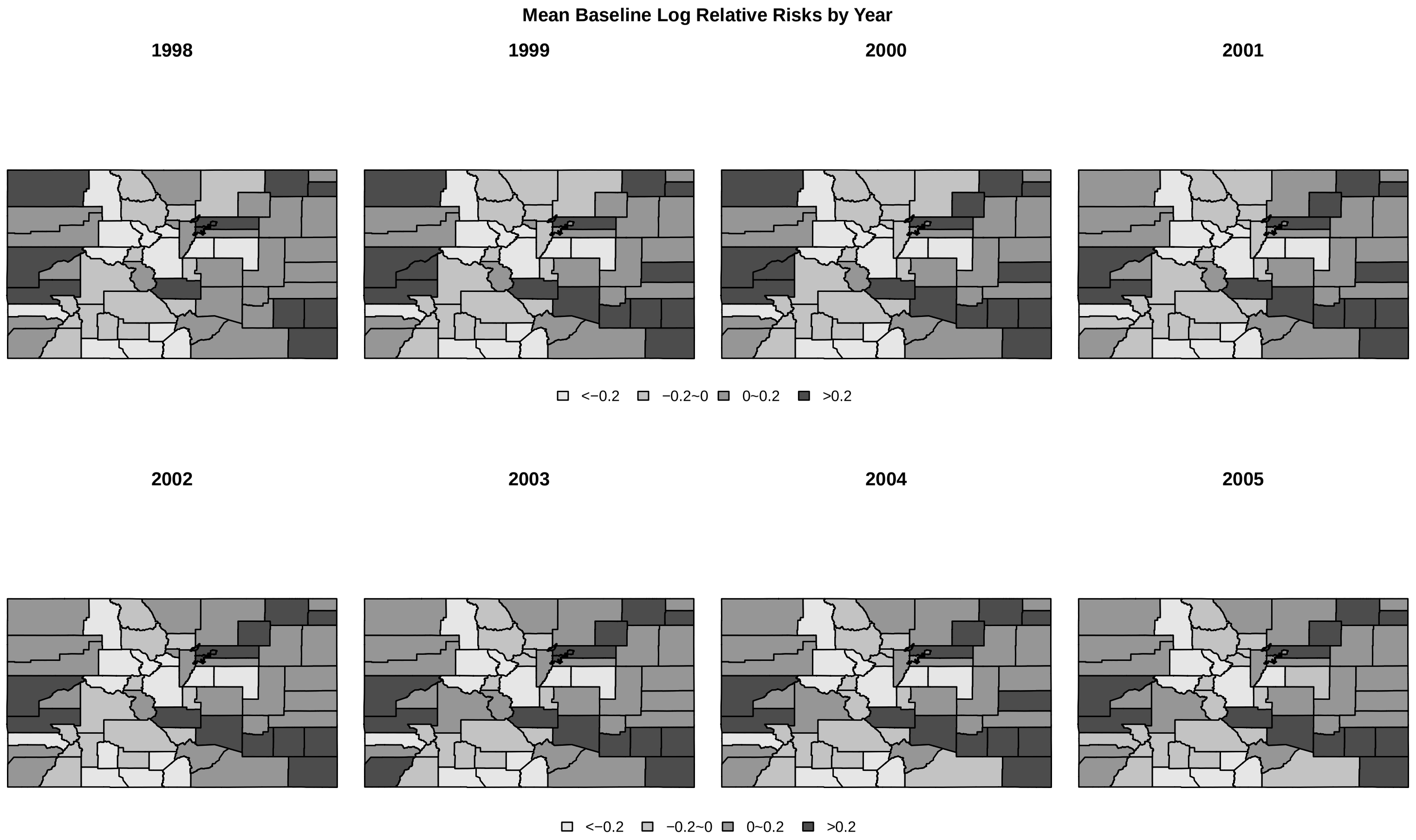


Figure 2.1: County-level Mean Log Baseline Relative Risks by Reporting Year, Colorado

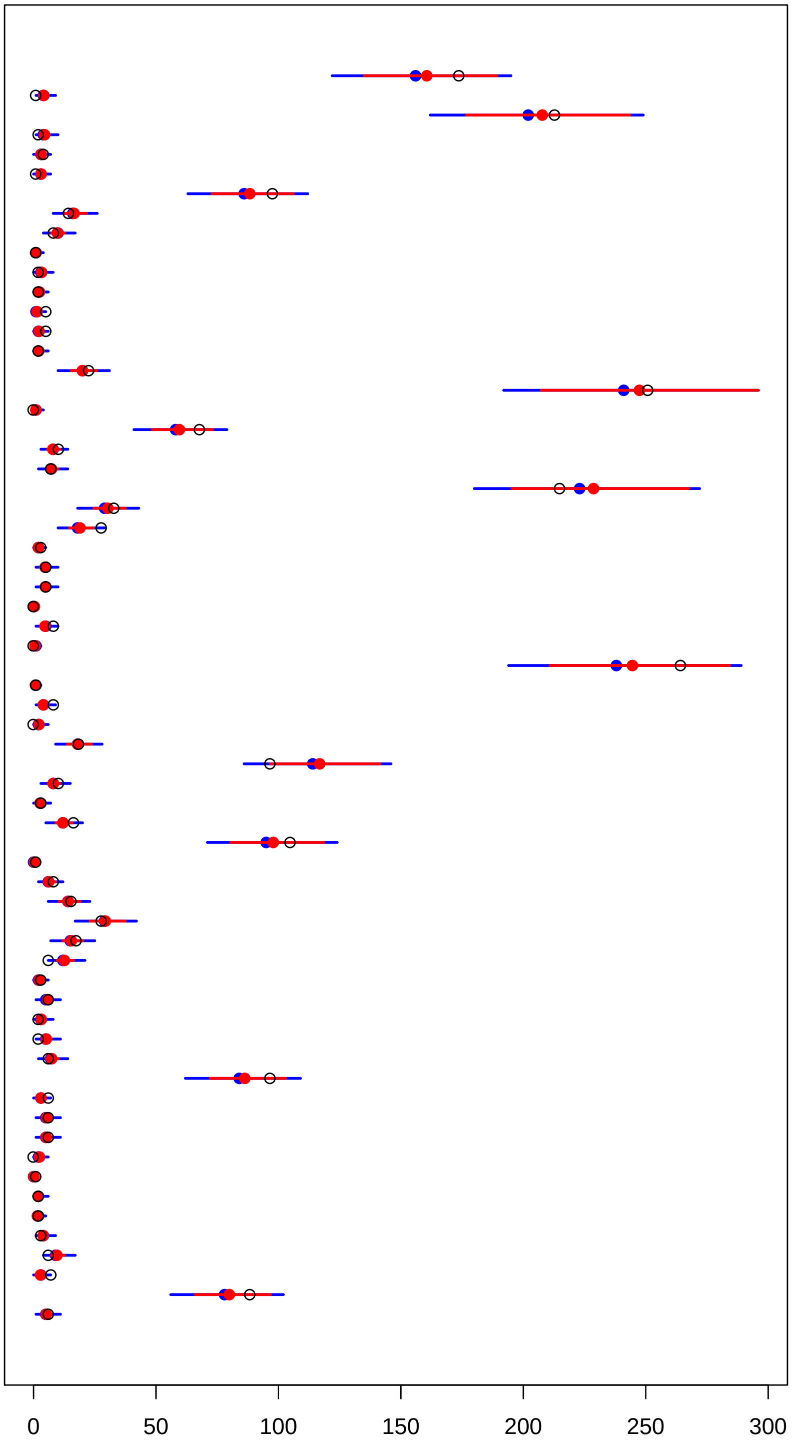


Figure 2.2: Expected(red) and projected (blue) lung and bronchus cancer cases by county from Model 7, with the observed cancer cases shown by the hollow circles, Colorado 2006

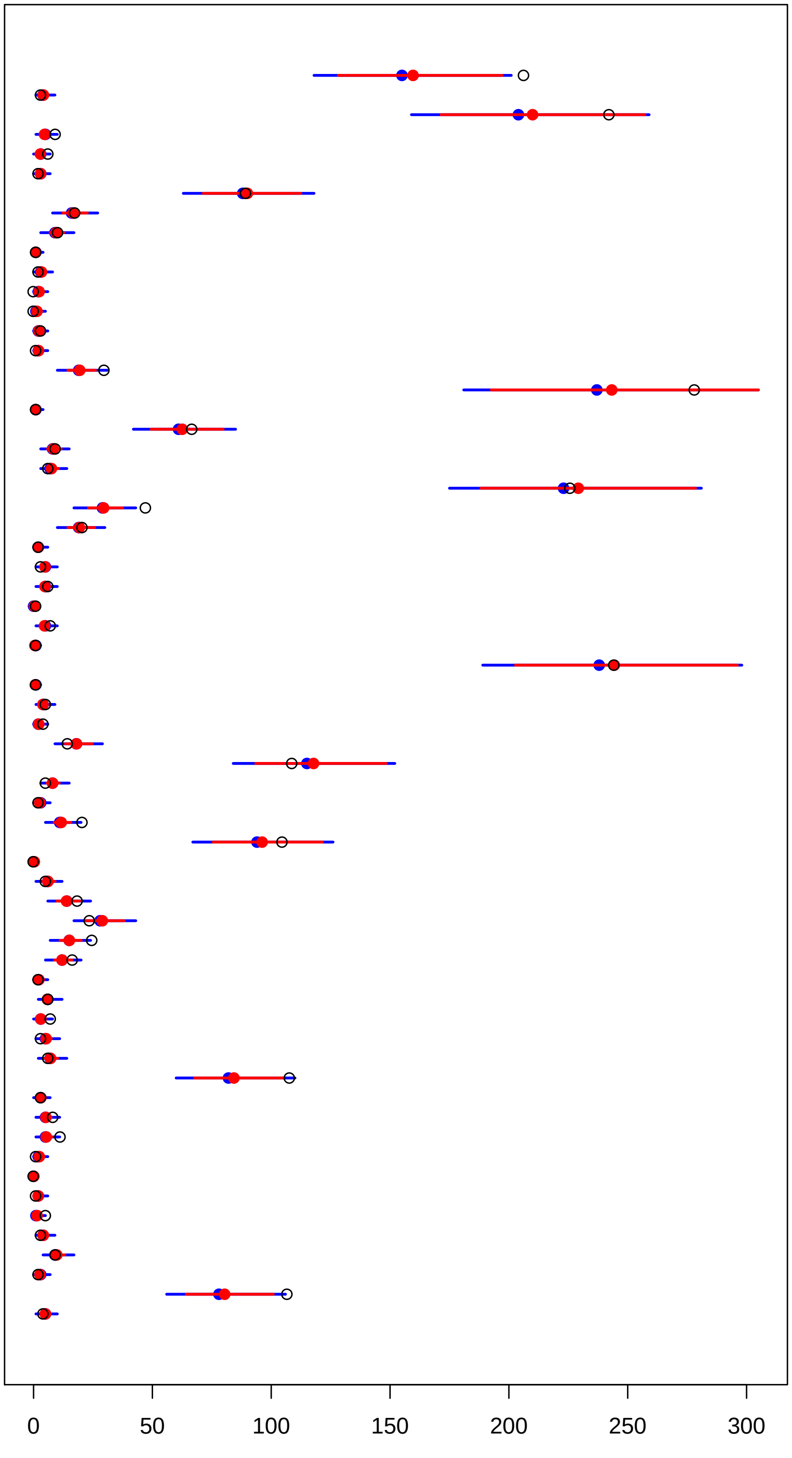


Figure 2.3: Expected(red) and projected (blue) lung and bronchus cancer cases by county from Model 7, with the observed cancer cases shown by the hollow circles, Colorado 2007

2.6 Discussion

Through comparison of the seven Bayesian hierarchical models through DIC and pD values, we observed that a Bayesian spatial-temporal autoregressive model fits the data better, which also provided better cancer incidence projection at the county level. The baseline log relative risks [Figure 1] demonstrated strong spatial correlation of the risk across time. On the other hand, spatial conditional autoregressive model smooths the cancer data by fitting a random-effect Poisson model allowing the spatial correlation to borrow strength from county neighborhoods and time. Omitting this important feature in cancer registry data will inevitably adversely affect the projection outcomes, especially of the corresponding variances.

There are additional challenges common to the analysis of cancer incidence that were not considered in this preliminary study. The first challenge arises from the fact that excess number of zeros exists often in area count data, where the observed number of zeros significantly exceeds the expected frequency given the Poisson distribution assumption. Such excess numbers of zeros are often observed in cancer incidence data. In the next chapter, we will consider zero-inflated Poisson mixture models to account for potential zero-inflated cancer data. The second challenge is to model multiple cancer types efficiently and simultaneously, such that the correlation of different cancer incidence trends can be considered in the modeling. For this purpose, we can propose other models, such as multivariate CAR model. We believe the Bayesian spatial-temporal models examined in this chapter can be conveniently extended and applied to other cancer sites, other state registries, or even the complete NPCR database.

By comparing projection performance for year 2007 in table 2.4 vs. that for year 2006 in table 2.3, we see that year 2007 posterior projection has a wider variation than that from 2006. The increased uncertainty in projection could be due to the method of sequential estimation in model 7: since spatial-temporal random effect for year

2007 is based on parameter estimation from 2006, uncertainty can be easily increases as we move further from the data. The higher RMSE, ASE can also reflect the fact that the most recent 2007 data include greater proportion of unreported cancer cases than that in year 2006 [Midthune et al., 2005].

Chapter 3

STUDY 2: BAYESIAN SPATIAL ZERO-INFLATED MIXTURE MODELS ACCOUNTING FOR ZERO-INFLATION AND OVERDISPERSION IN AREAL COUNT DATA

3.1 Introduction

When modeling count data, Poisson regression assumes equality of the conditional mean and variance of the response (*equidispersion*). However, this assumption is often violated where the variance could either be larger than the mean(*overdispersion*), or smaller than the mean*(underdispersion*). When the variance is not equal to the mean, the regression coefficient estimates in a Poisson regression model are still consistent, but inference based on the estimated standard errors is no longer valid.

One common source of overdispersion in Poisson regression in many disciplines, such as econometrics[Heilbron, 1994] and health services research, is zero inflation. Excess zeros are said to be present in data when the observed number of zeros significantly exceeds the expected frequency given the Poisson distribution assumption.

This high occurrence of zeros can reach a level such that standard Poisson regression exhibits lack-of-fit even after adjusting for covariates or introducing random effects in the model[Ghosh et al., 2012, Neelon et al., 2010].

Cancer registry data are spatial-temporal and often come with many zeros when examined within relatively small areas, such as at the county or census tract level. While some of the zero counts are true zeros of cancer incidence, the other can represent non-reports[McClintock, 2012], such as those due to possible cancer reporting delay or reporting error[Midthune et al., 2005].

This paper presents an analysis utilizing data submitted to SEER from the Iowa Central Cancer Registry (ICCR) between the period 1998 to 2005. ICCR entered the SEER program in 1973, and has a long history of high data quality. We only used data covering a recent period of time, such that the method developed could be implemented toward other cancer registries which have a shorter history of data collection. We focus on lung and bronchus cancer, the most common cause of cancer-related death worldwide and responsible for an estimated 1.37 million deaths annually[Organization, 2012]. Since US cancer surveillance, including both the SEER program and CDC's National Program of Cancer Registries (NPCR), has a standard 2-year delay in reporting cancer cases to federal partners, this work focuses on projecting short-term cancer incidence such that the most recent US cancer burden may be evaluated.

In this study, we consider spatial zero-inflated mixture models based on the Poisson distribution for areal data to account for potential excess number of zero incidence. Since [Lambert, 1992], zero-inflated count Poisson (ZIP) models have been extensively investigated in statistical research and utilized in application. There is a considerable literature on frequentist approaches to fit generalized linear regression of zero-inflated data, not only for cross-sectional studies, but also for longitudinal studies[Hall, 2000, Min and Agresti, 2005]. Recently, several authors have proposed

Bayesian hierarchical models to fit zero-inflated count data. For example, Agarwal et al. [2002] first introduced the "spatial ZIP regression" that incorporates spatial random effects within a Bayesian framework. Ver Hoef and Jansen [2007] further proposed a space-time ZIP model which includes first-order autoregressive temporal effects in a Bayesian hierarchical model. Besides zero-inflated models, we also consider two other models for count data: a negative binomial and a generalized Poisson model which are alternative methods for accommodating overdispersion and underdispersion. We will study whether they are flexible in accommodating overdispersion due to an excess number of zeros compared to the Poisson mixture approach.

3.2 Methods

3.2.1 Iowa Central Cancer Registry Data, 1998-2007

Cancer incidence data from the Iowa Central Cancer Registry (ICCR) were submitted to the National Cancer Institute's (NCI) SEER program in 2009[NCI, 2010]. Primary cancer sites were coded according to the International Classification of Diseases for Oncology (ICD-O) 3^{rd} Edition (available at http://seer.cancer.gov /siterecode /icdo3_d01272003/), and only lung and bronchus cancer cases reported for diagnosis year between 1998 to 2007 were analyzed in this study. Only malignant tumors were included, while *in situ* and other benign tumors were excluded. We did not exclude Death Clearance Only (DCO) cases in the analyses. Individual level demographic variables were also available, including patient's diagnosis year, race, sex, age and county of residence. Since minority race groups account for a very small proportion of the total incidence data (2.5%) for Iowa, only whites (including both Hispanic and non-Hispanic) were used in this study. Population count data for year 1998 to 2007 were obtained from the CDC's National Center of Health Statistics (NCHS).

3.2.2 Statistical Models

We developed Poisson models, negative binomial models, generalized Poisson models and zero-inflated Poisson models using the Iowa Cancer incidence data diagnosed between 1998 to 2005. Given the 10 years of incidence data available, we used the first 8 years (1998-2005) for model fitting and the last 2 years (2006, 2007) for evaluating model projection performance. Let s index patient's residency county, $s = 1, 2, \cdots, S$ ($S = 99$ for Iowa), and let t index reporting year. We categorized age into 6 groups ($\leq$ 44, 45-54, 55-64, 65-74, 75-84, and $\geq$ 85 years old). We aggregated the incidence data by all combinations of sex, and age groups. About 43% of the combinations of county, reporting year, sex and age group had no cases reported, which were treated as 0.

We first considered the standard Poisson model used in disease mapping. For county s, we observed O_{stk}, the number of cancer cases reported during year t in stratum k, and let P_{stk} denote the corresponding at-risk population size.

$$O_{stk} \sim Poisson(\lambda_{stk}),$$

where $\lambda_{stk} = P_{skt}e^{\mu_{stk}}$. The log relative risk, μ_{stk}, is further partitioned into the following different terms:

$$\mu_{stk} = \mu + \beta_0(t) + \beta_0(s) + \mathbf{X}_{stk}\boldsymbol{\beta}, \tag{3.1}$$

where μ is the overall average log baseline relative risk; $\mathbf{X}_{stk}$ denotes the vector of indicators for a particular age and sex groups to capture potential disparities in incidence; $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients. $\beta_0(t)$ is the purely temporal component that describes the temporal trend in cancer incidence. We used binary indicators for the first 4 years, and a linear trend for the last 4 years to capture the most recent temporal change in rate; $\beta_0(s)$ is the purely spatial component that describes the spatial trend in cancer incidence, and follows a Gaussian Conditionally Autoregressive (CAR) distribution, and its precision (the inverse of variance, σ_θ^2) has a prior of $gamma(0.5,\ 0.001)$. The CAR distribution defines a joint distribution of Markov random fields, which facilitates spatial smoothing and identifies spatial/temporal patterns. To simplify our models, We assume there is no space-time interaction in this model of this study. Prior distributions for other parameters were given previously (Chapter 2).

We then consider a negative binomial model

$$O_{stk} \sim NB\left(\gamma,\ p_{stk}\right). \tag{3.2}$$

with mean $E(O_{stk}) = \lambda_{stk} = \frac{(1-p_{stk})\gamma}{p_{stk}}$, and variance $V(O_{stk}) = \frac{(1-p_{stk})\gamma}{p_{stk}^2} = \frac{\lambda_{stk}}{p_{stk}}$. Since $p_{stk} \in (0,1)$, its variance is larger than the expected value. Therefore, a negative binomial distribution can account for the overdispersion problem in count data. We chose $gamma(0.1,\ 0.1)$ as the hyperprior for γ. To model cancer incidence, we parametrized $p_{stk} = \frac{\gamma}{\gamma+\lambda_{stk}}$, where $\lambda_{stk} = P_{skt}e^{\mu_{stk}}$ and μ_{stk} has the same components as in (3.1). Overdispersion, the ratio between its variance and mean, is estimated by $\frac{1}{p_{stk}} = \left(1 + \frac{\lambda_{stk}}{\gamma}\right)$.

We then consider a generalized Poisson distribution which has the advantage that it can be fitted accounting for both overdisperison and underdispersion:

$$O_{stk} \sim GenPoisson\left(\alpha,\ \lambda_{stk}\right). \tag{3.3}$$

and the probability density function as in [Ismail and Jemain, 2007] was specified as

$$Pr(O_{stk} = y) = \left(\frac{\lambda_{stk}}{1+\alpha\lambda_{stk}}\right)^y \frac{(1+\alpha y)^{y-1}}{y!} exp\left(-\frac{\lambda_{stk}\left(1+\alpha y\right)}{1+\alpha\lambda_{stk}}\right), \quad y = 0, 1, \cdots B(\alpha), \tag{3.4}$$

where $B(\alpha) = \lfloor -\frac{1}{\alpha} \rfloor$ (floor function, maps a real number to its previous largest integer) if $\alpha < 0$, otherwise $B(\alpha) = \infty$ if $\alpha \geq 0$. The generalized Poisson distribution has mean $E(O_{stk}) = \lambda_{stk} > 0$, and variance $V(O_{stk}) = \lambda_{stk}\left(1 + \alpha\lambda_{stk}\right)^2$. Note that because α can take both negative and positive values, its variance can account for both over- and under-dispersion. If $\alpha = 0$, generalized Poisson distribution reduces to a Poisson distribution [Fuentes et al., 2006]. We chose $beta(0.5,\, 0.5)$ as the hyperprior for α. Furthermore, we assume that $\alpha \sim \text{uniform}(l,\ u)$. The lower bound of α was set to be $-\frac{1}{O_{max}}$ such that $(1 + \alpha O_{stk}) > 0$, and the upper bound was set to 1. We transformed α to y (see appendix) for easy MCMC estimation.

The zero-inflated model assumes the excess of zeros is comprised of a mixture of a degenerate distribution at 0 and a standard Poisson, denoted as $ZIP(p, \lambda)$:

$$\begin{aligned} &P(Y = 0|\lambda_i) = p + (1-p)\pi(0|\lambda_i)\,, \qquad 0 \leq p < 1 \\ &P(Y = y_i|y_i > 0, \lambda_i) = (1-p)\pi(y_i|\lambda_i) \end{aligned} \tag{3.5}$$

where $\pi(y|\lambda_i)$ denotes the standard Poisson distribution probability mass function (PMF), and p denotes the proportion of the additional point mass at zero (structural zeros). When $p = 0$, the model reduces to the $Poisson(y|\lambda)$. When $0 < p < 1$, the model takes into account the inflated zeros $p + (1-p)\pi(0|\lambda_i) > \pi(0|\lambda_i)$.

The mean and variance of the ZIP model are given by:

$$\begin{aligned} &E(Y|p, \mu_i) = (1-p)\lambda_i \\ &V(Y|p, \mu_i) = p(1-p)\lambda_i^2 + (1-p)\lambda\mu_i \quad , \end{aligned} \tag{3.6}$$

and we see that that expected value for $ZIP(p, \lambda)$ is always smaller than its variance by $p(1-p)\lambda_i^2$, so *equidispersion* is no longer a necessary assumption for ZIP model.

To introduce covariate information in both the zero and non-zero components, the canonical links are used. Following equation (3.6), let $\lambda_{stk} = P_{stk}e^{\mu_{stk}}$, and μ_{stk} has the same partition terms as in (3.1). Let p_{stk} represent the probability of the degenerate distribution at 0.

$$\begin{aligned} log\,(\mu_{stk}) &= \mu + \beta_0(t) + \beta_0(s) + \boldsymbol{X}_{stk}\boldsymbol{\beta} \\ logit(p_{stk}) &= \mu^{'} + \alpha_0(t) + \alpha_0(s) + \boldsymbol{X}^{'}_{stk}\boldsymbol{\alpha} \end{aligned} \tag{3.7}$$

Based on preliminary analysis, $\boldsymbol{X}_{stk}$ included age groups and sex. Given the above regression specification, three models with different spatial random effect specifications were studied. The first is a model without $\alpha_0(s)$, the purely spatial component. This model assumes that the probability of zero incidence does not vary across counties after being adjusted for the temporal and covariate effects. The second model has a purely spatial component $\alpha_0(s)$, independent from $\beta_0(s)$ in the log link partition in equation (3.7).

The last ZIP model we studied includes correlated spatial random effects between the Bernoulli and Poisson components modeled by a joint bivariate intrinsic Conditional Autoregressive distribution ($biCAR(\boldsymbol{\Sigma})$) for $\alpha_0(s)$ and $\beta_0(s)$ as described by Neelon et al. [2013]. Joint modeling of the spatial random effects can reduce bias of the spatial covariance parameters and the intercept of the Poisson component as well. To be specific, let s denote the area unit spatial location, and let $\boldsymbol{\Phi^T} = (\Phi_{1,}\Phi_{2,} \cdots, \Phi_s, \cdots \Phi_S)$ where each $\Phi_s = (\alpha_s, \beta_s)^{'}$ is a 2×1 vector and has the following joint conditional distribution:

$$\Phi_s|\Phi_{(-s)} \quad \sim \quad N_2\left(\frac{1}{m_s}\sum_{l\in\partial_s}\Phi_l,\ \frac{1}{m_s}\boldsymbol{\Sigma}\right),$$

where m_s is the number of neighbors of area unit s, ∂_s is the set of neighbors for unit s, and $\boldsymbol{\Sigma}$ is a 2×2 variance-covariance matrix. In addition, we assume a $IW(3, \boldsymbol{I}_2)$ prior for $\boldsymbol{\Sigma}$. To simplify our notation, we denote $\boldsymbol{\Phi} = (\Phi_1 \cdots \Phi_S) \sim biCAR(\boldsymbol{\Sigma})$.

3.3 Results

There were a total of 22,246 lung and bronchus cancer cases in the study population between 1998 and 2007 reported to SEER program from ICCR by November 2009[NCI, 2010], with a range of 2,131 to 2,341 incident cases per year. Among the cases, 57.5% are male. And the age distribution was as follows: 1.9% less than 45 years old, 7.9% between 45 and 54, 19.5% between 55 and 64, 32.8% between 65 and 74, 29.6% between 75 and 84, and 8.3% older than 84. The total population of Iowa started from 2,795,851 in 1998, which climbed in the first 2 years and then slipped back to 2,800,261 by 2003. The population then grew to a peak of 2,826,220 in 2007. When the data are aggregated at the county level and stratified by age groups($\leq$ 44, 45-54, 55-64, 65-74, 75-84, and $\geq$ 85 years old) and sex, the percentage of annual number of zeros ranged from 41.3% to 45.4% (see table 3.1).

Table 3.2 summarizes the 6 model specifications, and gives an assessment of fit based on their deviance information criterion (DIC) and the effective number of parameters (pD) [Spiegelhalter et al., 2002]. All models included county-level spatial effects for cancer relative risks of the means. The three ZIP models differ by whether the structural zeros were modeled with a spatial effect (ZIP2), and whether the structural zero spatial effects were correlated with spatial effects for cancer relative risks (ZIP3). The negative binomial model and generalized Poisson model had the smallest DIC and pD values indicating the best fit. Model ZIP3 (model 6) has the largest DIC and pD values. This model comparison suggested that the Iowa cancer incidence data exhibited overdispersion which can not be explained solely by excess zeros.

We utilized the following 5 statistics to evaluate the predictive performance of the projections obtained from different spatio model specifications: the Average Absolute Relative Deviation (AARD), Root-Mean-Square Error (RMSE), Mean Absolute Error (MAE), 95% Posterior Interval (PI) length, and its empirical coverage probability (See Chapter 2). The 5 statistics are summarized in Table 3.3 and 3.4 , with results stratified by counties with less than or greater than 20 cases reported in 2006 and 2007. In the stratum of counties with less than 20 cases reported, we have an average of 10.8 cases observed for both 2006 and 2007; in the other stratum of counties with at least 20 cases observed, there were an average of 61.9 cases observed for 2006 and 59.8 cases for 2007. Overall, we found that all models performed similarly in term of coverage probability. Despite its DIC value, predictions from the negative binomial model had the largest PI length, AARD, RMSE, and MAE. Across models, greater prediction errors were also associated for year 2007 compared to year 2006 as we move further from the data. We observed that that negative binomial model (model 2) and generalized Poisson model (model 3) have wider PI from both 2006 and 2007 compared with other models, which resulted from the model's accommodation of overdispersion in the data.

Figure 3.1 plots the posterior means of the county-level log baseline relative risks, $\beta_0(s)$, in equation (3.1), from the Poisson model (model 1), the negative binomial model(model 2), the generalized Poisson model(model 3), and the zero-inflated Poisson model with no spatial effect for structured zeros (model 4). It shows that the counties along the state borders had generally higher baseline relative risks. Finally, Figure 3.2 and Figure 3.3 plot the expected ($P_{stk}e^{\mu_{stk}}$) and the projected lung and bronchus cancer cases for each county. Overall we found good agreement between the projections and the observed counts. The figures also include 95% posterior intervals (PI) and our model appears well calibrated. The observed cancer cases are shown by the hollow circles.

The parameter of γ in equation (3.2), which represents the "number of failures until the experiment is stopped" in negative binomial model (Model 2) was estimated to have median of 41.0 and 95% posterior interval is [28.8, 62.0] (posterior sample provided in Figure 3.4).

The parameter α, given in equation (3.3), in Generalized Poisson model (Model 3) was estimated to have posterior median 0.0264 (95% PI 0.0118 - 0.0419). Both negative binomial model and generalized Poisson model indicate the presence of overdispersion in the Iowa Lung and Bronchus cancer incidence data.

The parameter p, the proportion of structured zeros in the zero-inflated Poisson model (Model 4), was estimated for male and female separately. For male the median of p was estimated to be 0.659 and 95% PI (0.532, 0.758). For female the median of p was estimated to be 0.431 and 95% PI (0.366, 0.518).

Table 3.5 gives the posterior means and 95% PIs of the parameters in the negative Binomial model. In this study, we found that higher cancer incidence was also associated with males and older age groups. There was also a recent decreasing temporal trend.

3.4 Discussion

In general, Bayesian inference has advantages of incorporation of prior information, and avoidance of asymptotic assumptions. Specifically for small dataset with many zeros, Bayesian ZIP models have shown to provide better performance compared to the maximum likelihood in terms of both bias and precision [Ghosh et al., 2006].

In addition including spatial correlation using biCAR prior, Fuentes et al. [2006] illustrates the importance of inclusion of the temporal correlation when doing Bayesian spatial-temporal modeling of count variables. This approach was evaluated and did not provide significant improvement in model fit, and the results were not shown here.

Table 3.1: Percent of Zero Counts in Iowa Lung and Bronchus Cancer Data Stratified by County, Age Groups and Sex, White Population

Diagnosis Year	Total Records	Zero Case Percentage
1998	1188	45.4%
1999	1188	43.9%
2000	1188	42.3%
2001	1188	45.1%
2002	1188	42.3%
2003	1188	41.3%
2004	1188	42.8%
2005	1188	41.4%
2006	1188	41.5%
2007	1188	42.6%

Table 3.2: Model Specification and Comparisons

Model	No.	Spatial effect	pD	DIC
Poisson	1	mean	72.0	24416
Negative Binomial	2	mean	69.7	24400
Generalized Poisson	3	mean	70.3	24400
Zero-inflated Poisson (ZIP1)	4	no spatial effect for structured zeros	68.2	24413
ZIP2	5	spatial effects for structured zeros	67.6	24428
ZIP3	6	spatial effect for p correlated with RR spatial effects	79.6	24530

Table 3.3: Average county-level root mean squared error (RMSE), mean absolute error(AME), 95% posterior interval (PI) average length and coverage probability, and average absolute relative deviation (AARD) by county incidence size ($<$ 20 vs. $\geq$ 20). Predictions are for the year 2006.

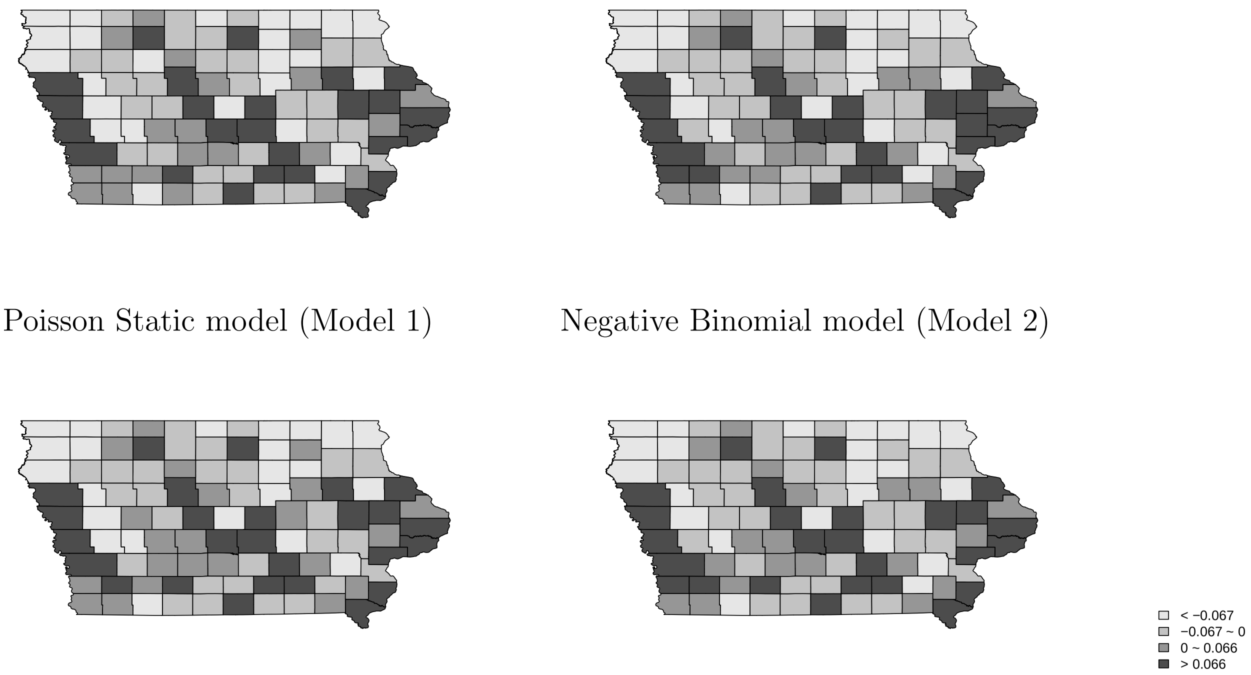
Model	95%Coverage		PI length		AARD		RMSE		AME	
	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20
1	97.57%	98.89%	7.48	3.08	0.52	0.66	7.45	1.00	1.68	0.72
2	97.92%	98.8%	8.29	3.14	0.53	0.66	7.81	1.00	1.70	0.72
3	97.57%	98.67%	7.65	3.17	0.52	0.66	7.40	1.00	1.67	0.72
4	97.22%	98.78%	7.45	3.08	0.52	0.66	7.48	1.00	1.68	0.72
5	97.57%	98.67%	7.45	3.08	0.52	0.66	7.47	1.00	1.68	0.72
6	97.22%	98.78%	7.52	3.08	0.52	0.65	7.42	0.99	1.68	0.72

Table 3.4: Average county-level aggregated root mean squared error (RMSE), mean absolute error(AME), 95% posterior interval (PI) average length and coverage probability, and average absolute relative deviation (AARD) by county incidence size ($<$ 20 vs. $\geq$ 20). Predictions are for the year 2007.

Model	95%Coverage		PI length		AARD		RMSE		AME	
	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20	$\geq$ 20	$<$ 20
1	94.79%	98.11%	7.57	3.10	0.69	0.64	8.15	0.94	1.91	0.70
2	96.18%	98.44%	8.42	3.17	0.69	0.64	8.58	0.95	1.94	0.70
3	95.49%	98.67%	7.78	3.18	0.69	0.64	8.11	0.95	1.90	0.70
4	94.44%	98.33%	7.60	3.11	0.69	0.64	8.18	0.94	1.91	0.70
5	94.79%	98.56%	7.63	3.11	0.69	0.64	8.13	0.94	1.91	0.70
6	94.44%	98.22%	7.67	3.11	0.68	0.63	8.07	0.94	1.90	0.70

Table 3.5: Posterior mean and 95% posterior interval for parameters from the negative binomial model (Model 2) with a last 4-years linear trend

Parameters	mean	2.5%	97.5%
intercept	-8.17	-8.25	-8.10
male	0.61	0.57	0.64
age			
$<$ 45	-2.86	-2.99	-2.74
45 $\sim$ 54	1.33	1.27	1.40
55 $\sim$ 64	2.16	2.09	2.22
65 $\sim$ 74	2.33	2.27	2.39
75 $\sim$ 84	1.98	1.89	2.05
$\geq$ 85		reference	
year			
1998	0.0011	-0.0572	0.0633
1999	-0.0119	-0.0735	0.0494
2000	0.0036	-0.0624	0.0631
2001	-0.0331	-0.0945	0.0261
linear	0.0099	-0.0109	0.0301
σ_θ^2	13.5	8.7	20.4
γ	42.1	28.8	62.0



Poisson Static model (Model 1)

Negative Binomial model (Model 2)

Generalized Poisson model (Model 3)

Zero Inflated model (Model 4)

Figure 3.1: County-level Baseline Relative Risks

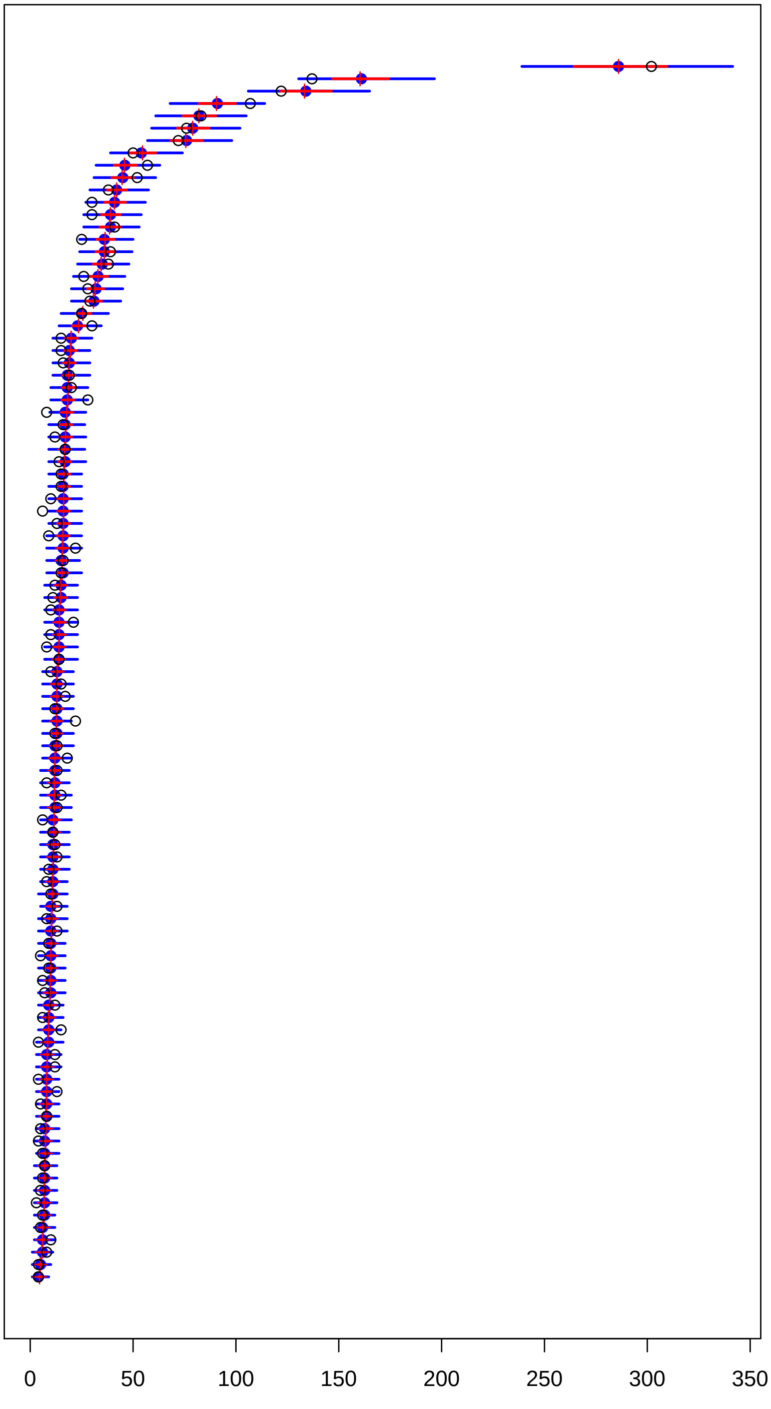


Figure 3.2: Expected(red) and projected (blue) lung and bronchus cancer cases by county with 95% posterior prediction intervals from negative binomial model (Model 2). The observed cancer cases were shown by the hollow circles, Iowa 2006

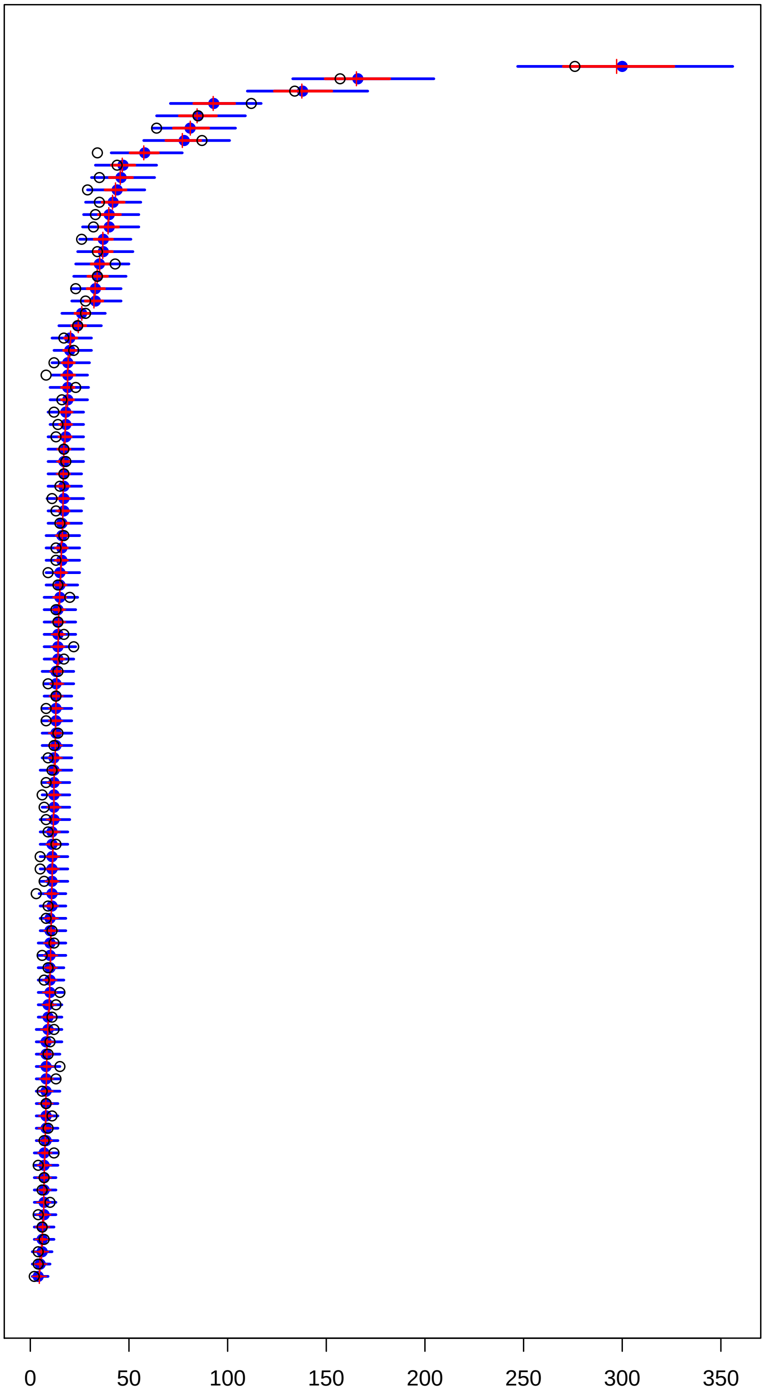


Figure 3.3: Expected(red) and projected (blue) lung and bronchus cancer cases by county with 95% posterior prediction intervals from negative binomial model (Model 2). The observed cancer cases were shown by the hollow circles, Iowa 2007

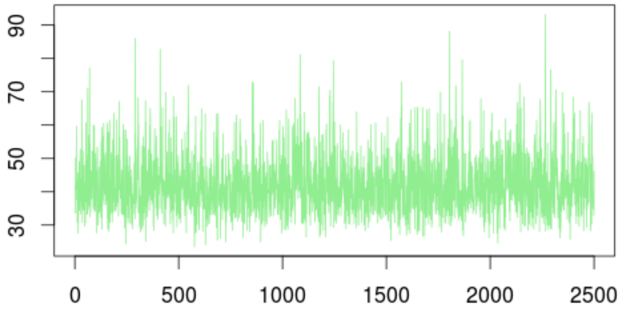


Figure 3.4: Plot of γ for Negative Binomial model (Model 2)

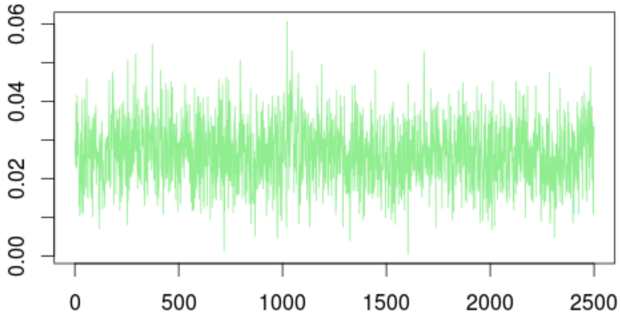


Figure 3.5: Plot of α for Generalized Poisson model (Model 3)

Chapter 4

STUDY 3: Bayesian Dynamic Spatial-temporal Hurdle Models for Zero-inflated Count Data

4.1 Introduction

When it comes to accounting for zero-inflation in count data, statisticians have other options beside the zero-inflated Poisson (ZIP) model. One of the widely used methods in this area is the *Poisson hurdle model.* The Poisson hurdle model, also called a two-stage model in econometrics[Heilbron, 1994], considers a point mass at zero and a truncated Poisson distribution for the nonzero observation, such that it does not distinguish between zeros from the degenerate distribution versus those from the Poisson distribution as in the ZIP model. Since Poisson hurdle approach models zeros and nonzeros separately, it has the flexibility of accommodating both zero-inflated and zero-deflated data, while the ZIP model can only accommodate zero-inflation(see equation 3.6).

Motivated by a study by Neelon et al. [2013] exploring spatial-temporal trends in

Emergency Department (ED) use, we develop a class of two-part hurdle models in this research for the analysis of zero-inflated areal count data. Different from surveillance data, ED visit data has no “false zeros” [McClintock, 2012]. Thence, hurdle models fit the health administrative data better than zero-inflated mixture models.

ED visit records, provided from Duke University Decision Support Repository (DSR) database, contain demographic, diagnostic and treatment information on over 4 million patients of Duke University Health System. There are many unique challenges in the analysis of these data: the data are spatial-temporal which have demonstrated considerable geographic and temporal variation in ED use. The DSR data come with abundant zeros (about 70% patients had no ED visits annually); the positive counts can go as high as 90 times a year for some patients. And finally, we need to improve small-area estimation by providing adequate spatial and temporal smoothing. The unique challenges associated with DSR data are of particular interest to address as part of this dissertation.

Our aims in this study are to develop spatial-temporal Bayesian models to address zero-inflation and potential overdispersion in the ER records from the DSR database submitted to Duke University Health System. We seek to identify areas where ED use remained persistently high, fluctuated from year to year, or increased systematically over time. To be consistent with the other two studies in my dissertation, we only used data covering a recent period of time, such that the method developed could be implemented toward other count data which have only short history of data collection. This paper will also present a cross-validation of posterior-predicted ER usage counts based on the model fit.

We evaluate a class of two-part hurdle models which are specifically designed to address those unique challenges in ER visit data. The hurdle model consists of two components: a Bernoulli component that models the probability of any ED use (i.e., have any ED visits in a given year) and a truncated count component that models the number of repeat visits among users. Together, these components accommodate both the high proportion of zeros and the right-skewness observed among the nonzero counts. To address potential overdispersion in the positive counts, we consider three distributional specifications for the nonzero observations: the truncated Poisson, the truncated negative binomial, and the truncated generalized Poisson distribution. Taking advantage of the unique hierarchical structure of the DSR data, our models incorporate both patient- and region-level predictors, as well as spatially and temporally correlated random effects for each model component.

We also seek to develop models that could accommodate the correlation between the two components of hurdle models. The previous study by [Neelon et al., 2013] has shown the probability of ED use was correlated with the expected number of ED visits among users. Therefore, the random effects are modeled via multivariate conditionally autoregressive priors that induce dependence between the components and provide smoothing across adjacent space and time periods.

4.2 Methods

4.2.1 The DSR Data, 2007-2011

The Duke University Decision Support Repository (DSR) has been in existence for over a decade, which holds 17 years of demographic, diagnostic, and billing data on over 4 million patients of the Duke University Health System. As part of a ongoing exploring study of ED use, researchers recently reviewed ED admission records for Durham County residents who were seen at either an ED or non-ED clinic between 2007 and 2011, the most recent years for which records were available. The records were listed by residential address which were subsequently linked at the Census block level to the 2005-2009 American Community Survey[US Census Bureau, 2010] data. The final dataset contained over 122,000 records from the 129 Census block groups in Durham County, and included information on the annual number of ED visits for each patient, patient-level demographics, such as age, race, gender and insurance status, and median household income of each block group.

4.2.2 The Hurdle Model

For the analysis of the DSR data, we consider a broad class of two-part hurdle models to address both zero inflation and potential overdispersion of the nonzero counts. Hurdle models are two-part mixtures consisting of a point mass at zero followed by a zero-truncated count distribution (base distribution) for the positive observations [Neelon et al., 2013, Mullahy, 1986]. Letting Y denote a count-valued response, the generic structure of the hurdle model is given by

$$Pr(Y = y) = \begin{cases} 1 - \pi, & 0 \leq \pi \leq 1,\, y = 0 \\ \frac{\pi p(y;\mu)}{1 - p(0;\mu)}, & y = 1, 2, \cdots \end{cases}, \tag{4.1}$$

where $\pi = Pr(Y > 0)$ is the probability of a nonzero response; $p(y; \mu)$ is an untruncated base probability density function with parameter θ; and $p(0; \mu)$ is the probability of probability density function (PDF) evaluated at 0. When $1 - \pi = p(0; \mu)$, the hurdle model reduces to its base distribution; when $1 - \pi > p(0; \mu)$, the zeros are inflated relative to the base distribution; and when $1 - \pi < p(0; \mu)$, there is zero deflation. When $\pi = 1$, there are no zeros and the model reduces to truncated base distribution; when $\pi = 0$, the model is degenerate at zero. However, we assume that π is strictly between 0 and 1, so that there is a nonzero utilization probability for all individuals under study.

4.2.3 Spatial-temporal Hurdle Model

For analyses in this study, we first included spatial-temporal components into model (4.1) through the dynamic space-time model developed by Chang et al. [2012]. Let y_{ijk} denote the number of annual ED visits for the k^{th} patient in block group i and year j. A general form of the spatial-temporal hurdle model is given specified as following:

$$\begin{aligned} Pr(y_{ijk}) &= \begin{cases} 1-\pi_{ijk}, & 0 \leq \pi_{ijk} \leq 1,\ y_{ijk}=0 \\ \frac{\pi_{ijk} p(y_{ijk};\mu_{ijk})}{1-p(0;\mu_{ijk})}, & y_{ijk}=1,2,\cdots \end{cases} \\ g(\pi_{ijk}) &= \boldsymbol{x}_{ijk}^{'}\boldsymbol{\alpha} + f_1(z_{ijk}) + \phi_{1i} + \nu_{1j} + \delta_{1ij} \\ ln(\mu_{ijk}) &= \boldsymbol{x}_{ijk}^{'}\boldsymbol{\beta} + f_2(z_{ijk}) + \phi_{2i} + \nu_{2j} + \delta_{2ij}, \end{aligned} \tag{4.2}$$

where $\pi_{ijk} = Pr(Y_{ijk} > 0)$; μ_{ijk} is the conditional mean of the base distribution given a set of spatial-temporal random effects; $g(.)$ denotes the logit link; $\boldsymbol{x}_{ijk}$ is a $p \times 1$ vector of fixed-effects, including both individual- and region-level predictors (here assumed identical for both model components); $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $p \times 1$ vectors of fixed-effect regression coefficients for the two components respectively; $f_1(z_{ijk})$ and $f_2(z_{ijk})$ are optional smoothing functions of a continuous predictor z_{ijk} (e.g., patient age) to be modeled via cubic B-splines; $\boldsymbol{\phi}_i = (\phi_{1j}, \phi_{2j})^{'}$ is a vector representing the purely spatial "main effects" for the i^{th} block, which are assumed to be spatially dependent in this analysis; $\boldsymbol{\nu}_j = (\nu_{1j}, \nu_{2j})^{'}$ is a vector of purely temporal "main effects" for year j; and $\boldsymbol{\delta}_{ij} = (\delta_{1ij}, \delta_{2ij})^{'}$ denotes a vector of space-time interactions, with $\boldsymbol{\delta}_{i1} \equiv 0$ for identifiability. Thus, we partition the spatial-temporal effects into three parts: a purely spatial component, represented by $\boldsymbol{\phi}_i$; a purely temporal component, represented by $\boldsymbol{\nu}_j$; and a residual interaction term, $\boldsymbol{\delta}_{ij}$. Together, these parameters capture subtle, unobserved block-group effects over time.

As a comparison, we also considered a static model where the spatial random effects are constant in time without space-time interaction ($\boldsymbol{\delta}_{ij} = (\delta_{1ij}, \delta_{2ij})' \equiv \mathbf{0}$), such that,

$$\begin{aligned} g(\pi_{ijk}) &= \boldsymbol{x}_{ijk}'\boldsymbol{\alpha} + f_1(z_{ijk}) + \phi_{1i} + \nu_{1j} \\ ln(\mu_{ijk}) &= \boldsymbol{x}_{ijk}'\boldsymbol{\beta} + f_2(z_{ijk}) + \phi_{2i} + \nu_{2j}. \end{aligned}$$

4.2.4 Choice of Base Distribution

To accurately capture the dispersion of the positive counts, we consider three choices for the base distribution: the Poisson, negative binomial, and generalized Poisson distribution. The *spatial-temporal Poisson hurdle model* is expressed as

$$Pr(Y_{ijk} = y_{ijk}|\boldsymbol{\phi}_i, \boldsymbol{\nu}_j, \boldsymbol{\delta}_{ij}) = (1-\pi_{ijk})\mathbf{1}_{(y_{ijk}=0)} + \frac{\pi_{ijk}\lambda_{ijk}^{y_{ijk}} e^{-\lambda_{ijk}}}{y_{ijk}!(1-e^{-\lambda_{ijk}})}\mathbf{1}_{(y_{ijk}>0)},$$

where λ_{ijk} is the conditional mean of the Poisson base distribution, and λ_{ijk} and π_{ijk} are modeled as in (4.2). The Poisson distribution implies equivalence of the conditional mean and variance. In many applications, this assumption is restrictive and can result in poor model fit.

An alternative is to select a negative binomial base distribution, giving rise to the *spatial-temporal negative binomial hurdle model*:

$$\begin{aligned} Pr(Y_{ijk} = y_{ijk}|\boldsymbol{\phi}_i, \boldsymbol{\nu}_j, \boldsymbol{\delta}_{ij}) = \quad & (1-\pi_{ijk})\mathbf{1}_{(y_{ijk}=0)} + \tfrac{\pi_{ijk}}{1-(\frac{\alpha}{\mu_{ijk}+\alpha})^{\alpha}} \tfrac{\Gamma(y_{ijk}+\alpha)}{\Gamma(\alpha) y_{ijk}!} \\ & \times \left(\tfrac{\mu_{ijk}}{\mu_{ijk}+\alpha}\right)^{y_{ijk}} \left(\tfrac{\alpha}{\mu_{ijk}+\alpha}\right)^{\alpha} \mathbf{1}_{(y_{ijk}>0)}, \alpha > 0, \end{aligned}$$

where μ_{ijk} is the conditional mean of the negative binomial base distribution, α is a dispersion parameter, π_{ijk} and μ_{ijk} are modeled as in (4.2). The negative binomial base distribution is appealing if there is evidence of overdispersion relative to the Poisson, i.e., a variance exceeding the mean. In particular, if $X \sim NegBin(\mu, \alpha)$, then $E(X) = \mu$ and $V(X) = \mu(1 + \mu/\alpha)$, hence μ/α is a measure of overdispersion. As $\alpha \to \infty$, the negative binomial converges to a Poisson distribution with mean and variance equal to μ. The added flexibility of the negative binomial in accommodating heterogeneity can yield improved model fit for highly dispersed count data.

Lastly, we consider the *spatial-temporal generalized Poisson hurdle model*:

$$\begin{aligned} Pr(Y_{ijk} = y_{ijk} | \boldsymbol{\phi}_i, \boldsymbol{\nu}_j, \boldsymbol{\delta}_{ij}) = \quad & (1 - \pi_{ijk})\mathbf{1}_{(y_{ijk}=0)} + \frac{\pi_{ijk}}{1 - exp(-\frac{\mu_{ijk}}{1+\alpha\mu_{ijk}})} \left(\frac{\mu_{ijk}}{1+\alpha\mu_{ijk}}\right)^{y_{ijk}} \\ & \times \left(\frac{1+\alpha y_{ijk}}{y_{ijk}!}\right)^{y_{ijk}-1} exp\left\{-\frac{\mu_{ijk}(1+\alpha y_{ijk})}{1+\alpha\mu_{ijk}}\right\} \mathbf{1}_{(y_{ijk}>0)}, \alpha > 0, \end{aligned}$$

for $y_{ijk} = 0, 1, \cdots, C(\alpha)$ [Ismail and Jemain, 2007, Fuentes et al., 2006]. Here, μ_{ijk} denotes the conditional mean of the generalized Poisson distribution and $\alpha \in (-1/y_{max}, \infty)$ is the dispersion parameter, where y_{max} is the maximum observed response; $C(\alpha) = \llcorner -1/\alpha \lrcorner$ for $\alpha < 0$ and $C(\alpha) = \infty$ otherwise, where $\llcorner\lrcorner$ denotes the floor function; and π_{ijk} and μ_{ijk} are modeled as in (4.2). As in the negative binomial case, α functions as a heterogeneity parameter accommodating departures from equidispersion. In particular, if $X \sim GPois(\mu, \alpha)$, then $E(X) = \mu$ and $V(X) = \mu(1 + \alpha\mu)^2$. When $\alpha = 0$, the generalized Poisson reduces to the Poisson distribution; when $\alpha > 0$, $V(X) > E(X)$ and there is overdispersion; and when $\alpha < 0$, $V(X) < E(X)$ and there is underdispersion. Thus, unlike the negative binomial, the generalized Poisson allows for underdispersion. Moreover, while both distributions accommodate overdispersion, the generalized Poisson has a heavier tail compared to a negative binomial with the same first two moments, and is therefore well-suited for highly skewed data such as ours[Joe and Zhu, 2005].

4.2.5 Computation Details

Autoregressive priors are subsequently used to provide spatial-temporal smoothing and "sharing" of information across neighboring block groups and adjacent years.

Previous work by Neelon et al. [2013] shown the probability of ED use was associated with the expected number of visits given use (i.e., the model components were correlated), and explicitly modeling this between-component correlation improved inferences. To accommodate this association in the current study, and to provide adequate spatial and temporal smoothing, we assume bivariate intrinsic CAR (biCAR) priors for the spatial random effects $\boldsymbol{\phi}_i = (\phi_{1j}, \phi_{2j})'$ [Mardia, 1988]. Please see the previous chapter for the biCAR prior specification.

For the temporal main effects, we consider models with fixed annual effects; we assign independent $N(0,\, 100)$ priors to ν_{1j} and $\nu_{2j} (j = 2, \cdots 5)$, with ν_{11} and ν_{21}, set to 0 in correspondence with the reference year 2005.

Since previous studies have suggested a nonlinear effect for patient age[Niska et al., 2010], we modeled age using cubic B-splines with interior knots at the first, second, and third quartiles of the age distribution (20, 38 and 55 years, respectively).

The models also included patient race, gender, age, and insurance, and block-group median income as predictors. We assign improper priors to the fixed-effect intercepts, and diffuse normal priors to the remaining fixed effects and spline coefficients, inverse-Wishart (IW) priors to covariance matrices, and a $U(0,\, 1)$ prior to the temporal autoregressive dependency parameter, ρ. For the negative binomial hurdle model, we assign a Gamma prior $gamma(0.01,\, 0.01)$ to α, and for the generalized Poisson hurdle model, we assume $\alpha \sim \text{uniform}(-1/y_{max},\, M)$ for a suitably large $M > 0$ that ensures α is bounded away from the lower limit, where in our study, $y_{max} = 91$ and M was set to be 10.

Posterior computation proceeds via Markov chain Monte Carlo (MCMC), which can be implemented easily within WinBUGS. However, since WinBUGS does not have a pre-designated function for truncated count distributions, we apply the "zeros trick" to explicitly define the hurdle likelihood [Spiegelhalter et al., 2007]. The BICAR prior can be specified with the *mv.car* function, and the remaining MCMC steps are readily coded using standard WinBUGS syntax. The WinBUGS codes for some spatio-temporal hurdle models for DSR analyses are provided in the Appendix.

We monitor MCMC convergence using trace plots and Geweke's z-test, which assesses the distributional similarity of disjoint portions of the sampler. For model comparison, we keep using the deviance information criterion (DIC) for model comparison purpose as proposed by Spiegelhalter et al. [2002] in previous chapters.

To further evaluate model fit, we implemented a series of posterior predictive assessments, whereby the observed data were compared to data replicated from the posterior predictive distribution. If the model fits well, the replicated data should resemble the observed data. To quantify the degree of similarity, one typically chooses a "discrepancy statistic", such as a sample moment or quantile, that captures some important aspect of the data. For the DSR analysis, we adopt three discrepancy measures: the sample proportion of zeros and the sample mean and variance among the positive observations. For each measure, we compute the posterior predictive mean and 95% credible interval. A 95% credible interval that includes the observed sample statistic suggests adequate model fit. In addition to the above measures, for the final model we also produce a histogram comparing the observed and posterior-predictive counts of ED visits.

The models were fit in WinBUGS 1.4.3 and called into R using the function R2WinBUGS. We ran the sampler for 15,000 iterations, discarding the first 5,000 as burn-in. Trace plots and Geweke diagnostics indicated rapid convergence and efficient mixing of the chains.

4.3 Results

Summary statistics of patients and geographic block were provided in Table 4.1 for the five study years period. 59% of the patients were female and 46% and 42% of patients are non-Hispanic white and non-Hispanic black, respectively. The median age was 38 years. About 60% had private medical insurance, 11% as part of a University-sponsored plan. The median household income was just over $45K, approximately $5000 below the national average [DeNavas-Walt et al., 2013]. The median block group sample size, combined over five years, was 776.

Table 4.2 presents the model comparison results across models. The negative binomial and generalized Poisson hurdle models substantially outperformed the Poisson models with respect to DIC. Overall, the static generalized Poisson model with fixed map had the lowest DIC value (210983). In terms of posterior predictions, all models accurately reproduced the observed proportion of zeros and the conditional mean among the positive values, while none of the models did especially well in predicting the observed conditional variance. The ordinary Poisson models showed the poorest fit, confirming overdispersion exists and need to be accounted for in modeling these data.

Table 4.3 presents the posterior means and 95% credible intervals for the three hurdle models with fixed annual effects. The effect estimates and intervals for the binary component were similar across models, which is expected since this component has the same structure in all three models. Male gender, non-Hispanic black and Hispanic race/ethnicity, and non-private insurance, including Medicaid, Medicare and Self-paid, were associated with increased probability of ED use, while patients of Asian race and Duke insurance were associated with decreased probability of use. Patients' median household income had minimal impact on ED use.

In contrast to the binary component, the parameter estimates in the count component varied substantially across the models (Table 4.4), indicating that the choice of base distribution has a significant impact on estimating covariate effects . For example, non-Hispanic black race showed a much stronger effect for the negative binomial and generalized Poisson models than for the ordinary Poisson models. A similar, although less transparent, phenomenon occurred for the federal and self-insurance categories: while all models showed a positive effect, the effect was most pronounced in the two overdispersed models. Interestingly, for all models, the estimates for male gender and Hispanic race reversed direction between the binary and count components. Hispanics, for example, were more likely than non-Hispanic whites to visit the ED at least once; however, among ED users, they tended to make fewer repeat visits than whites. This points to a potential difference between the way Hispanics and non-Hispanic whites use ED services. In particular, although modest ED use seems to be more ubiquitous among Hispanics, they are less inclined than whites to use EDs repeatedly. And finally, there was moderate correlation between the components for the spatial main effects ($\rho_\phi = 0.46$), suggesting a modest benefit to modeling the between-component association.

Table 4.5 also shows that the binary components in the three hurdle models have compatible spatial covariance (Sigma.phi[1,1]), while the spatial covariance for the count component differs from each other quite significantly. The generalized Poisson hurdle model has the biggest spatial covariance of 0.39, compared to the spatial covariance of approximate 0.30 for the Poisson hurdle and Negative Binomial models. The covariance for dynamic CAR covariance is also similar across the three models at 0.03 for the binary components for the three models. The dynamic CAR covariance for the count component also differs from each other dramatically, with the Poisson hurdle model having the hightest at 0.16 and the negative binomial having the lowest at 0.06.

Figure 4.1 displays the predicted spatial-temporal effects, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ from the generalized hurdle model, where $\eta_{1ij} = \nu_{1j} + \phi_{1i} + \delta_{1ij}$ and $\eta_{2ij} = \nu_{2j} + \phi_{2i} + \delta_{2ij}$. The highest ED activity occurred among the central block groups and the lowest among the block groups in the southwest corner of the county. Across years, the most significant change took place between 2007 and 2008, with several central block groups transitioning into the highest usage category (represented by the darkest shade) and the southwestern block groups transitioning into the lowest category (represented by the lightest shade). The spatial pattern stabilized following 2008 with only minor fluctuations in select block groups.

4.4 Discussion

Our study has introduced a series of two-part hurdle models for the spatial-temporal analysis of zero-inflated count data. The proposed models have several attractive features: spatial and temporal smoothing was incorporated in the modeling in order to improve small-area estimation; they also incorporate individual and regional-level information to explain spatial-temporal trends; and the generalized Poisson hurdle model and negative binomial hurdle models can address potential over- or underdispersion in the counts. In addition, the models can be conveniently implemented in freely available packages such as WinBUGS.

In our other paper, we also presented a spatial-temporal dynamic models, in which for the space-time interactions, we assume a first-order dynamic biCAR prior[Gelfand et al., 2005], whereby $\boldsymbol{\delta}_{ij} = \rho\boldsymbol{\delta}_{i(j-1)} + \boldsymbol{\psi}_{ij}$ as in equation (4.2)and $\boldsymbol{\psi}_{ij}$ is a biCAR. For identifiability, we set $\boldsymbol{\psi}_{i1} = 0 \ \forall i$. Unlike with the annual main effects, temporal smoothing is needed here to improve small-area estimation, particularly when one considers that the minimum block group sample size in a given year is 5, occurring in 2011. Along with the dynamic space-time interaction, we implemented two temporal effects as introduced in 4.2.5: independent $N(0, 100)$ priors to $\boldsymbol{\nu}$, the same method as in this paper. For the purely time random effects, we experimented with assigning a biCAR prior with $IW(3, \boldsymbol{I}_2)$ prior for the conditional covariance $\boldsymbol{\Sigma}_\nu$ to $\boldsymbol{\nu}_j(j = 1, 2, \cdots 5)$ analogous to the prior for the spatial effects. This choice is particularly beneficial when the temporal units are sparse, because it allows adjacent time periods to pool information to improve efficiency.

The Spatial and Dynamic CAR covariance comparisons among the three tables tell that the covariance for the binary components are similar across the three hurdle models, which is to our expectation. However, the difference among Spatial and Dynamic CAR covariance for the count components are quite big. The Poisson hurdle model has the largest the dynamic CAR covariance, the spatial-temporal random effects (such as the map shown in the Figure 4.1) change the most across the years in the Poisson hurdle model. As a result, the static Poisson hurdle model has a much worse DIC compared to that of the dynamic Poisson hurdle model. On the other hand, the generalized Poisson hurdle model with static spatio-temporal random effect shows a small DIC difference from the dynamic generalized Poisson hurdle model. The dynamic CAR autoregressive parameter ρ has a similar value of 0.6 across the three dynamic models which suggested the temporal dependency is moderate in all three models.

In our application, models accommodating overdispersion, and in particular the generalized Poisson hurdle model, substantially outperformed the ordinary Poisson hurdle model. Given that the negative binomial and generalized Poisson base distributions include the Poisson as either a limiting distribution (in the case of the negative binomial) or as a specific submodel (in the case of the generalized Poisson), their performance should be comparable to the Poisson for equidispersed data, while providing a distinct advantage for overdispersed data.

Since both distributions arise as a mixture of ordinary Poisson[Joe and Zhu, 2005], they reduce to the Poisson in the case of a degenerate mixture. In the overdispersed (or non-degenerate) setting, the choice between the negative binomial and generalized Poisson base distributions will depend on the structure of the data, with the generalized Poisson typically providing better fit for highly skewed data [Joe and Zhu, 2005].

Our analysis of the DSR data also yielded several important public health findings. Non-private insurance, male gender, and non-Hispanic black race were associated with increased ED use. Compared to non-Hispanic whites, Hispanics were more likely to use the ED at least once, but less inclined to make repeat visits. In all years, block groups in the center of county had the highest rates of ED use while those in the southwest had the lowest. The spatial pattern changed most noticeably between 2007 to 2008 before stabilizing in the following years.

In general, the models developed here are useful for the spatial-temporal analysis of overdispersed count data. The proposed Bayesian approach provides a practical framework for fitting such models.

Table 4.1: Characteristics of DSR Patients ($N = 122273$) and Census block groups ($n = 129$), 2007-2011

Variable	n	%
Male	49719	41
Race		
Non-Hispanic White	56,734	46
Non-Hispanic Black	51,528	42
Hispanic	7,523	6
Asian	3,165	3
Other	3,323	3
Insurance		
Duke Insurance	13,932	11
Other Private Insurance	58,918	48
Medicaid	17,761	15
Medicare	19,493	16
Self	12,169	10
	Median	Range
Age (Years)	38	(1, 103)
Median Household Income ($)	45,330	(5,980, 134,000)
Block Group Sample Size	776	(39, 3,212)

Table 4.2: Model Comparison Results

Base Distribution	Spatial Map	DIC	pD	Posterior Predictive Checks		
				$Pr(Y = 0)$	$E(Y\|Y > 0)$	$V(Y\|Y > 0)$
Poisson	Fixed Map	232542	258	0.709(0.706, 0.713)*	1.93(1.91, 1.95)	1.46(1.41, 1.51)
Poisson	Dynamic Map	232158	566	0.709(0.705, 0.714)	1.93(1.90, 1.95)	1.48(1.43, 1.55)
Negative Binomial	Fixed Map	211014	96	0.710(0.706, 0.713)	1.92(1.90, 1.95)	3.72(3.41, 4.06)
Negative Binomial	Dynamic Map	211198	367	0.709(0.705, 0.713)	1.93(1.89, 1.96)	3.76(3.42, 4.20)
Generalized Poisson	Fixed Map	210983	236	0.709(0.706, 0.713)	1.93(1.89, 1.96)	4.61(4.05, 5.27)
Generalized Poisson	Dynamic Map	211035	367	0.709(0.706, 0.713)	1.93(1.89, 1.97)	4.66(4.10, 5.46)
				Observed: 0.709	Observed: 1.94	Observed: 5.89

*Posterior median and 95% credible interval.

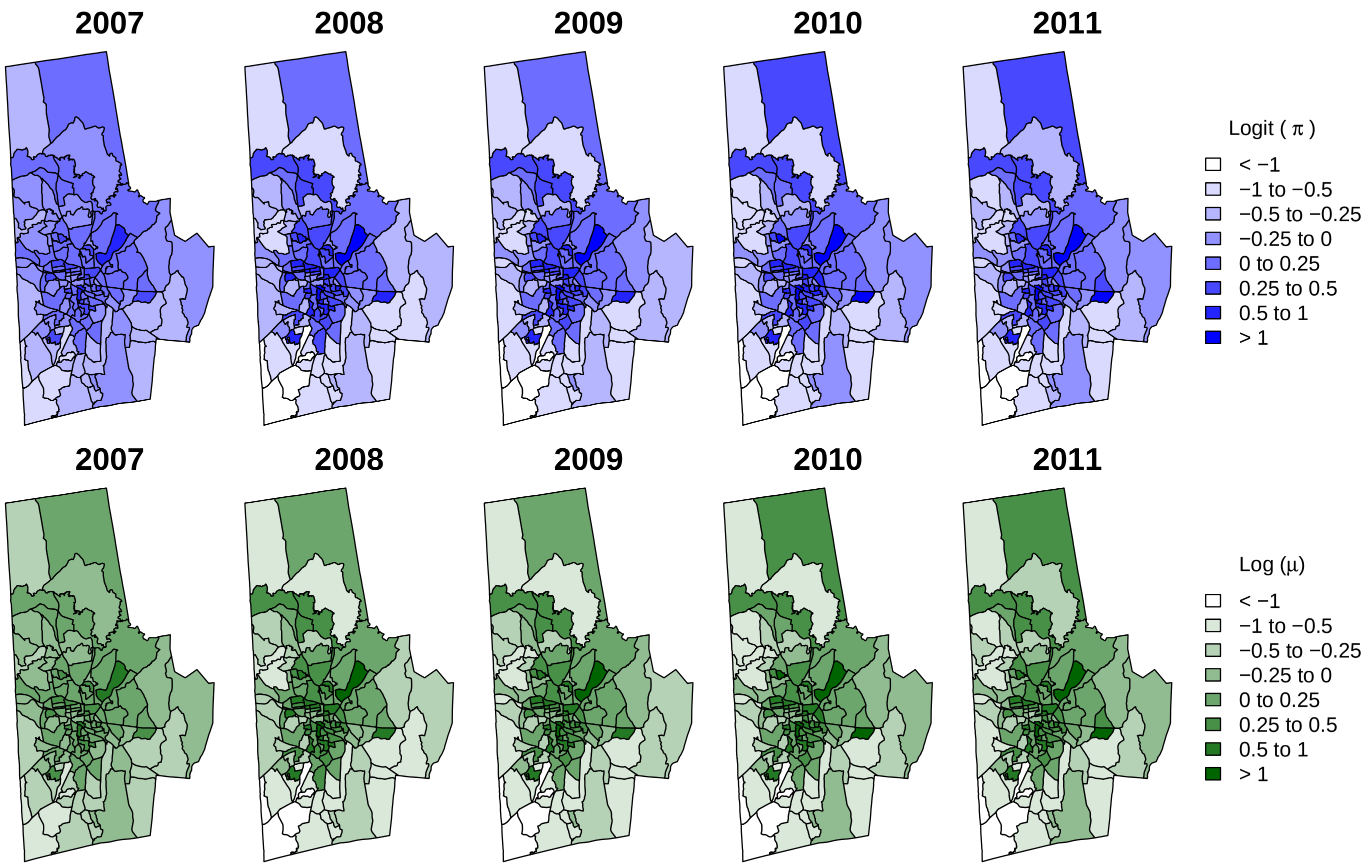


Figure 4.1: Spatial-temporal effects, η_1 and η_2, from the generalized Poisson model

Table 4.3: Posterior summaries for the Poisson, negative binomial and generalized Poisson hurdle models, Part(1)

	Poisson			Negative Binomial			Generalized Poisson		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Logit(π)									
Intercept	-1.206	-1.307	-1.12	-1.204	-1.358	-1.089	-1.209	-1.331	-1.087
Year[1]									
2008	-0.004	-0.049	0.041	-0.005	-0.05	0.041	-0.004	-0.049	0.04
2009	0.004	-0.037	0.050	0.003	-0.038	0.046	0.005	-0.041	0.053
2010	-0.024	-0.071	0.022	-0.025	-0.069	0.018	-0.024	-0.072	0.025
2011	-0.015	-0.058	0.027	-0.015	-0.062	0.032	-0.015	-0.063	0.033
Male	0.186	0.157	0.213	0.185	0.157	0.215	0.186	0.157	0.216
Race[2]									
Black	0.685	0.649	0.721	0.685	0.648	0.725	0.685	0.648	0.721
Hispanic	0.485	0.427	0.545	0.483	0.422	0.543	0.485	0.422	0.545
Asian	-0.424	-0.536	-0.307	-0.424	-0.552	-0.306	-0.427	-0.542	-0.309
Other	0.106	0.015	0.193	0.107	0.014	0.198	0.107	0.016	0.193
Insurance[3]									
Duke	-0.293	-0.349	-0.239	-0.293	-0.35	-0.241	-0.292	-0.346	-0.242
Medicaid	1.191	1.145	1.237	1.191	1.146	1.239	1.193	1.144	1.24
Medicare	0.81	0.748	0.871	0.808	0.751	0.867	0.810	0.743	0.869
Self	1.538	1.492	1.584	1.540	1.494	1.588	1.539	1.490	1.584
Median Income	-0.005	-0.007	-0.003	-0.005	-0.007	-0.003	-0.005	-0.007	-0.003

[1] Reference: Year 2007

[2] Reference: Non-Hispanic White

[3] Reference: Other Private Insurance

Table 4.4: Posterior summaries for the Poisson, negative binomial and generalized Poisson hurdle models, Part(2)

	Poisson			Negative Binomial			Generalized Poisson		
	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Log(μ)									
Intercept	0.311	0.225	0.39	-1.345	-1.679	-1.127	-4.017	-4.695	-3.392
Year[1]									
2008	0.045	0.007	0.084	0.078	0.017	0.145	0.085	0.013	0.159
2009	0.080	0.041	0.12	0.091	0.028	0.158	0.100	0.034	0.171
2010	0.100	0.061	0.142	0.122	0.049	0.190	0.138	0.061	0.215
2011	0.156	0.117	0.195	0.199	0.134	0.265	0.223	0.151	0.296
Male	-0.071	-0.092	-0.049	-0.082	-0.126	-0.037	-0.095	-0.144	-0.052
Race[2]									
Black	0.036	0.005	0.065	0.178	0.120	0.237	0.188	0.130	0.251
Hispanic	-0.606	-0.660	-0.554	-0.524	-0.610	-0.437	-0.611	-0.703	-0.513
Asian	-0.766	-0.962	-0.573	-0.698	-0.969	-0.443	-0.774	-1.061	-0.508
Other	-0.444	-0.539	-0.357	-0.371	-0.526	-0.223	-0.443	-0.609	-0.280
Insurance[3]									
Duke	-0.153	-0.216	-0.088	-0.142	-0.244	-0.032	-0.158	-0.270	-0.041
Medicaid	0.668	0.636	0.696	0.804	0.739	0.864	0.914	0.845	0.982
Medicare	0.739	0.701	0.777	0.768	0.688	0.844	0.885	0.797	0.972
Self	0.413	0.381	0.442	0.510	0.456	0.571	0.575	0.512	0.642
Median Income	-0.004	-0.006	-0.002	-0.003	-0.006	-0.001	-0.004	-0.006	-0.001

[1] Reference: Year 2007

[2] Reference: Non-Hispanic White

[3] Reference: Other Private Insurance

Table 4.5: Posterior summaries for the Poisson, negative binomial and generalized Poisson hurdle models, Part(3)

Spatial CAR Covariance									
Sigma.phi[1,1]	0.216	0.160	0.293	0.216	0.157	0.291	0.215	0.157	0.288
Sigma.phi[1,2]	0.119	0.066	0.191	0.141	0.086	0.211	0.162	0.098	0.243
Sigma.phi[2,2]	0.305	0.218	0.421	0.294	0.204	0.408	0.391	0.261	0.556
ρ_ϕ	0.464	0.289	0.636	0.561	0.374	0.716	0.560	0.372	0.716
Dynamic CAR Covariance									
Sigma.phi[1,1]	0.030	0.021	0.042	0.030	0.021	0.042	0.030	0.021	0.043
Sigma.phi[1,2]	-0.005	-0.026	0.015	-0.001	-0.013	0.011	-0.001	-0.014	0.012
Sigma.phi[2,2]	0.155	0.118	0.201	0.059	0.036	0.093	0.073	0.039	0.115
ρ_ϕ	-0.067	-0.352	0.239	-0.026	-0.279	0.246	-0.023	-0.304	0.246
Dynamic CAR Autoregressive Parameter									
ρ	0.635	0.445	0.782	0.587	0.252	0.838	0.623	0.275	0.883
Overdispersion Parameter									
α	–	–	–	2.747	2.297	3.412	0.009	0.005	0.015

Chapter 5

Summary

5.1 Conclusions

We developed a series of Bayesian hierarchical spatial and spatial-temporal models. Models were developed with increasing complexity, from simple static spatial random effects to dynamic spatial-temporal random effects, from Poisson response models to two component mixture models (zero-inflated and hurdle), from independent two component mixture models to introducing correlation between the two components.

Model comparison results have shown that Colorado cancer surveillance data have little space-time interaction in terms of spatial and temporal random effects. Models with simple static spatial random effects fits the data better than models considering the time-space interaction. However, the number of zero reported counts in those data impose lack-of-fit due to overdispersion if fitting the Poisson model. The generalized Poisson models and the negative binomial models were shown to be better alternatives to the Poisson models by providing better fit to both the cancer data and the Duke Emergency Department visit data.

5.2 Future Work

In this dissertation, we have evaluated modeling the correlation among the components of mixture models (zero-inflated models and hurdle models), as well as correlation among spatial and temporal random effects. One important future direction is to consider dependence among multivariate responses[Tzala and Best, 2008], such as modeling areal count data of multiple cancer types[Downing et al., 2008], or of multiple age groups simultaneously. Gamerman and Moreira [2004] discussed how multivariate regression models can be accomplished in Bayesian framework, while accommodating temporal and spatial variations in the covariate effects. Another advantage of building multivariate response model is the convenience in testing dependence between outcome responses.

For instance, if we assume there are q outcomes, and p independent variables, then we have the structure of the form:

$$\mathbf{A}\mathbf{y} = \mathbf{\Gamma}\mathbf{x} + \boldsymbol{\phi} \quad \text{and} \quad \boldsymbol{\phi} \sim \mathbf{N}\left(\mathbf{0},\ \mathbf{\Delta}\right), \tag{5.1}$$

with $\mathbf{y} = (y_1, \ldots, y_q)^{'}$, and $\mathbf{x} = (x_1, \ldots, x_p)^{'}$. Furthermore, $\mathbf{A}$ is a $q \times q$ matrix with linear relations between the outcomes, which provides a link between the q univariate regressions. $\mathbf{\Gamma}$ is a $q \times p$ matrix of regression coefficients and $\mathbf{\Delta}$ is a diagonal matrix with entries $\delta_1^2, \ldots, \delta_q^2$.

If $\mathbf{A}$ is full rank, the solution of $\mathbf{y}$ above is obtained:

$$\begin{aligned} \mathbf{y} &= \mathbf{A}^{-1}\mathbf{\Gamma}\mathbf{x} + \mathbf{A}^{-1}\boldsymbol{\phi} && (5.2) \\ &= \mathbf{B}\mathbf{x} + \boldsymbol{\epsilon}; && (5.3) \end{aligned}$$

with $\mathbf{B} = \mathbf{A}^{-1}\mathbf{\Gamma}$ and $\boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$.

To incorporate spatial and temporal dependence, Gamerman and Moreira [2004] have shown that 5.2 can be rewritten as:

$$\mathbf{y} = \mathbf{W}\mathbf{y} + \mathbf{B}\mathbf{x} + \boldsymbol{\mu}, \tag{5.4}$$

$$\boldsymbol{\mu} = \mathbf{W}\boldsymbol{\mu} + \mathbf{e}, \tag{5.5}$$

where $\mathbf{e} \sim N(\mathbf{0}, \delta^2\mathbf{I})$ and matrix $\mathbf{W}$ are defined the same as in section (1.2)

For modeling areal count data, such as cancer data or infectious disease counts from small areas, we show that Bayesian multivariate models can improve inference and also the short-term prediction performance when the correlation structure is correctly addressed.

Usually, data from the National Surveillance system, such as US cancer surveillance, are massive and different subtypes of diseases have diverse gender, racial, and age group cancer profiles. Comprehensive national analysis cannot be conducted for evaluating different Bayesian methods in the pilot case studies. Future work can extend the modeling strategy from regional data to national data. Despite the computational effort of Bayesian inference and model complexity, recent advances in Bayesian computation, such as the integrated nested Laplace approximations (INLA) [Rue and Martino, 2007, Schmid and Held, 2004], may further encourage the use of Bayesian space-time model for such large datasets. We believe the same methods can be conveniently extended and applied to model data in a large scale on a national level.

Appendices

Appendix A

Metropolis-Hastings algorithm for estimating the dispersion parameter in the generalized Poisson model.

Let

$$y = log\left(\frac{\alpha - l}{u - \alpha}\right)$$

such that Y could take a normal prior with parameter space covering$(-\infty, +\infty)$. Since $\alpha \leftrightarrow y$ is one-to-one monotonic transformation, we can derive the pdf of Y as:

$$F_Y(y) = Pr\left[log\left(\frac{X - l}{u - X}\right) \leq y\right] = Pr\left[X \leq \frac{l + ue^y}{1 + e^y}\right]$$

by taking derivative, we have

$$f_Y(y) = f_X\left(\frac{l + ue^y}{1 + e^y}\right)\frac{d}{dy}\left(\frac{l + ue^y}{1 + e^y}\right)$$

where since X follows a uniform distribution, $f_X(.) = \frac{1}{u-l}$. So we have

$$f_Y(y) = \frac{1}{u-l} \times \frac{e^y\,(u-l)}{(1+e^y)^2} = \frac{e^y}{(1+e^y)^2}$$

To provide MCMC updates by componentwise Metropolis-Hastings (M-H) algorithm, we generate new values of $\alpha^{'}$ through target distribution of $Y^{'} \sim N(y,\, \sigma^2)$. $f_Y(y)$ will be factored to provide the MCMC acceptance ratio calculation.

Bibliography

D.K. Agarwal, A.E. Gelfand, and S. Citron-Pousty. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9(4):341–355, 2002.

S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical modeling and analysis for spatial data*, volume 101. Chapman & Hall/CRC, 2004.

U.S. Cancer Statistics Working Group CDC. United states cancer statistics: 1999-2007 incidence and mortality web-based report., 2010. URL `http://www.cdc.gov/uscs`.

H. Chang, B. Reich, and M. Miranda. Spatial Time-to-Event Analysis of Air Pollution and Preterm Birt. *Journal of the Royal Statistical Society*, page In Press, 2011.

H.H. Chang, B.J. Reich, and M.L. Miranda. A spatial time-to-event approach for estimating associations between air pollution and preterm birth. *Journal of the Royal Statistical Society: Series C*, 2012. Published online ahead of print.

C DeNavas-Walt, BD Proctor, and JC Smith. Income, poverty, and health insurance coverage in the United States: 2011. September 2012, 2013.

Amy Downing, David Forman, Mark S Gilthorpe, Kimberley L Edwards, and Samuel OM Manda. Joint disease mapping using six cancers in the yorkshire region of england. *International journal of health geographics*, 7(1):41, 2008.

M.P. Fay, R.C. Tiwari, E.J. Feuer, and Z. Zou. Estimating average annual percent change for disease rates without assuming constant change. *Biometrics*, 62(3): 847–854, 2006.

C.M. Frey, E.J. Feuer, and M.J. Timmel. Projection of incidence rates to a larger population using ecologic variables. *Statistics in medicine*, 13(17):1755–1770, 1994.

M. Fuentes, H.R. Song, S.K. Ghosh, D.M. Holland, and J.M. Davis. Spatial association between speciated fine particles and mortality. *Biometrics*, 62(3):855–863, 2006.

Dani Gamerman and Ajax RB Moreira. Multivariate spatial regression models. *Journal of multivariate analysis*, 91(2):262–281, 2004.

Alan E Gelfand, Sudipto Banerjee, and Dani Gamerman. Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, 16(5):465–479, 2005.

A Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, 2 edition, 2004.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

K. Ghosh and R.C. Tiwari. Prediction of US cancer mortality counts using semiparametric Bayesian techniques. *Journal of the American Statistical Association*, 102 (477):7–15, 2007.

S.K. Ghosh, P. Mukhopadhyay, and J.C.J.C. Lu. Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, 136(4):1360–1375, 2006.

Souparno Ghosh, Alan E Gelfand, Kai Zhu, and James S Clark. The k-ZIG: Flexible Modeling for Zero-Inflated Counts. *Biometrics*, February 2012. ISSN 1541-0420. URL `http://www.ncbi.nlm.nih.gov/pubmed/22348816`.

Pierre Goovaerts and Hong Xiao. Geographical, temporal and racial disparities in late-stage prostate cancer incidence across Florida: a multiscale joinpoint regression analysis. *Int J Health Geogr*, 10:63, 2011. doi: 10.1186/1476-072X-10-63.

C.A. Gotway and L.J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

D.B. Hall. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039, 2000.

D.C. Heilbron. Zero-Altered and other Regression Models for Count Data with Added Zeros. *Biometrical Journal*, 36(5):531–547, 1994.

MD Hutton, LD Simpson, DS Miller, HK Weir, K McDavid, and HI Hall. Progress toward nationwide cancer surveillance: an evaluation of the national program of cancer registries, 1994–1999. *Journal Registry Manage*, 28(3):113–120, 2001.

Noriszura Ismail and Abdul Aziz Jemain. *Handling overdispersion with Negative Binomial and Generalized Poisson regression models.* Casualty Actuarial Society, 2007.

Harry Joe and Rong Zhu. Generalized Poisson Distribution: the Property of Mixture of Poisson and Comparison with Negative Binomial Distribution. *Biometrical Journal*, 47(2):219–229, 2005. ISSN 1521-4036. doi: 10.1002/bimj.200410102. URL `http://dx.doi.org/10.1002/bimj.200410102`.

H J Kim, M P Fay, E J Feuer, and D N Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Stat Med*, 19(3):335–51, February 2000.

D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, pages 1–14, 1992.

KV Mardia. Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284, 1988.

M. A. Martínez-Beneito, A. López-Quilez, and Botella-Rocamora. An Autoregressive approach to spatio-temporal disease mapping. *Statist. Med.*, 27:2874–2889, 2008.

Shannon K. McClintock. *Model-Based Statistical Methods for Public Health Surveillance Subject to Imperfect Observations.* PhD thesis, Emory University, Atlanta, GA, March 2012.

D.N. Midthune, M.P. Fay, L.X. Clegg, and E.J. Feuer. Modeling reporting delays and reporting corrections in cancer registry data. *Journal of the American Statistical Association*, 100(469):61–70, 2005.

Y. Min and A. Agresti. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19, 2005.

J Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365, 1986.

M. Mungiole, L.W. Pickle, K.H. Simonson, and AA White. Application of a weighted head-banging algorithm to mortality data maps. *Statistics in Medicine*, 18(23): 3201–3209, 1999.

National Cancer Institute NCI. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 17 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2009 Sub

(1973-2007 varying) - Linked To County Attributes - Total U.S., 1969-2007 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2010, based on the November 2009 submission., 2010.

B. Neelon, P. Ghosh, and P.F. Loebs. A Spatial Poisson Hurdle Model for Exploring Geographic Variation in Emergency Department Visits. *Journal of the Royal Statistical Society, Series A*, 176(389-413), 2013.

B.H. Neelon, A.J. O'Malley, and S.L.T. Normand. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical modelling*, 10(4):421–439, 2010.

R Niska, F Bhuiya, and J Xu. National Hospital Ambulatory Medical Care Survey: 2007 emergency department summary. Technical Report NHS Report no. 26, National Center for Health Statisticsn, Washington, D.C., 2010.

World Health Organization. Cancer Fact Sheets. Web, February 2012. URL `http://www.who.int/mediacentre/factsheets/fs297/en`. Retrieved March 9, 2012.

L. W. Pickle, L. A. Waller, and A. B. Lawson. Current practices in cancer spatial data analysis: a call for guidance. *Int J Health Geogr*, 4(1):3, 2005. Journal article International journal of health geographics Int J Health Geogr. 2005 Jan 13;4(1):3.

L. W. Pickle, Y. Hao, A. Jemal, Z. Zou, R. C. Tiwari, E. Ward, M. Hachey, H. L. Howe, and E. J. Feuer. A new method of estimating United States and state-level cancer incidence counts for the current calendar year. *CA Cancer J Clin*, 57(1): 30–42, 2007.

L.W. Pickle, E.J. Feuer, and B.K. Edwards. US predicted cancer incidence, 1999: Complete maps by county and state from spatial projection models. *National Cancer Institute, Cancer Surveillance Monograph No*, 5, 2003.

H. Rue and S. Martino. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of statistical planning and inference*, 137(10): 3177–3192, 2007.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2004.

V. Schmid and L. Held. Bayesian extrapolation of space-time trends in cancer registry data. *Biometrics*, 60(4):1034–42, 2004.

D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS User Manual, Version 1.4.3*, 2007. URL http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00353. URL http://dx.doi.org/10.1111/1467-9868.00353.

S.D. Stellman, Y. Chen, J.E. Muscat, I.V. Djordjevic, J.P. Richie, P. Lazarus, S. Thompson, N. Altorki, M. Berwick, M.L. Citron, et al. Lung cancer risk in white and black Americans. *Annals of epidemiology*, 13(4):294–302, 2003.

S.B. Thacker and R.L. Berkelman. Public health surveillance in the united states. *Epidemiologic Reviews*, 10(1):164–190, 1988. ISSN 0193-936X.

Ram C Tiwari, Kaushik Ghosh, Ahmedin Jemal, Mark Hachey, Elizabeth Ward, Michael J Thun, and Eric J Feuer. A new method of predicting US and state-level cancer mortality counts for the current calendar year. *CA Cancer J Clin*, 54(1): 30–40, 2004.

E. Tzala and N. Best. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical methods in medical research*, 17 (1):97–118, 2008.

US Census Bureau. *American Community Survey 2005–2009*. Washington, DC, 2010. URL `http://www.census.gov/acs/www`.

J.M. Ver Hoef and J.K. Jansen. Space–time zero-inflated count models of Harbor seals. *Environmetrics*, 18(7):697–712, 2007.

L.A. Waller, B.P. Carlin, H. Xia, and A.E. Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, pages 607–617, 1997.

H Xia and B P Carlin. Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Stat Med*, 17(18):2025–43, September 1998.

H. Zhang. Detecting change points and monitoring biomedical data. *Communications in Statistics-Theory and Methods*, 24(5):1307–1324, 1995.