8-2011

A comparison of spatio-temporal prediction methods of cancer incidence in the U.S

Michelle Hamlyn
University of Nevada, Las Vegas

A COMPARISON OF SPATIO-TEMPORAL PREDICTION METHODS OF CANCER INCIDENCE IN THE U.S.

By

Michelle Hamlyn

A thesis submitted in partial fulfillment

of the requirements for the

Master of Science in Mathematical Sciences

Department of Mathematical Sciences

College of Sciences

The Graduate College

University of Nevada, Las Vegas

August 2011

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend the thesis prepared under our supervision by

Michelle Hamlyn

entitled

A Comparison of Spatio-Temporal Prediction Methods of Cancer Incidence in the U.S.

be accepted in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences
Department of Mathematical Sciences

Kaushik Ghosh, Ph. D., Committee Chair

Sandra Catlin, Ph. D., Committee Member

Anton Westveld, Ph. D., Committee Member

Sheniz Moonie, Ph. D., Graduate College Representative

Ronald Smith, Ph. D., Vice President for Research and Graduate Studies
and Dean of the Graduate College

August 2011

ABSTRACT

A Comparison of Spatio-Temporal Prediction Methods of

Cancer Incidence in the U.S.

by

Michelle Hamlyn

Dr. Kaushik Ghosh, Examination Committee Chair

Assistant Professor of Biostatistics

University of Nevada, Las Vegas

Cancer is the cause of one out of four deaths in the United States, and in 2009, researchers expected over 1.5 million new patients to be diagnosed with some form of cancer. People diagnosed with cancer, whether a common or rare type, need to undergo treatments, the amount and kind of which will depend on the severity of the cancer. So how do healthcare providers know how much funding is needed for treatment? What would better enable a pharmaceutical company to determine how much to allocate for research and development of drugs, the amount of each drug to manufacture, or the time spent to improve or reformulate those drugs? How do government planners determine which cancers need more attention than others? To answer these questions, it becomes extremely important to get accurate predictions

of new cancer cases (also known as cancer incidences) that will occur in the future based on past data.

Past data on cancer incidences in the U.S. is available only at certain cancer registries. These registries did not all come online at the same time, resulting in varying lengths of incidence data. Prediction into the future would require one to account for these varying lengths. Additionally, since these registries do not cover the entire United States, one needs to incorporate some spatial projection methods. In this thesis, we develop a Bayesian spatio-temporal method of predicting future cancer incidences based on past data. A conditional autoregressive prior is used for the spatial component and an autoregressive model is used for the temporal part. We use standard Bayesian Markov chain Monte Carlo techniques to develop predictions four years into the future for individual states. The method is illustrated using incidence data for some rare and common cancers.

ACKNOWLEDGEMENTS

I have learned a great deal from those who have worked with me throughout my time at UNLV, especially Dr. Kaushik Ghosh. I am particularly grateful for the continuous guidance, encouragement and support he has provided. In addition, I would like to thank my committee members, Dr. Sandra Catlin, Dr. Sheniz Moonie and Dr. Anton Westveld for their comments, suggestions and continued assistance.

Finally, I would like to thank my family and friends for their love and encouragement throughout my graduate studies. Special thanks to my husband, Ian, for his patience, understanding and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cancer is the cause of one out of four deaths in the United States, and, in 2009, researchers expected over 1.5 million new patients to be diagnosed with some form of cancer [2]. According to the Centers for Disease Control and Prevention, breast cancer and lung cancer are two of the most common cancers. Based on rates from 2005 to 2007, the National Cancer Institute reported that one in eight women in the U.S. will be diagnosed with breast cancer in their lifetime [4]. In 2010, the American Cancer Society estimated that there would be about 222,520 new cases of lung cancer, the second most common cancer [3]. The rates from 2005 to 2007 lead experts to expect that almost seven percent of men and women will be diagnosed with lung cancer in their lifetime [9]. On the other side of the spectrum, less than half a percent of the population born today will be diagnosed with cancer of the small intestine. Hence the number of cancer cases diagnosed can vary widely.

People diagnosed with cancer, whether an ordinary or rare type, need to undergo

treatments, the amount and kind of which will depend on the severity of the cancer. So how do healthcare providers know how much funding is needed for future treatment? What would better enable a pharmaceutical company to determine how much to allocate for research and development of drugs, the amount of each drug to manufacture, or the time spent to improve or reformulate those drugs? How do government planners determine which cancers need more attention than others?

To answer these questions, it becomes extremely important to get accurate predictions of new cancer cases (also known as cancer incidence) that will occur in the future. Statisticians continue to search for models that will accurately predict future cancer incidences based on historical data and allow healthcare providers to set aside a reasonable amount of funding for research, detection, and treatment of new cases [13]. The main focus of this thesis will be to use observed incidence data to project future counts for the entire U.S.

1.2 Data

Currently, data on cancer incidence is collected at several cancer registries, which are spread across the United States and cover only a small fraction of the total population. The data for this study was obtained from registries affiliated with the National Cancer Institute's Surveillance, Epidemiology and End Results program, henceforth to be called SEER registries. The SEER program has been collecting incidence data since 1973, when it started with seven registries [10]. Over the years, it has grown to its current size of seventeen registries shown in the map in Figure 1.1.

Since not all registries went online at the same time, some registries have thirty years worth of data while others may have only ten.

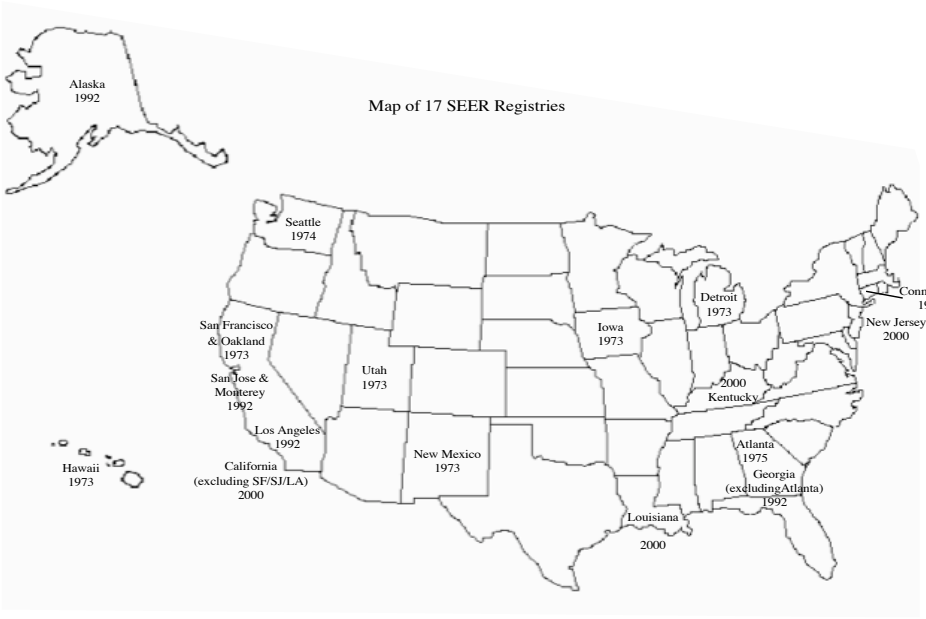

Figure 1.1: U.S. map showing the 17 SEER registries.

Currently, the incidence and survival data published by SEER covers about twenty-eight percent of the U.S. population and includes many demographics, such as race, which includes African American, Hispanic and Asian subgroups. Among the routinely collected data for each incidence, SEER is the only program in the U.S. to include the stage of cancer at the time of diagnosis. Not only does SEER allow access to the data for analysis by different organizations, it is also committed to continually improving methods so that complete and accurate data are collected. The

program's goals, statistics, data and information for registrars can be found on the SEER website, `http://seer.cancer.gov/` [8].

The SEER data used in this study provided incidences for seventeen different registries broken down by cancer type. SEER collects data on new cases as well as other variables such as whether the person was a smoker or had a family history of cancer. For this study we have used only the incidence counts to keep our model simple.

Some registries had collected data for the whole state, while some were focused on specific locations. For example, California had four different registries: one for the Los Angeles area, one for the San Jose area, one for San Francisco and Oakland and the last one collected data for the remaining part of California. Unfortunately, these registries were not all online at the same time, so we were unable to get a clear picture on total California incidences. To simplify our model, we decided to study only states that had a single registry collecting data anytime between 1973 and 2003. Those nine states were Connecticut, Hawaii, Iowa, New Mexico, Utah, Alaska, Kentucky, Louisiana and New Jersey. Although this may seem like a small sample, these registries provide valuable information on cancer incidences across the country.

For this study, we have chosen to focus and test our model on two common cancers – breast and lung cancer, and one rare type – cancer of the small intestine. The original data can be seen in Tables 5.1 - 5.3 in the Appendix. As can be seen in Figure 1.2 and Figure 1.3, the number of new breast and lung cancer cases has for the most part been increasing over the years, although a slight decline can be seen after 2000. Three of the four states from the west side of the country – New Mexico,

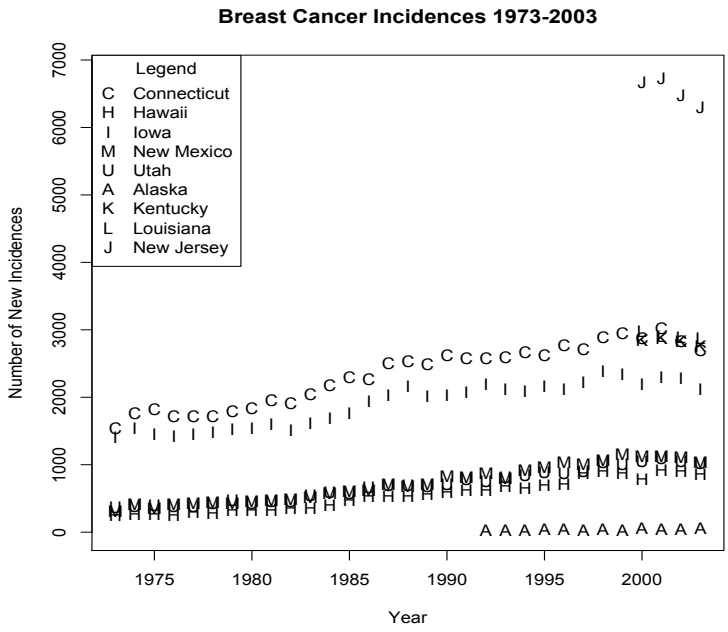

Figure 1.2: SEER data for breast cancer incidence from 1973 to 2003.

Utah and Hawaii – have close to the same number of new cases each year. However, based on Figure 1.2 and Figure 1.3, we could not conclude that region alone drives the number of new incidences, since New Jersey has a significantly higher number of incidences even though it is in closer proximity to Connecticut than Hawaii is to Utah or New Mexico.

In Figure 1.4, the number of new small intestine cancer cases show an overall increase, but in general from year to year, seems to fluctuate between increasing and decreasing more than the breast and lung cancer incidences. As we saw with the other cancer types, the incidences for cancer of the small intestine in New Jersey are also significantly higher than those in the other regions.

As can be seen in all figures, not all states had observed incidences every year.

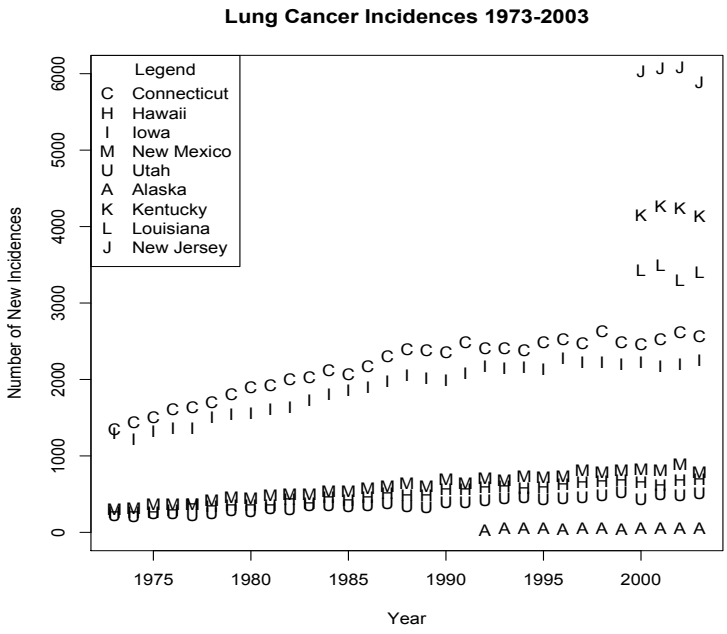

Figure 1.3: SEER data for lung cancer incidence from 1973 to 2003.

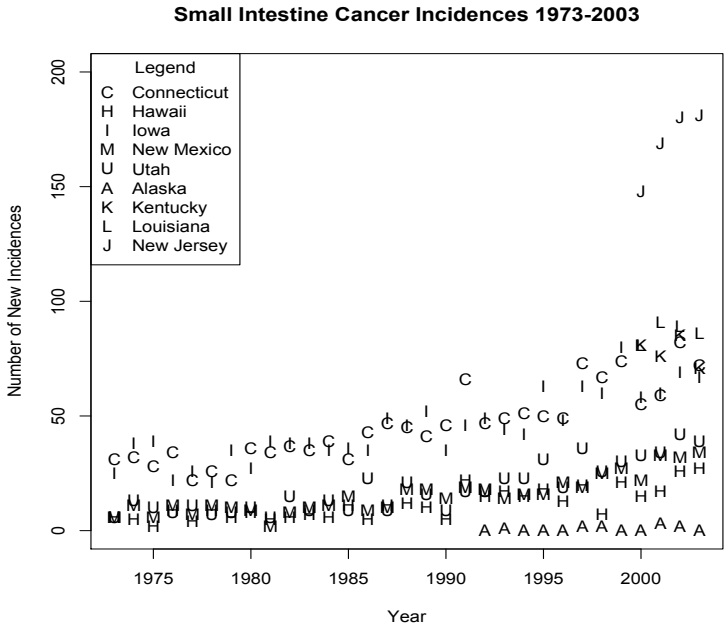

Figure 1.4: SEER data for small intestine cancer incidence from 1973 to 2003.

Due to the incomplete and irregular nature of the past data, incidence prediction presents several challenges. First, in any particular year, incidence data are available only at the SEER registries, leaving counts outside those locations unobserved. This requires some sort of spatial projection to "fill-in" the unobserved sites. Next, the data need to be temporally projected. Some feel that any incomplete data should just be removed. However, this can lead to a loss of power and in our case would remove almost twenty years worth of data [13]. Therefore, we plan to research and investigate methods for projecting future incidence counts in the SEER as well as the unobserved non-SEER regions and then compare them to methods currently being used by the American Cancer Society (ACS).

1.3 Past Work

Currently, ACS is responsible for incidence projection using a two step method as follows. First, for any year, it spatially extrapolates the incidence counts to non-SEER regions. To estimate the incidences for every state in that particular year, ACS assumes that the number of new cases in county i and age-group j has a Poisson distribution with the intensity $\lambda_{i,j}$ having the following log-linear structure

$$\ln(\lambda_{ij}|\alpha,\beta,\gamma,\delta,\zeta) = \alpha_r + f(a_j)\beta + \ln(m_{ij})\gamma + X_i'\delta + Y_i'\zeta,$$

where α_r is the intercept for region r ($r = 1, 2, 3, 4$) where county i is located, and a_j is the centered midpoint for age group j. To accommodate potential downturns in cancer rates among older patients, $f(a_j)$ was taken to be a cubic function of age a_j. The mortality rate is represented by m_{ij} and the vectors X_i and Y_i represent demographic

covariates and lifestyle covariates respectively [11]. Next, the output from the spatial projection is used in a temporal model that projects incidence counts four years into the future. Based on a study that compared four different types of temporal methods, ACS determined that a piecewise linear regression method, also known as a joinpoint method, was most accurate [10]. For observations $(x_1, y_1), \cdots, (x_n, y_n)$ with $x_1 \leq \cdots \leq x_n$, the general joinpoint model can be written as [6]

$$E(y \mid x) = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \cdots + \delta_k (x - \tau_k)^+,$$

where τ_k's are unknown joinpoints and

$$a_+ = \begin{cases} a & \text{when } a > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Although the ACS determined it to be the best method for projecting new cases in SEER and non-SEER regions, there are several drawbacks to this method, the biggest one being that one must use two separate methods to model the spatial and temporal components. As a result, it is difficult to provide accurate measures of overall uncertainty for the projected counts.

Therefore, our primary goal in this project will be to combine the spatial and temporal estimation into one single model and study the effectiveness of the proposed model. For simplicity, we will only use observed incidence counts as inputs to our model and ignore other covariates such as mortality counts, average income, etc.

In Chapter 2, we propose a temporal prediction model to be fitted using Bayesian techniques and apply it to the three cancers in Chapter 3. A spatial improvement is introduced and tested in Chapter 4. Finally, we compare our predictions for the

year 2006 to some of the predictions published by the ACS and also discuss further research.

CHAPTER 2

A TEMPORAL PROJECTION MODEL

2.1 Introduction

As previously mentioned, SEER started in 1973 with seven registries and over the years has grown to its current size of seventeen registries. Not all affiliated registries came online at the same time, giving rise to an incomplete data problem. Below we propose a model that is able to accommodate vectors of different data lengths and use it to generate predictions.

Consider a specific cancer. Let $Y_{i,t}$ be the incidence count in state i at time t and $y_{i,t}$ be its observed counterpart. We assume that the time series of incidence counts for each state is driven by a common underlying time series x_t with state-specific multipliers of θ_i. Let $\mathbf{Y}_i = (y_{i,1}, \cdots, y_{i,n})$, $\mathbf{X} = (x_1, \cdots, x_n)$ and σ^2 be the variance. We assume $Y_{i,t} \sim N(\theta_i x_t, \sigma^2)$. For the five states with data for all thirty-one years, (Connecticut, Hawaii, Iowa, New Mexico, and Utah) the likelihood contribution for state i is

$$f(\mathbf{Y_i}|\mathbf{X}, \theta_i, \sigma^2) = f(y_{i,1}|\theta_i, x_1, \sigma^2) \cdots f(y_{i,n}|\theta_i, x_n, \sigma^2),$$

where $i = 1, 2, 3, 4, 5$ and $n = 31$. Alaska did not start recording incidences until 1992, so its conditional distribution only had twelve years of data and therefore its likelihood contribution is of the form

$$f(\mathbf{Y_6}|\mathbf{X}, \theta_6, \sigma^2) = f(y_{6,t_0+1}|\theta_6, x_{t_0+1}, \sigma^2) \cdots f(y_{6,n}|\theta_6, x_n, \sigma^2),$$

where $t_0 = 19$. The remaining three states, Kentucky, Louisiana and New Jersey only began collecting data in 2000 so they could only contribute four years of data and in general their likelihood contribution looks like

$$f(\mathbf{Y_i}|\mathbf{X}, \theta_i, \sigma^2) = f(y_{i,t_0+9}|\theta_i, x_{t_0+9}, \sigma^2) \cdots f(y_{i,n}|\theta_i, x_n, \sigma^2),$$

for $i = 7, 8, 9$.

We now focus on the prior distribution of the model parameters, $\mathbf{X}, \theta_1, \cdots, \theta_9$ and σ^2. As a first pass, we assume that for $t = 1, 2, \cdots, n$,

$$X_t \sim N(x_{t-1}, \tau^2),$$

so we expect the conditional on the value at the current year to be the same as the current value. The joint distribution of $\mathbf{X}$ then looks like

$$f(\mathbf{X}|x_0, \tau^2) = f(x_1|x_0, \tau^2) f(x_2|x_1, \tau^2) \cdots f(x_n|x_{n-1}, \tau^2).$$

We also assume that the state-specific factors θ_i are independent of each other as well as the underlying process $\mathbf{X}$. Hence,

$$f(\mathbf{X}, \theta_1, \cdots, \theta_9, \sigma^2|x_0, \tau^2) = f(\mathbf{X}|x_0, \tau^2) f(\theta_1) \cdots f(\theta_9) f(\sigma^2).$$

Then the joint distribution of all the variables is

$$f(\mathbf{Y}, \mathbf{X}, \theta_1, \cdots, \theta_9, \sigma^2, x_0, \tau^2)$$

$$= f(\mathbf{Y}|\mathbf{X}, \theta_1, \cdots, \theta_9, \sigma^2) f(\mathbf{X}, \theta_1, \cdots, \theta_9, \sigma^2 | x_0, \tau^2) f(x_0, \tau^2).$$

The joint distribution can be used to obtain the posterior distribution of the model parameters. Since there is no explicit closed-form expression, a Markov chain Monte Carlo approach will be used to sample from the posterior distribution. In particular, a Gibbs sampler will be used, as the univariate posterior conditionals have simple closed-form expressions. In order to calculate the full conditional distributions, the prior information had to first be identified and are as follows:

$$\sigma^2 \sim \text{IG}(a_1, b_1),$$

where a_1 and b_1 are known,

$$\tau^2 \sim \text{IG}(a_2, b_2),$$

where a_2 and b_2 are known,

$$X_0 \sim \text{N}(\mu_0, w^2),$$

with μ_0 and w^2 both known, and θ_i has a non-informative prior

$$f(\theta_i) = \text{constant}.$$

Here we use N to denote the normal distribution and IG to denote the Inverse Gamma distribution which has a pdf of

$$f(x|a, b) = \frac{(1/b)^a}{\Gamma(a)} y^{-a-1} \exp\left\{-\frac{1/b}{y}\right\},$$

for $y > 0$.

2.2 Posterior Sampling

The univariate conditional posterior distributions for each parameter can be obtained from the joint distribution in the previous section. Beginning with $\mathbf{X}$, which allows us to account for the time dependency for new incidences, the posterior conditional distribution for X_t where $t = (1, \cdots, n-1)$ is obtained as

$$f(x_t|\cdots) \propto$$

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_{i,t}-\theta_i x_t)^2\right\}\exp\left\{-\frac{1}{2\tau^2}(x_{t+1}-x_t)^2\right\}\exp\left\{-\frac{1}{2\tau^2}(x_t-x_{t-1})^2\right\},$$

with m equal to the total number of states and $f(x_t|\cdots)$ denoting the distribution of X_t given all other quantities. That is,

$$X_t|\cdots\sim \mathrm{N}\left(\frac{\frac{\sum_{i=1}^{m} y_{i,t}\theta_i}{\sigma^2}+\frac{x_{t+1}+x_{t-1}}{\tau^2}}{\frac{\sum_{i=1}^{m}\theta_i^2}{\sigma^2}+\frac{2}{\tau^2}}, \left(\frac{\sum_{i=1}^{m}\theta_i^2}{\sigma^2}+\frac{2}{\tau^2}\right)^{-1}\right).$$

Similarly the distribution of X_n is

$$X_n|\cdots\sim \mathrm{N}\left(\frac{\frac{\sum_{i=1}^{9} y_{i,n}\theta_i}{\sigma^2}+\frac{x_{n-1}}{\tau^2}}{\frac{\sum_{i=1}^{9}\theta_i^2}{\sigma^2}+\frac{1}{\tau^2}}, \left(\frac{\sum_{i=1}^{9}\theta_i^2}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\right).$$

The posterior conditional distribution for X_0 is

$$X_0|\cdots\sim \mathrm{N}\left(\frac{\frac{x_1}{\tau^2}+\frac{\mu_0}{w^2}}{\frac{1}{\tau^2}+\frac{1}{w^2}}, \left(\frac{1}{\tau^2}+\frac{1}{w^2}\right)^{-1}\right).$$

For θ_i $(i = 1, 2, 3, 4, 5)$ we have

$$\theta_i|\cdots\sim \mathrm{N}\left(\frac{\sum_{t=1}^{n} y_{i,t}x_t}{\sum_{t=1}^{n} x_t^2}, \frac{\sigma^2}{\sum_{t=1}^{n} x_t^2}\right).$$

As previously mentioned, four states had less than thirty-one years of observed data so the posterior reflected only the years for which data was collected. For θ_6 (corresponding to Alaska) which had data for the years $t_0+1, t_0+2, \cdots, n$, the posterior

had the following normal distribution

$$\theta_6|\cdots\sim \mathrm{N}\left(\frac{\sum_{t=t_0+1}^{n} y_{6,t}x_t}{\sum_{t=t_0+1}^{n} x_t^2}, \frac{\sigma^2}{\sum_{t=t_0+1}^{n} x_t^2}\right).$$

Only four years worth of data was available for Kentucky, Louisiana, and New Jersey corresponding to θ_7, θ_8 and θ_9 but they still had similar distributions which in general can be shown to be

$$\theta_i|\cdots\sim \mathrm{N}\left(\frac{\sum_{t=t_0+9}^{n} y_{i,t}x_t}{\sum_{t=t_0+9}^{n} x_t^2}, \frac{\sigma^2}{\sum_{t=t_0+9}^{n} x_t^2}\right),$$

where $i = 7, 8, 9$. Finally, for the variances, σ^2 and τ^2, the posteriors had Inverse Gamma distributions. For σ^2, we have

$$\sigma^2|\cdots\sim IG(a_1^*, b_1^*),$$

where

$$a_1^* = \frac{9n - 4t_0 + 24}{2} + a_1,$$

and

$$b_1^* = \left\{\frac{1}{2}\sum_{i=1}^{5}\sum_{j=1}^{n}(y_{i,j} - \theta_i x_j)^2 + \frac{1}{2}\sum_{i=t_0+1}^{n}(y_{6,j} - \theta_6 x_j)^2 + \frac{1}{2}\sum_{i=t_0+9}^{n}\sum_{j=7}^{9}(y_{i,j} - \theta_i x_j)^2 + \frac{1}{b_1}\right\}^{-1}$$

and for τ^2 we have ,

$$\tau^2|\cdots\sim \mathrm{IG}\left(\frac{n}{2} + a_2, \left\{\frac{1}{b_2} + \frac{\sum(x_i - x_{i-1})}{2}\right\}^{-1}\right).$$

The Gibbs sampler samples each parameter from its conditional posterior distribution and generates a dependent sequence that eventually converges to the joint posterior distribution of interest. For example, if at the s^{th} iteration, the parameter

values are given by ($\mathbf{x}^{(s)}$, ${\sigma^{(s)}}^2$, ${\tau^{(s)}}^2$, $\boldsymbol{\theta}^{(s)}$), the next iteration of parameter values will be obtained by generating $\boldsymbol{\theta}^{(s+1)}$ from

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}^{(s)}, {\sigma^{(s)}}^2, {\tau^{(s)}}^2),$$

$\mathbf{X}^{(s+1)}$ from

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(s+1)}, {\sigma^{(s)}}^2, {\tau^{(s)}}^2),$$

${\sigma^{(s+1)}}^2$ from

$$p(\sigma^2|\mathbf{y}, \boldsymbol{\theta}^{(s+1)}, \mathbf{x}^{(s+1)}, {\tau^{(s)}}^2),$$

and ${\tau^{(s+1)}}^2$ from

$$p(\tau^2|\mathbf{y}, \boldsymbol{\theta}^{(s+1)}, \mathbf{x}^{(s+1)}, {\sigma^{(s+1)}}^2).$$

This process is repeated for each parameter until the Gibbs sampler converges.

2.3 Predictions

Once the Gibbs sampler has converged, predictions into the future could be made based on samples drawn from the posterior. Suppose we want to predict new incidences for Connecticut in 2004, based on data from 1973 to 2003. Then we are essentially looking for a 1-year-ahead prediction. The prediction will be given by the mean of the one-year-ahead predictive density. In general, if $\mathbf{Y}$ represents all available prior year data, we will denote the one-year-ahead prediction for region i by $\hat{Y}_{i,n+1}$

which will be calculated as

$$
\begin{aligned}
E(Y_{i,n+1}|\mathbf{Y}) &= \int y_{i,n+1} f(y_{i,n+1}|\mathbf{y}) dy_{i,n+1} \\
&= \int\int\int y_{i,n+1} f(y_{i,n+1}|\theta_i, x_{n+1}, \mathbf{y}) f(\theta_i, x_{n+1}|\mathbf{y}) dy_{i,n+1} d\theta_i dx_{n+1} \quad (2.1) \\
&= \int\int \theta_i x_{n+1} f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} \quad (2.2) \\
&= \int\int \theta_i x_{n+1} f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} \\
&= \int\int\int \theta_i x_{n+1} f(\theta_i, x_{n+1}, x_n|\mathbf{y}) d\theta_i dx_{n+1} dx_n \\
&= \int\int\int \theta_i x_{n+1} f(x_{n+1}|x_n, \theta_i, \mathbf{y}) f(\theta_i, x_n|\mathbf{y}) d\theta_i dx_{n+1} dx_n \\
&= \int\int \theta_i x_n f(\theta_i, x_n|\mathbf{y}) d\theta_i dx_n \\
&= E(\theta_i X_n|\mathbf{y}),
\end{aligned}
$$

where we move from Step (2.1) to Step (2.2) using

$$Y_{i,n+1}|\theta_i, X_{n+1} \sim N(\theta_i x_{n+1}, \sigma^2).$$

Then $E(\theta_i X_n|\mathbf{y})$ can be approximated by

$$\frac{1}{M}\sum_{m=1}^{M} \theta_i^{(m)} x_n^{(m)},$$

where M is equal to the total number of iterations of the Gibbs sampler.

Using the same principle, the two-year-ahead prediction for state i, denoted by

$\hat{Y}_{i,n+2}$ will be calculated as

$$\begin{aligned}
\hat{Y}_{i,n+2} &= E(Y_{i,n+2}|\mathbf{Y}) \\
&= \int y_{i,n+2} f(y_{i,n+2}|\mathbf{y}) dy_{i,n+2} \\
&= \int\int\int y_{i,n+2} f(y_{i,n+2}|\theta_i, x_{n+2}, \mathbf{y}) f(\theta_i, x_{n+2}|\mathbf{y}) dy_{i,n+2} d\theta_i dx_{n+2} \\
&= \int\int \theta_i x_{n+2} f(\theta_i, x_{n+2}|\mathbf{y}) d\theta_i dx_{n+2} \\
&= \int\int\int \theta_i x_{n+2} f(\theta_i, x_{n+2}, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+2} dx_{n+1} \\
&= \int\int\int \theta_i x_{n+2} f(x_{n+2}|x_{n+1}, \theta_i, \mathbf{y}) f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+2} dx_{n+1} \\
&= \int\int \theta_i x_{n+1} f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} \\
&= \int\int\int \theta_i x_{n+1} f(\theta_i, x_{n+1}, x_n|\mathbf{y}) d\theta_i dx_{n+1} dx_n \\
&= \int\int\int \theta_i x_{n+1} f(x_{n+1}|x_n, \theta_i, \mathbf{y}) f(\theta_i, x_n|\mathbf{y}) d\theta_i dx_{n+1} dx_n \\
&= \int\int \theta_i x_n f(\theta_i, x_n|\mathbf{y}) d\theta_i dx_n \\
&= E(\theta_i X_n|\mathbf{y}),
\end{aligned}$$

which is the same as the one-year-ahead prediction. Proceeding similarly, it can be shown that $\hat{Y}_{i,n+k} = E(Y_{i,n+k}|\mathbf{Y}) = E(\theta_i X_n|\mathbf{y})$ for $k = 2, 3, \cdots$. That is, the k-year-ahead prediction is the same for all k.

However the variance for each year will show how the uncertainty changes as predictions are made further away from the last year observed. Consider the variance for a one-year-ahead prediction for region i given by,

$$Var(Y_{i,n+1}|\mathbf{Y}) = E(Y_{i,n+1}^2|\mathbf{Y}) - \{E(Y_{in+1}|\mathbf{Y})\}^2.$$

First, we can reduce $E(Y^2_{i,n+1}|\mathbf{y})$ using the definition and integrating

$$\begin{aligned} E(Y^2_{i,n+1}|\mathbf{y}) &= \int y^2_{i,n+1} f(y_{i,n+1}) dy_{i,n+1} \\ &= \int\int\int y^2_{i,n+1} f(y_{i,n+1}|\theta_i, x_{n+1}, \mathbf{y}) f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} dy_{i,n+1} \\ &= \int\int\int y^2_{i,n+1} f(y_{i,n+1}|\theta_i, x_{n+1}) f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} dy_{i,n+1}. \end{aligned} \tag{2.3}$$

Now

$$\begin{aligned} &\int y^2_{i,n+1} f(y_{i,n+1}|\theta_i, x_{n+1}) dy_{i,n+1} \\ &= Var(y_{i,n+1}|\theta_i, x_{n+1}) + \{E(y_{i,n+1}|\theta_i, x_{n+1})\}^2 \\ &= \sigma^2 + (\theta_i x_{n+1})^2, \end{aligned}$$

since $Y_{i,n+1}|\theta_i, x_{n+1} \sim N(\theta_i x_{n+1}, \tau^2)$.

Then (2.3) becomes

$$\begin{aligned} &\int\int\int y^2_{i,n+1} f(y_{i,n+1}|\theta_i, x_{n+1}, \mathbf{y}) f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} dy_{i,n+1} \\ &= \int\int (\sigma^2 + \theta_i^2 x^2_{n+1}) f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} \\ &= \sigma^2 + \int\int \theta_i^2 x^2_{n+1} f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} \\ &= \sigma^2 + \int\int \theta_i^2 x^2_{n+1} f(x_{n+1}|\theta_i, \mathbf{y}) f(\theta_i|\mathbf{y}) dx_{n+1} d\theta_i \\ &= \sigma^2 + \int\int\int \theta_i^2 x^2_{n+1} f(x_{n+1}|\theta_i, x_n, \mathbf{y}) f(x_n|\theta_i, \mathbf{y}) f(\theta_i|\mathbf{y}) dx_{n+1} dx_n d\theta_i \\ &= \sigma^2 + \int\int\int \theta_i^2 x^2_{n+1} f(x_{n+1}|x_n) f(x_n|\theta_i, \mathbf{y}) f(\theta_i|\mathbf{y}) dx_{n+1} dx_n d\theta_i. \end{aligned}$$

We know by definition that

$$\int x^2_{n+1} f(x_{n+1}|x_n) dx_{n+1} = E(X^2_{n+1}|x_n),$$

which can be broken down into

$$Var(X_{n+1}|x_n) + \{E(X_{n+1}|x_n)\}^2,$$

which is equal to

$$\tau^2 + x_n^2,$$

since

$$X_{n+1}|X_n \sim N(X_n, \tau^2).$$

Therefore,

$$\begin{aligned}
&\sigma^2 + \int\int\int \theta_i^2 x_{n+1}^2 f(x_{n+1}|x_n) f(x_n|\theta_i, \mathbf{y}) f(\theta_i|\mathbf{y}) dx_{n+1} dx_n d\theta_i \\
&= \sigma^2 + \int\int \theta_i^2 (\tau^2 + x_n^2) f(x_n|\theta_i, \mathbf{y}) f(\theta_i|\mathbf{y}) dx_n d\theta_i \\
&= \sigma^2 + \tau^2 \int\int \theta_i^2 f(x_n|\theta_i, \mathbf{y}) f(\theta_i|\mathbf{y}) dx_n d\theta_i + \int\int \theta_i^2 x_n^2 f(x_n|\theta_i, \mathbf{y}) f(\theta_i|\mathbf{y}) dx_n d\theta_i \\
&= \sigma^2 + \tau^2 E(\theta_i^2|\mathbf{y}) + E(\theta_i^2 x_n^2|\mathbf{y}).
\end{aligned}$$

Next,

$$\begin{aligned}
\{E(Y_{i,n+1}|\mathbf{y})\} &= \int y_{i,n+1} f(y_{i,n+1}|\mathbf{y}) dy_{i,n+1} \\
&= \int\int\int y_{i,n+1} f(y_{i,n+1}|\theta_i, x_{n+1}, \mathbf{y}) f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} dy_{i,n+1}. \quad (2.4)
\end{aligned}$$

Since $Y_{i,n+1}|\theta_i, x_{n+1} \sim N(\theta_i x_{n+1}, \sigma^2)$, we have Equation (2.4) equal to,

$$\begin{aligned}
&\int\int\int y_{i,n+1} f(y_{i,n+1}|\theta_i, x_{n+1}, \mathbf{y}) f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} dy_{i,n+1} \\
&= \int\int \theta_i x_{n+1} f(\theta_i, x_{n+1}|\mathbf{y}) d\theta_i dx_{n+1} \\
&= \int\int\int \theta_i x_{n+1} f(\theta_i, x_{n+1}, x_n|\mathbf{y}) d\theta_i dx_{n+1} dx_n \\
&= \int\int\int \theta_i x_{n+1} f(x_{n+1}|x_n, \theta_i, \mathbf{y}) f(x_n, \theta_i|\mathbf{y}) d\theta_i dx_{n+1} dx_n.
\end{aligned}$$

Since $X_{n+1}|x_n \sim N(x_n, \tau^2)$, we have

$$\begin{aligned}
&\int\int\int \theta_i x_{n+1} f(x_{n+1}|x_n, \theta_i, \mathbf{y}) f(x_n, \theta_i|\mathbf{y}) d\theta_i dx_{n+1} dx_n \\
&= \int\int \theta_i x_n f(x_n, \theta_i|\mathbf{y}) d\theta_i dx_n,
\end{aligned}$$

which by definition is equal to $E(\theta_i X_n|\mathbf{y})$. Therefore,

$$Var(Y_{i,n+1}|\mathbf{y}) = \sigma^2 + \tau^2 E(\theta_i^2|\mathbf{y}) + E(\theta_i^2 X_n^2|\mathbf{y}) - \{E(\theta_i X_n|\mathbf{y})\}^2,$$

which can be approximated by

$$\frac{1}{M}\sum_{m=1}^{M} \sigma^{(m)^2} + \frac{1}{M}\sum_{m=1}^{M} \tau^{(m)^2}\theta_i^{(m)^2} + \frac{1}{M}\sum_{m=1}^{M} \theta_i^{(m)^2} x_n^{(m)^2} - \left(\frac{1}{M}\sum_{m=1}^{M} \theta_i^{(m)} x_n^{(m)}\right)^2.$$

In a similar manner, we can obtain expressions for the variances for the two-year-ahead and three-year-ahead predictions. In general, for k-year-ahead predictions, we will have

$$Var(Y_{i,n+k}|\mathbf{Y}) = \sigma^2 + k\tau^2 E(\theta_i^2|\mathbf{y}) + E(\theta_i^2 X_n^2|\mathbf{y}) - \{E(\theta_i X_n|\mathbf{y})\}^2,$$

which can be approximated by

$$\frac{1}{M}\sum_{m=1}^{M} \sigma^{2(m)} + k(\frac{1}{M}\sum_{m=1}^{M} \tau^{2(m)}\theta_i^{(m)^2}) + \frac{1}{M}\sum_{m=1}^{M} \theta_i^{(m)^2} x_n^{(m)^2} - (\frac{1}{M}\sum_{m=1}^{M} \theta_i^{(m)} x_n^{(m)})^2.$$

Since the parameters and their distributions have all been identified, we can now run our model and begin making predictions.

CHAPTER 3

TESTING THE MODEL

3.1 Predictions

We applied the model developed earlier to the three data sets. First, we ran the Gibbs sampler in R to get samples from posterior distributions and then used the posterior samples to approximate the expected values found in Chapter 2 to make predictions for 2001 to 2003. The predictions were then compared to the observed values by finding a quantile interval for the prediction. Observed values that fell within the interval were considered "good" predictions.

We used the following prior distributions for model parameters as outlined in Chapter 2:

$$\sigma^2 \sim \text{IG}\left(2.01, \frac{1}{1.01}\right),$$

and

$$\tau^2 \sim \text{IG}\left(2.01, \frac{1}{1.01}\right).$$

These choices reflect uncertainty in the prior information and were guided by the fact

that if $X \sim IG(a, b)$ we have

$$E(X) = \frac{1}{b(a-1)},$$

and

$$Var(X) = \frac{1}{\{b(a-1)\}^2(a-2)},$$

when $a > 2$. We also chose

$$X_0 \sim \mathrm{N}(0, 5000),$$

where the large variance was choosen to make up for the uncertainty of X_0. State-specific θ_i–values were chosen to have a non-informative prior

$$f(\theta_i) = \text{constant},$$

again to reflect uncertainty in prior information.

We used the data from each state to get starting values for $\boldsymbol{\theta}$ and $\mathbf{X}$. Since $\mathbf{X}$ represents year-to-year dependency over all the states, the starting value for $\mathbf{X}$ was estimated at the mean number of incidences at each year. Then $\boldsymbol{\theta}$ was estimated using the mean number of incidences per state divided by the mean of the x_t starting values. The starting values for τ^2, σ^2 and X_0 were randomly drawn from the prior distributions mentioned previously.

The Gibbs sampler was coded in R. It was run for 25,000 iterations and the first 500 iterations were discarded since the sampler had not yet converged. Convergence was verified using traceplots where the parameter values were plotted against the iteration number for each parameter σ^2 and τ^2. When looking at the traceplot of θ_i and x_t they show non-convergence due to lack of identifiability but a plot of $\theta_i x_t$ for

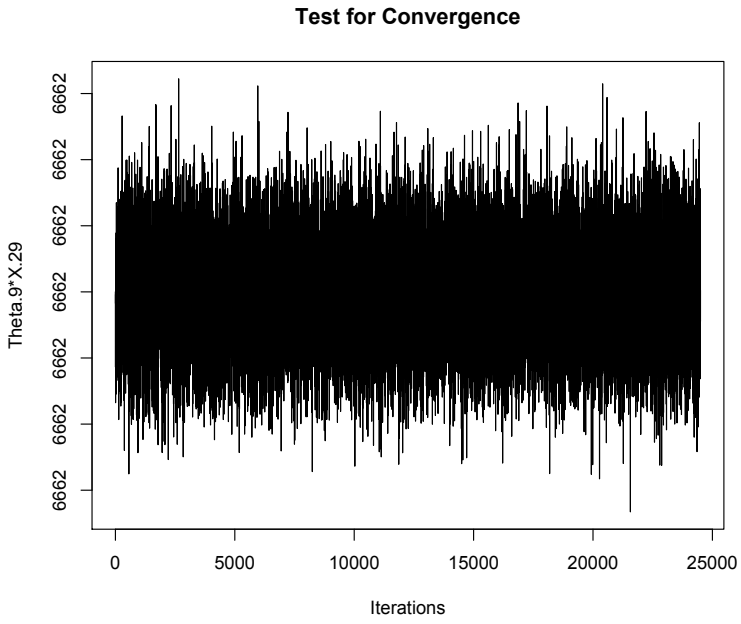

Figure 3.1: Convergence check for Gibbs sampler to predict 2001-2003.

$i = 1, 2, \cdots, 9$ and $t = 1, 2, \cdots, 28$ showed convergence. Figure 3.1 shows an example of the traceplots when the sampler converged. One should not be able to identify a pattern; that is we expect to see a line jumping back and forth but no obvious increase or decrease.

The model was then tested using lung, breast and small intestine cancer data. We used our model to predict for 2001 based on data from 1973 to 2000 and then compared those predictions to the observed counts. As can be seen in Table 3.1, for lung cancer, the model seems to be most often under-predicting. To determine if the difference between the prediction and the actual count is significant, we should look at a prediction interval.

Based on the variances discussed in Chapter 2 the approximate 95% prediction

Table 3.1: 2001 Observed and predicted lung cancer incidences for 2001 using data up to 2000.

State	Prediction	Observed	Difference
CT	2539	2529	10
HI	576	611	-35
IA	2206	2174	32
NM	700	813	-113
UT	427	503	-76
AK	46	53	-7
KY	4142	4263	-121
LA	3422	3494	-72
NJ	6024	6065	-41

intervals $\hat{Y}_{i,n+1} \pm 1.96\sqrt{Var(Y_{i,n+1}|Y)}$ should also show how the uncertainty changes as we predict further ahead. However, as the samples of σ^2 were very large as shown in Figure 3.2, the prediction intervals constructed in this fashion were not practically useful.

Therefore it was decided to calculate the 2.5% and 97.5% quantiles for each prediction since this would give us similar results as using an interval two standard deviations from the mean. Table 3.2, Table 3.3 and Table 3.4 show the predicted value along with the 95% interval and the observed incidences for 2001, for three different cancers. As we can see, the observed incidences fall within the interval in over half the states for each cancer type. The observed incidences for New Mexico

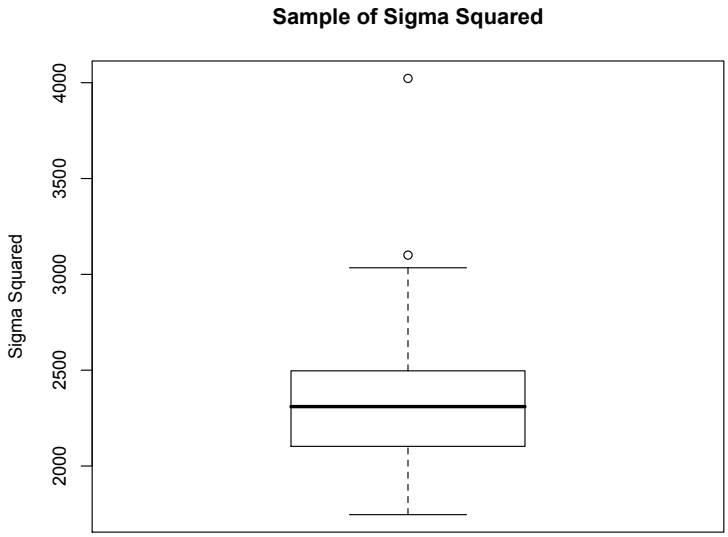

Figure 3.2: Boxplot of a sample of 100 iterations of σ^2.

Table 3.2: Predictions of breast cancer incidences for 2001 using data up to 2000.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	2796	2917	3037	3026
HI	628	679	731	913
IA	2265	2368	2468	2291
NM	848	906	964	1125
UT	800	855	912	1095
AK	-24	42	1086	46
KY	2654	2842	3034	2872
LA	2787	2975	3163	2906
NJ	6473	6662	6851	6725

Table 3.3: Predictions of lung cancer incidences for 2001 using data up to 2000.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	2476	2539	2601	2529
HI	551	576	601	611
IA	2149	2206	2261	2174
NM	673	700	727	813
UT	404	427	451	503
AK	14	46	78	53
KY	4047	4142	4239	4263
LA	3327	3422	3517	3494
NJ	5928	6024	6119	6065

Table 3.4: Predictions of small intestine cancer incidences for 2001 using data up to 2000.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	55	61	67	59
HI	12	15	18	17
IA	54	60	65	60
NM	17	20	23	33
UT	21	24	27	34
AK	-2	1	4	3
KY	72	81	90	76
LA	72	81	90	91
NJ	139	148	157	169

and Utah did not fall within the interval for any type of cancer. Data for these two states show some unusual fluctuations taking place beginning about 1995 and 1996 that aren't happening in the other states. According to the observed values, there seem to be some significant unexpected increases and decreases in the number of cases which maybe causing some inaccuracies in our predictions.

Although this method does a decent job of predicting incidences, the predictions are only falling in the intervals about half the time. As shown in Chapter 2, the mean for the k-year-ahead prediction will be the same as the one-year-ahead prediction. This is a strong limitation of the model since for most cancers we would expect to see an increasing or decreasing trend. Since we weren't able to use the variances to calculate the prediction method, we no longer have the ability to see how the uncertainty changes when two and three-year predictions are made. Therefore, this model seems to only be useful for one-year-ahead predictions.

3.2 Modified Model

In an attempt to improve the predictions and see how the two- and three-year-ahead predictions would change, we incorporated an autoregression coefficient ϕ into the distribution of X_t. Starting with the original model tested

$$Y_{it} \sim \mathrm{N}(\theta_i x_t, \sigma^2),$$

where we assumed

$$X_t \sim N(x_{t-1}, \tau^2),$$

we will now assume

$$X_t \sim N(\phi x_{t-1}, \tau^2).$$

With the additional parameter ϕ, the posterior distributions in our Gibbs sampler will need to be updated. The univariate posterior conditional distribution for X_t where $t = 1, 2, \cdots, n-1$ becomes

$$X_t|\cdots \sim \mathrm{N}\left(\frac{\frac{\sum_{i=1}^{9} y_{i,t}\theta_i}{\sigma^2} + \frac{\phi(x_{t+1}+x_{t-1})}{\tau^2}}{\frac{\sum_{i=1}^{9}\theta_i^2}{\sigma^2} + \frac{1+\phi^2}{\tau^2}}, \left(\frac{\sum_{i=1}^{9}\theta_i^2}{\sigma^2} + \frac{1+\phi^2}{\tau^2}\right)^{-1}\right).$$

The posterior conditional of X_n is given by

$$X_n|\cdots \sim \mathrm{N}\left(\frac{\frac{\sum_{i=1}^{9} y_{i,n}\theta_i}{\sigma^2} + \frac{\phi x_{n-1}}{\tau^2}}{\frac{\sum_{i=1}^{9}\theta_i^2}{\sigma^2} + \frac{1}{\tau^2}}, \left(\frac{\sum_{i=1}^{9}\theta_i^2}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right).$$

The posterior conditional distribution for X_0 and τ^2 will also have slight modifications as can be seen below. First, the posterior conditional of X_0 is,

$$X_0|\cdots \sim \mathrm{N}\left(\frac{\frac{\phi x_1}{\tau^2} + \frac{\mu_0}{w^2}}{\frac{\phi^2}{\tau^2} + \frac{1}{w^2}}, \left(\frac{\phi^2}{\tau^2} + \frac{1}{w^2}\right)^{-1}\right).$$

Then the posterior conditional distribution for τ^2 will look like

$$\tau^2|\cdots \sim \mathrm{IG}\left(\frac{n}{2} + a_2, \left(\frac{1}{b_2} + \frac{\sum_{i=1}^{n}(x_i - \phi x_{i-1})^2}{2}\right)^{-1}\right).$$

Finally, we will also need to update ϕ with each iteration. Assuming a non-informative prior for ϕ, the posterior can be found using the definition and is equal to

$$\begin{aligned}
&f(\phi|\cdots) \propto f(x_1|\phi, x_0, \tau^2)\cdots f(x_n|\phi, x_{n-1}, \tau^2) \\
&\propto \exp\left\{-\frac{1}{2\tau^2}\sum(x_1 - \phi x_0)^2\}\cdots\exp\{-\frac{1}{2\tau^2}\sum(x_n - \phi x_{n-1})^2\right\} \\
&= \exp\left\{-\frac{1}{2\tau^2}\sum(x_1^2 - 2\phi x_1 x_0 + \phi^2 x_0^2)\right\}\cdots\exp\left\{-\frac{1}{2\tau^2}\sum(x_n^2 - 2\phi x_n x_{n-1} + \phi^2 x_{n-1}^2)\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\phi\left(\frac{x_1 x_0}{\tau^2} + \cdots + \frac{x_n x_{n-1}}{\tau^2}\right) + \phi^2\left(\frac{x_0^2}{\tau^2} + \cdots + \frac{x_{n-1}^2}{\tau^2}\right)\right]\right\}.
\end{aligned}$$

This gives us a posterior distribution of

$$\mathrm{N}\left(\frac{x_1x_0 + x_2x_1 + \cdots + x_nx_{n-1}}{x_0^2 + \cdots + x_{n-1}^2}, \frac{\tau^2}{x_0^2 + \cdots + x_{n-1}^2}\right).$$

With the addition of ϕ, predictions in general will be

$$\hat{Y}_{i,n+k} = E(Y_{i,n+k}|\mathbf{Y}) = E(\theta_i\phi^k x_n|\mathbf{Y})$$

which is estimated using

$$E(\theta_i\phi^k x_n|\mathbf{Y}) \approx \frac{1}{M}\sum_{m=1}^{M}\theta_i^{(m)}\phi^{(m)^k}x_n^{(m)},$$

where i represents the region, M is the number of iterations and k is number of years ahead for which prediction is desired.

A histogram of posterior samples of ϕ for breast cancer data up to 2000 is presented in Figure 3.3. The corresponding summary statistics are in Table 3.5. Since the 95% credible interval does not include one, we can conclude that ϕ is actually greater than one, supporting a growing trend for the incidence counts.

Table 3.5: Posterior summary of ϕ for breast cancer incidences.

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.9867	1.0190	1.0220	1.0220	1.0250	1.0640

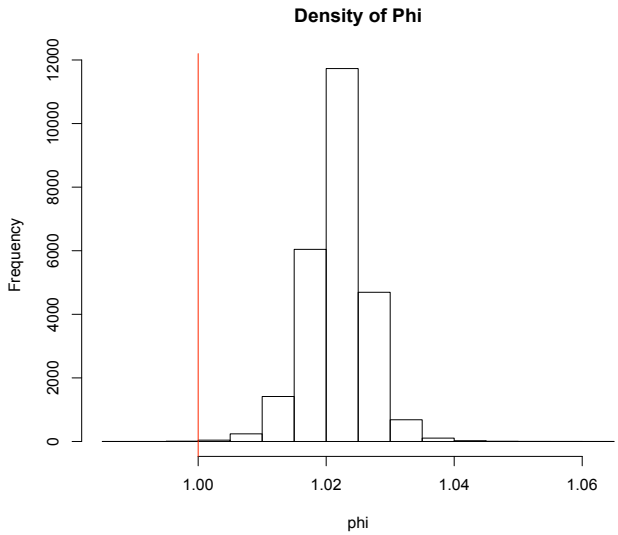

Figure 3.3: Histogram of posterior samples of ϕ for breast cancer incidences. The value of $\phi = 1$ shown in red.

3.3 Modified Predictions

We next used our modified model to develop predictions and prediction intervals for 2001 to 2003. As before, we applied our method to the data on the three cancer sites. For intervals, we continued to use the quantile-based method. Results are presented in Tables 3.6 - 3.14.

When we look at the predictions for 2001, we can see that for each cancer, the actual incidence counts for at least five or six of the nine states fell within the prediction interval. As we move further out and begin predicting two or three years ahead, the predictions become less accurate for the common cancers. Looking at 2002, cancer of the small intestine has just under half of the actual incidences within the prediction

interval and for 2003, over half the states' predictions fell in the interval. For New Mexico and Utah, there was only one prediction that was within the prediction interval. However, Kentucky, Louisiana, and New Jersey were fairly accurate considering they had the least amount of data.

Table 3.6: Predictions of breast cancer incidences for 2001 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	2917	3047	3166	3026
HI	655	709	764	913
IA	2364	2473	2573	2291
NM	886	946	1005	1125
UT	834	893	952	1095
AK	-26	44	112	46
KY	2710	2905	3098	2872
LA	2849	3041	3234	2906
NJ	6604	6807	7010	6725

Table 3.7: Predictions of breast cancer incidences for 2002 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	2965	3113	3248	2824
HI	668	725	783	904
IA	2402	2527	2639	2278
NM	902	967	1030	1109
UT	850	913	975	1046
AK	-26	45	115	46
KY	2765	2968	3172	2827
LA	2907	3107	3312	2889
NJ	6722	6956	7188	6479

Table 3.8: Predictions of breast cancer incidences for 2003 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	3010	3182	3335	2693
HI	680	740	802	853
IA	2440	2582	2709	2121
NM	917	988	1057	1036
UT	864	933	1000	1018
AK	-27	46	117	61
KY	2819	3033	3250	2761
LA	2962	3175	3395	2879
NJ	6832	7108	7382	6294

Table 3.9: Predictions of lung cancer incidences for 2001 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	2540	2611	2680	2529
HI	566	592	620	611
IA	2205	2268	2329	2174
NM	691	720	748	813
UT	415	440	464	503
AK	14	48	80	53
KY	4117	4219	4321	4263
LA	3387	3485	3586	3494
NJ	6027	6135	6244	6065

Table 3.10: Predictions of lung cancer incidences for 2002 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	2574	2659	2740	2613
HI	575	603	632	679
IA	2235	2310	2381	2196
NM	702	733	764	891
UT	422	448	474	478
AK	15	49	82	53
KY	4179	4297	4417	4244
LA	3439	3550	3664	3301
NJ	6110	6248	6389	6077

Table 3.11: Predictions of lung cancer incidences for 2003 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	2606	2708	2805	2559
HI	583	615	646	695
IA	2263	2353	2438	2248
NM	711	747	782	787
UT	428	456	484	514
AK	15	50	83	53
KY	4235	4376	4521	4134
LA	3485	3615	3748	3397
NJ	6185	6364	6548	5888

Table 3.12: Predictions of small intestine cancer incidences for 2001 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	53	62	70	59
HI	12	15	18	17
IA	52	60	68	60
NM	16	20	23	33
UT	20	24	28	34
AK	-2	1	4	3
KY	71	82	93	76
LA	71	82	93	91
NJ	135	150	165	169

Table 3.13: Predictions of small intestine cancer incidences for 2002 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	50	62	75	82
HI	12	15	19	26
IA	49	61	73	69
NM	16	20	25	32
UT	19	25	30	42
AK	-2	1	4	2
KY	68	83	100	85
LA	67	83	100	89
NJ	126	152	179	180

Table 3.14: Predictions of small intestine cancer incidences for 2003 using data up to 2000 and the modified model.

State	2.5% Quantile	Prediction	97.5% Quantile	Observed
CT	47	63	81	72
HI	11	16	21	27
IA	46	62	80	67
NM	15	20	27	34
UT	18	25	33	39
AK	-3	1	4	0
KY	63	85	108	71
LA	63	85	108	86
NJ	117	154	195	181

3.4 Discussion

It appears that the model is predicting more accurately for cancer of the small intestine, which seems unusual since we assumed a normal distribution, and for a rare cancer, a Poisson distribution may be a more appropriate fit. It also appears that predictions seemed better for the states with the least amount of prior year data. Overall, the model with ϕ predicted slightly higher for the one-year-ahead predictions than the model where $\phi = 1$. For example, looking at breast cancer incidences in Table 3.15, most of the predicted values using the modified model are closer to the observed values than the original model where $\phi = 1$. This supports our previous conclusion that ϕ is greater than one. Then we can see the growing trend for incidence counts when we look at the two and three-year ahead predictions. Note however that the methods discussed so far do not allow one to predict incidences for those states without prior data. We attempt to do so in the next chapter.

Table 3.15: Prediction comparison of breast cancer incidences for 2001 using data up to 2000.

State	ϕ=1	Modified Model	Observed
CT	2917	3047	3026
HI	679	709	913
IA	2368	2471	2291
NM	906	946	1125
UT	855	893	1095
AK	42	44	46
KY	2842	2905	2872
LA	2975	3041	2906
NJ	6662	6807	6725

CHAPTER 4

ADDING THE SPATIAL COMPONENT

4.1 Spatial Model

The previous models did not take into account the spatial structure of the states. We now incorporate such information to improve the predictions. This final piece will also allow us to use information from neighboring states to help predict incidences for states where data had not been collected. In addition, this will help improve predictions of states with prior data by sharing of information.

We will look at incorporating neighborhood information using the intrinsic Gaussian Markov random field (IGMRF) described in Rue and Held [12]. We will use the idea of first-order IGMRFs on regular lattices. Let

$$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{51})$$

be the θ-paramters for the fifty states and District of Columbia. For neighboring states i and j, we will assume normal "increments"

$$\theta_i - \theta_j \sim \mathrm{N}(0, \eta^2).$$

Assuming that "increments" are independent, the IGMRF model is given by

$$\pi(\boldsymbol{\theta}) \propto \left(\frac{1}{\eta^2}\right)^{(p-1)/2} \exp\left\{-\frac{1}{2\eta^2}\sum_{j\sim i}(\theta_i - \theta_j)^2\right\},$$

where $p = 51$ for our case and $j\sim i$ represents the unordered pairs of neighbors, with two states defined as neighbors if they share a border. If we let

$$\boldsymbol{\theta}_{-i} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots, \theta_p),$$

it follows tha t

$$\theta_i|\boldsymbol{\theta}_{-i}, \eta^2 \sim \mathrm{N}\left(\frac{\sum_{j\sim i}\theta_j}{n_i}, \frac{\eta^2}{n_i}\right),$$

where n_i represents the number of neighbors of state i. If prior year data are available for state i, the posterior conditional of θ_i is

$$f(\theta_i|\cdots) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_{i,t} - \theta_i x_t)^2\right\} \exp\left\{-\frac{n_i}{2\eta^2}\left(\theta_i - \frac{\sum_{j\sim i}\theta_j}{n_j}\right)^2\right\},$$

which can be written in the form

$$\exp\left\{-\frac{a^*}{2}\left(\theta_i - \frac{b^*}{a^*}\right)^2\right\}.$$

Therefore,

$$\theta_i|\cdots \sim \mathrm{N}\left(\frac{b_3^*}{a_3^*}, \frac{1}{a_3^*}\right),$$

with

$$a_3^* = \frac{\sum x_t^2}{\sigma^2} + \frac{1}{\eta^2}n_i,$$

and

$$b_3^* = \frac{\sum y_{i,t}x_t}{\sigma^2} + \frac{1}{\eta^2}\sum_{j\sim i}\theta_j.$$

If state i does not have prior year information, the posterior will be

$$\mathrm{N}\left(\frac{\sum_{j\sim i}\theta_j}{n_i}, \frac{\eta^2}{n_i}\right).$$

Thus, for states with past data, θ values are updated based on the past data as well as data from neighboring states. For states with no past data, θ values are updated based on θ's of those neighboring states only. The case could exist where state i does not have prior year data and does not share a border with region j for all j; for example an island. In this instance we would have a difficult time updating the value of θ_i. However for our study, the only two states without neighbors are Alaska and Hawaii but we do have data from prior years, so this situation did not arise.

4.2 Spatial Predictions

Since the code is now quite complex, the iterations are running slower and the chain is taking longer to converge. Therefore, we updated the Gibbs sampler using $15,000$ iterations with the first $5,000$ iterations removed for burn-in. As before, we used samples from the posterior distribution to get $\hat{Y}_{i,n+k}$ as well as the 95% prediction intervals. ACS predictions for 2006 were estimated using data up through 2002, so our prediction results for 2006 use data up to 2002 as well. Prediction results, ACS predictions and observed values provided by the National Cancer Institute are presented in Table 4.1, Table 4.2, and Table 4.3.

Looking at the spatial predictions found in Tables 4.1 - 4.3, it seems that the predictions are now relatively close to the predictions made by the ACS. For breast cancer (shown in Table 4.1), states that had prior year data look good, and for some

states (Hawaii, Louisiana and New Jersey), our model is performing better than the ACS model. Similar results can be seen for lung cancer in Table 4.2. For cancer of the small intestine, because it is a rare form of cancer, the results by state were not published by ACS. When we compare our predictions with the observed values in Table 4.3, there are a few cases where fairly significant variations can be seen. It appears though that the predictions for cancer of the small intestine are more consistent and accurate than the predictions for lung or breast cancer.

Overall, the addition of the spatial component has improved the predictions made in Chapter 3. For states where we had prior year data, being able to use information from neighboring states allowed us to make predictions that were comparable to the ACS predictions, and in some cases predictions were better than those reported by ACS. We were also able to use the neighboring region's information to make fairly accurate predictions for states that did not have prior year information. There are still some states where significant departures can be seen between the predicted and the observed values. For these cases, further research needs to be done to identify weighted values by state that could be incorporated as part of our neighborhood information.

Table 4.1: 2006 spatial predictions of breast cancer incidences for all states using data from 1973 to 2002.

State	Observed	ACS Pred	Spatial Pred	2.5% Quan	97.5% Quan
Alabama	3226	3740	3693	-16788	23906
Alaska	342	310	47	-22	115
Arizona	Missing	3740	1644	-13049	16175
Arkansas	1847	2030	2653	-13111	18226
California	22085	21200	2188	-15457	19913
Colorado	2863	2650	2507	-11037	16541
Connecticut	2860	2600	3195	3026	3358
Delaware	599	570	2291	-17438	21618
D.C.	436	470	2476	-20575	26080
Florida	12862	13360	2757	-23773	28933
Georgia	5474	5920	3472	-11402	17901
Hawaii	836	680	770	707	833
Idaho	921	940	2219	-12159	16925
Illinois	8843	9250	2473	-14450	18372
Indiana	3965	4680	2419	-16606	21036
Iowa	2156	2230	2584	2445	2718
Kansas	2011	2080	2076	-14560	19255

Continued on next page ...

Table 4.1: 2006 spatial breast predictions contd.

State	Observed	ACS Pred	Spatial Pred	2.5% Quan	97.5% Quan
Kentucky	2979	3220	3059	2888	3228
Louisiana	2763	4000	3139	2963	3313
Maine	1107	1040	2339	-33362	35794
Maryland	3608	4310	2402	-13242	18351
Massachusetts	5083	4680	2466	-12701	18238
Michigan	6965	7070	2498	-17307	22299
Minnesota	3575	3070	2476	-14813	19919
Mississippi	1701	2290	2314	-13763	18407
Missouri	4041	4570	2087	-10733	14787
Montana	642	620	2386	-15958	20556
Nebraska	1263	1200	2334	-12347	16847
Nevada	1405	1660	2037	-12202	16459
New Hampshire	993	940	2120	-17800	22514
New Jersey	6489	8110	7112	6794	7402
New Mexico	1144	1090	1016	946	1089
New York	14211	14400	2353	-13207	17790
North Carolina	6299	6290	2434	-17503	22353
North Dakota	460	470	2701	-17240	21924
Ohio	7935	9610	2218	-14232	19196

Continued on next page ...

Table 4.1: 2006 spatial breast predictions contd.

State	Observed	ACS Pred	Spatial Pred	2.5% Quan	97.5% Quan
Oklahoma	2455	2860	1955	-11238	15664
Oregon	2791	2810	1851	-15696	19091
Pennsylvania	9831	12320	2494	-11702	16459
Rhode Island	850	780	2763	-19872	25615
South Carolina	3027	3170	2418	-24910	30724
South Dakota	476	520	2581	-12366	17854
Tennessee	4166	4630	2458	-11987	17172
Texas	12750	13150	2040	-14159	18930
Utah	1153	1200	961	893	1031
Vermont	526	520	2575	-19033	24383
Virginia	5167	6080	2319	-12794	17434
Washington	4449	4000	2418	-21438	27668
West Virginia	1317	1400	2263	-14254	18878
Wisconsin	Missing	4000	2646	-16066	21556
Wyoming	317	260	2063	-11962	16420

Table 4.2: 2006 spatial predictions of lung cancer incidences for all states using data from 1973 to 2002.

State	Observed	ACS Pred	Spatial Pred	2.5% Quan	97.5% Quan
Alabama	3784	3530	3520	-33709	40379
Alaska	324	240	52	18	86
Arizona	Missing	3140	1558	-25290	28183
Arkansas	2452	2350	2648	-25948	31040
California	16872	14900	2031	-30827	34319
Colorado	1995	1790	2556	-21829	27504
Connecticut	2631	2000	2768	2667	2869
Delaware	783	550	2259	-33985	38136
D.C.	340	290	2636	-39169	45596
Florida	15891	13280	2561	-45381	49411
Georgia	5734	4860	3227	-23368	29731
Hawaii	744	500	635	602	667
Idaho	756	670	2177	-24711	28624
Illinois	9012	7290	2429	-27641	31773
Indiana	4955	4620	2497	-31975	35992
Iowa	2283	1850	2401	2312	2489
Kansas	1960	1650	1979	-28375	33057

Continued on next page ...

Table 4.2: 2006 spatial lung predictions contd.

State	Observed	ACS Pred	Spatial Pred	2.5% Quan	97.5% Quan
Kentucky	4345	3760	4529	4376	4684
Louisiana	3344	3170	3657	3528	3789
Maine	1256	1030	2358	-61514	64657
Maryland	3489	3320	2484	-25835	31043
Massachusetts	4505	4070	2817	-25176	31205
Michigan	7589	6240	2606	-33206	38771
Minnesota	2882	2610	2538	-29204	33927
Mississippi	2280	2200	2180	-26546	31123
Missouri	4914	4130	2028	-21392	24899
Montana	650	620	2290	-31204	35219
Nebraska	1142	1000	2359	-24356	28335
Nevada	1732	1520	1870	-23880	28037
New Hampshire	928	770	2088	-34323	39170
New Jersey	5975	4960	6503	6289	6713
New Mexico	871	820	778	742	815
New York	13004	9900	2628	-25785	30669
North Carolina	6798	5480	2446	-33395	38046
North Dakota	378	330	2333	-34030	37640
Ohio	9096	7900	2177	-28214	32569

Continued on next page ...

Table 4.2: 2006 spatial lung predictions contd.

State	Observed	ACS Pred	Spatial Pred	2.5% Quan	97.5% Quan
Oklahoma	3097	2560	1864	-22388	26614
Oregon	2554	2290	1647	-29747	32830
Pennsylvania	10432	8450	2524	-23250	27991
Rhode Island	822	680	3515	-37412	44992
South Carolina	3290	3040	2432	-47650	53410
South Dakota	524	440	2631	-24683	30464
Tennessee	5332	4680	2262	-23680	28799
Texas	12312	10780	2057	-27323	32576
Utah	542	480	472	443	501
Vermont	554	390	2718	-36288	41507
Virginia	4952	4840	2323	-25488	29864
Washington	4054	3540	2576	-40769	49034
West Virginia	2038	1640	2179	-27701	31995
Wisconsin	Missing	3040	2743	-30961	36489
Wyoming	257	290	1960	-23656	27676

Table 4.3: 2006 spatial predictions of small intestine cancer incidences for all states using data from 1973 to 2002.

State	Observed	Spatial Pred	2.5% Quan	97.5% Quan
Alabama	126	106	-369	580
Alaska	Suppressed	1	-3	5
Arizona	Missing	50	-290	391
Arkansas	66	76	-287	449
California	552	67	-349	494
Colorado	80	75	-235	393
Connecticut	86	84	63	108
Delaware	20	69	-386	535
D.C.	Suppressed	72	-464	622
Florida	385	77	-529	686
Georgia	143	101	-237	455
Hawaii	26	21	16	28
Idaho	30	66	-275	408
Illinois	262	75	-304	452
Indiana	140	73	-361	513
Iowa	71	81	61	104
Kansas	59	59	-328	458

Continued on next page ...

Table 4.3: 2006 spatial small intestine predictions contd.

State	Observed	Spatial Pred	2.5% Quan	97.5% Quan
Kentucky	105	98	74	126
Louisiana	90	106	80	136
Maine	30	71	-732	867
Maryland	116	73	-288	440
Massachusetts	158	75	-277	433
Michigan	242	74	-380	545
Minnesota	123	74	-327	481
Mississippi	61	65	-300	443
Missouri	107	63	-229	367
Montana	Suppressed	71	-358	498
Nebraska	62	70	-271	410
Nevada	38	63	-273	401
New Hampshire	24	64	-394	538
New Jersey	204	203	154	258
New Mexico	38	29	21	38
New York	457	72	-285	433
North Carolina	183	72	-384	528
North Dakota	Suppressed	79	-385	544
Ohio	254	66	-327	461

Continued on next page ...

Table 4.3: 2006 spatial small intestine predictions contd.

State	Observed	Spatial Pred	2.5% Quan	97.5% Quan
Oklahoma	74	57	-256	374
Oregon	74	57	-340	452
Pennsylvania	315	78	-241	409
Rhode Island	24	85	-446	632
South Carolina	104	71	-572	724
South Dakota	21	76	-274	425
Tennessee	121	70	-257	403
Texas	443	60	-314	453
Utah	32	35	26	45
Vermont	Suppressed	77	-417	569
Virginia	141	70	-289	437
Washington	132	71	-492	659
West Virginia	45	67	-308	454
Wisconsin	Missing	80	-343	515
Wyoming	Suppressed	61	-264	397

CHAPTER 5

CONCLUSION

In this thesis, we have attempted to develop a spatio-temporal model for projecting U.S. cancer incidence counts into the future based on SEER registry data. Using a normal distribution and making assumptions regarding prior distributions allowed us to find conditional posterior distributions for θ_i (the effect of region i from year to year), x_t (which captures the dependency of counts for time t for all regions), and the variance σ^2. While first year predictions seemed realistic, this model did not consistently provide a reasonable two or three-year-ahead predictions. The predictions were then improved upon by the addition of the autoregressive variable ϕ into the distribution of x_t.

Finally, we found a way to incorporate the spatial structure of states by the IGMRF model. Overall, with this addition, we saw an improvement in the predictions, especially for small intestine cancer. Therefore, we can conclude that we have a decent model that can predict cancer incidences for rare or common cancers across the U.S. using prior year data when it is available. If prior year data are unavailable, the model uses information from neighboring states to make predictions. As the

ability to register new incidences becomes easier and more consistent from state to state, researchers will continue to look for ways to improve the prediction of cancer incidences.

To help enhance our model some further research could be done. For example, had the data included other information such as ethnicity, smoking rate, etc., we may have been able to use a regression model to improve our predictions. It is possible that by knowing if the patient had a family history of cancer, was a smoker or was exposed to other elements that increase the risk for cancer, we could determine if there was a correlation between the variables and the number of new incidences. If a relationship was found, it could also help the health care industry to better educate the community on early detection and ways to avoid cancer causing risks.

As we mentioned in Chapter 1, not all the data provided from the SEER registry was used in this study. Some of the registries were for major cities instead of the entire state so this study used only the data collected for an entire state. Although a registry for Atlanta may not give us a complete picture of cancer incidences in Georgia, it would still provide some information and might allow a more accurate prediction for Georgia as well as the neighboring states.

In the final model which incorporated the spatial component, we could also look at the effect of the predictions when μ and η^2 are updated. In Chapter 4 we let μ_i equal the state mean and μ_j be the neighboring mean. We then assumed that μ_i was equal to μ_j and did not change as θ converged. However, as the seen in Figure 1.2, Figure 1.3 and Figure 1.4 there were significant differences in the number of incidences between some states. Therefore, it seems reasonable that by identifying and updating

the mean for each state we may improve our predictions.

Finally, a normal distribution was used throughout the model but since the cancer incidences are count data, a Poisson model might better represent that data. Lawson (2009) described a model by Besag (1975) which uses an autologistic model on binary data in a spatiotemporal setting [7]. The model is able to capture spatial correlation effects as well as allowing conditioning on time labeled neighborhood counts using a pseudolikelihood. Such a model could be modified to use a Poisson distribution but would need to be explored further.

APPENDIX

SEER Incidence Data Used

Table 5.1: Breast cancer incidences from SEER registries

Year	CT	HI	IA	NM	UT	AK	KY	LA	NJ
1973	1543	247	1411	311	365	NA	NA	NA	NA
1974	1766	267	1542	413	388	NA	NA	NA	NA
1975	1817	266	1451	358	391	NA	NA	NA	NA
1976	1713	256	1426	217	384	NA	NA	NA	NA
1977	1721	296	1448	432	382	NA	NA	NA	NA
1978	1714	284	1475	442	422	NA	NA	NA	NA
1979	1790	318	1523	424	473	NA	NA	NA	NA
1980	1831	319	1540	461	432	NA	NA	NA	NA
1981	1949	318	1606	479	459	NA	NA	NA	NA
1982	1914	359	1518	488	464	NA	NA	NA	NA
1983	2044	356	1612	521	549	NA	NA	NA	NA

Continued on next page ...

Table 5.1: Breast cancer incidence contd.

Year	CT	HI	IA	NM	UT	AK	KY	LA	NJ
1984	2183	401	1688	575	591	NA	NA	NA	NA
1985	2293	466	1769	606	576	NA	NA	NA	NA
1986	2273	530	1939	624	671	NA	NA	NA	NA
1987	2502	536	2024	711	698	NA	NA	NA	NA
1988	2538	534	2158	677	693	NA	NA	NA	NA
1989	2482	556	2010	707	686	NA	NA	NA	NA
1990	2629	593	2032	822	710	NA	NA	NA	NA
1991	2578	614	2077	816	776	NA	NA	NA	NA
1992	2574	628	2191	875	750	26	NA	NA	NA
1993	2596	683	2125	795	815	32	NA	NA	NA
1994	2667	655	2094	916	836	35	NA	NA	NA
1995	2623	699	2160	964	887	48	NA	NA	NA
1996	2772	704	2121	1030	868	47	NA	NA	NA
1997	2707	872	2226	1008	904	35	NA	NA	NA
1998	2895	905	2385	1069	1034	48	NA	NA	NA
1999	2950	878	2337	1161	1008	32	NA	NA	NA
2000	2869	783	2188	1118	1053	56	2842	2975	6662
2001	3026	913	2291	1125	1095	46	2872	2906	6725
2002	2824	904	2278	1109	1046	46	2827	2889	6479

Continued on next page ...

Table 5.1: Breast cancer incidence contd.

Year	CT	HI	IA	NM	UT	AK	KY	LA	NJ
2003	2693	853	2121	1036	1018	61	2761	2879	6294

Table 5.2: Lung cancer incidence from SEER registries

Year	CT	HI	IA	NM	UT	AK	KY	LA	NJ
1973	1344	264	1298	305	216	NA	NA	NA	NA
1974	1443	268	1223	310	203	NA	NA	NA	NA
1975	1512	261	1327	364	242	NA	NA	NA	NA
1976	1615	288	1364	366	248	NA	NA	NA	NA
1977	1635	365	1366	364	228	NA	NA	NA	NA
1978	1704	339	1504	419	251	NA	NA	NA	NA
1978	1704	339	1504	419	251	NA	NA	NA	NA
1979	1805	348	1545	463	292	NA	NA	NA	NA
1980	1895	355	1557	451	281	NA	NA	NA	NA
1981	1921	358	1615	490	315	NA	NA	NA	NA
1982	2009	432	1631	497	299	NA	NA	NA	NA
1983	2029	376	1732	494	358	NA	NA	NA	NA
1984	2121	464	1807	535	350	NA	NA	NA	NA
1985	2070	414	1854	534	339	NA	NA	NA	NA

Continued on next page ...

Table 5.2: Lung cancer incidence contd.

Year	CT	HI	IA	NM	UT	AK	KY	LA	NJ
1986	2174	459	1898	571	358	NA	NA	NA	NA
1987	2304	517	1978	598	374	NA	NA	NA	NA
1988	2389	480	2051	638	346	NA	NA	NA	NA
1989	2388	488	2013	599	331	NA	NA	NA	NA
1990	2359	563	1993	688	395	NA	NA	NA	NA
1991	2485	568	2076	637	393	NA	NA	NA	NA
1992	2414	595	2173	707	418	31	NA	NA	NA
1993	2405	598	2148	685	440	49	NA	NA	NA
1994	2383	580	2164	729	453	50	NA	NA	NA
1995	2485	593	2131	723	434	53	NA	NA	NA
1996	2521	623	2274	728	440	45	NA	NA	NA
1997	2480	658	2228	808	458	50	NA	NA	NA
1998	2628	672	2226	791	487	48	NA	NA	NA
1999	2483	687	2196	807	523	40	NA	NA	NA
2000	2459	650	2225	830	435	48	4142	3422	6024
2001	2529	611	2174	813	503	53	4263	3494	6065
2002	2613	679	2196	891	478	53	4244	3301	6077
2003	2559	695	2248	787	514	53	4134	3397	5888

Table 5.3: Small intestine cancer incidence from SEER registries

Year	CT	HI	IA	NM	UT	AK	KY	LA	NJ
1973	31	6	25	6	6	NA	NA	NA	NA
1974	32	5	38	11	13	NA	NA	NA	NA
1975	28	2	39	6	10	NA	NA	NA	NA
1976	34	9	22	11	8	NA	NA	NA	NA
1977	22	4	26	7	11	NA	NA	NA	NA
1978	26	9	21	11	7	NA	NA	NA	NA
1979	22	6	35	10	8	NA	NA	NA	NA
1980	36	8	27	9	10	NA	NA	NA	NA
1981	34	5	39	2	6	NA	NA	NA	NA
1982	37	6	38	8	15	NA	NA	NA	NA
1983	35	7	38	10	9	NA	NA	NA	NA
1984	39	6	35	11	13	NA	NA	NA	NA
1985	31	11	36	15	9	NA	NA	NA	NA
1986	43	5	35	9	23	NA	NA	NA	NA
1987	47	11	49	10	9	NA	NA	NA	NA
1988	45	12	46	18	21	NA	NA	NA	NA
1989	41	10	52	18	16	NA	NA	NA	NA

Continued on next page ...

Table 5.3: Small intestine cancer incidence contd.

Year	CT	HI	IA	NM	UT	AK	KY	LA	NJ
1990	46	5	35	14	9	NA	NA	NA	NA
1991	66	22	46	19	17	NA	NA	NA	NA
1992	47	15	49	18	17	0	NA	NA	NA
1993	49	17	44	14	23	1	NA	NA	NA
1994	51	15	42	16	23	0	NA	NA	NA
1995	50	18	63	16	31	0	NA	NA	NA
1996	49	13	48	21	19	0	NA	NA	NA
1997	73	20	63	19	36	2	NA	NA	NA
1998	67	7	60	25	26	2	NA	NA	NA
1999	74	21	80	27	30	0	NA	NA	NA
2000	55	15	58	22	33	0	81	81	148
2001	59	17	60	33	34	3	76	91	169
2002	82	26	69	32	42	2	85	89	180
2003	72	27	67	34	39	0	71	86	181

R Code

```
library(MCMCpack)  #for rinvgamma function

rivg<-function(a, b){

rinvgamma(1, a, 1/b)

}

##data

d <- read.table("BreastNY.txt", header=TRUE)

d <- d[1:30,]  #only used for testing predictions

#MH 04.28

#neighborhood matrix

n.m <- read.table("StateMatrix3.txt",header=TRUE)

c.mean=apply(d, 2, mean, na.rm=T)

c.var=apply(d, 2, var, na.rm=T)

r.mean = apply(d, 1, mean, na.rm=T)

r.var = apply(d, 1, var, na.rm=T)

#means used to calc starting values for x & theta

n <- nrow(d)

set.seed(1)

#these lines generate the initial values of the parameters x0,

tau^2, sigma^2, x1 - x19, theta1 - theta9

mu.0 <- 0
```

```
t.0 <- 19

a<-2.01; b<-1/1.01

t.2 <- rivg(a,b) #tau^2

s.2 <- rivg(a,b) #sigma^2

w.2 <- 5000  #var for w^2

eta.2 <- 5000

phi <- 1

r = n+1

x.0 <- rnorm(1,mu.0,sqrt(w.2))

x <- c(x.0, r.mean)

theta <- c.mean/mean(r.mean)

#starting values for theta

#MH 04.28

theta.n <- c(theta,runif(42,0,1))

##starting values

S <- 500 #number of iterations

PHI <- matrix(nrow=S, ncol=r)

PHI[1,] <- x

PHI.2 <- matrix(nrow=S, ncol = 3)

PHI.2[1,] <- c(t.2, s.2,phi)

#MH 04.28

PHI.4 <- matrix(nrow=S, ncol = 51)

#matrix of 51 theta's updated using spatial techniques
```

```
PHI.4[1,] <- theta.n


##Gibbs Sampling

for(s in 2:S)   {

cat("s=", s,"\n")

#### phi

ss.x <- sum(x[-r]^2)

s2.s <- t.2/ss.x

mu.t<-sum(x[1:n]*x[2:r])/ss.x

phi <- rnorm(1,mu.t, sqrt(s2.s))

#cat("phi=", phi, "\n")

#### x0

mu.t <- (phi*x[2]/t.2 + mu.0/w.2)/(phi^2/t.2 + 1/w.2)

#posterior mean of x0

s2.s <- 1/(phi^2/t.2 + 1/w.2) #posterior variance of x0

x[1] <- rnorm(1, mu.t, sqrt(s2.s))

#### x1 - x19

s2.s <- 1/((sum(theta.n[1:5]^2)/s.2 + (1+phi^2)/t.2))

#posterior variance of x

for (i in 1:t.0) {

mu.t <- (sum(d[i,1:5]*theta.n[1:5])/s.2 + phi*(x[i+2] + x[i])/t.2)

*s2.s    #posterior mean of x

x[i+1] <- rnorm(1, mu.t, sqrt(s2.s)) #sample from the posterior of x
```

```
}

#### x20 - x27

s2.s <- 1/((sum(theta.n[1:6]^2)/s.2 + (1+phi^2)/t.2))

for (i in (t.0+1):(t.0+8)) {

mu.t <- (sum(d[i,1:6]*theta.n[1:6])/s.2 + phi*(x[i+2] + x[i])/t.2)

*s2.s

x[i+1] <- rnorm(1, mu.t, sqrt(s2.s))

}

#### x28 - x29

s2.s <- 1/((sum(theta.n[1:9]^2)/s.2 + (1+phi^2)/t.2))

for (i in (t.0+9):(n-1)) {

mu.t <- (sum(d[i,1:9]*theta.n[1:9])/s.2 + phi*(x[i+2] + x[i])/t.2)

*s2.s

x[i+1] <- rnorm(1, mu.t, sqrt(s2.s))

}

#### xn: n=30 (through 2002)

s2.s <- 1/((sum(theta.n[1:9]^2)/s.2 + 1/t.2))

mu.t <- (sum(d[n, 1:9]*theta.n[1:9])/s.2 + (phi*x[n]/t.2))*s2.s

x[n+1] <- rnorm(1, mu.t, sqrt(s2.s))

#### tau.2

mu.t <- (n/2) + a  #posterior mean of tau^2

s2.s<-1/(1/b + sum((x[2:r]-phi*x[1:n])^2)/2)

#posterior var of tau^2
```

```
t.2<-rivg(mu.t, s2.s)

#### sigma.2

mu.t <- ((5*n + n - t.0) + 3*(n - (t.0 + 8)))/2 + a

 #posterior mean of sigma^2

term<-sum((d[,(1:5)]-outer(x[-1], theta.n[1:5]))^2)

term<-term+sum((d[(t.0+1):n,6] - theta.n[6]*x[(t.0+2):(n+1)])^2)

term<-term+sum((d[(t.0+9):n,7:9]

  - outer(x[(t.0+10):(n+1)], theta.n[7:9]))^2)

s2.s <- 1/(term/2 + 1/b)  #posterior var of sigma^2

s.2<-rivg(mu.t, s2.s)

#cat("tau_sq=", t.2, "sigma_sq=", s.2,"\n")

#MH 04.28

#no y information available - theta being updated using information

from neighbors - 41 states, theta.10 to theta.51

for(j in 10:51){

mu.t <- sum(n.m[j,3:53]*theta.n)/sum(n.m[j,3:53])

s2.s <- eta.2/sum(n.m[j,3:53])

theta.n[j] <- rnorm(1,mu.t,sqrt(s2.s))

}

#MH 04.28

#update theta 1 - 9 using spatial posterior

#theta 1 - 5

x.sqsum <- sum(x[-1]^2)
```

```
for(j in 1:5){

s2.s <- 1/(x.sqsum/s.2 + (1/eta.2)*sum(n.m[j,3:53]))

mu.t <- ((sum(d[,j]*x[-1]))/s.2 + sum(n.m[j,3:53]*theta.n)/eta.2)

*s2.s

theta.n[j] <- rnorm(1,mu.t,sqrt(s2.s))

}

#### theta 6

x.sqsum <- sum(x[(t.0+2):(n+1)]^2)

s2.s <- 1/(x.sqsum/s.2 + (1/eta.2)*sum(n.m[6,3:53]))

mu.t <- ((sum(d[(t.0+1):n,6]*x[(t.0+2):(n+1)]))/s.2

      + sum(n.m[6,3:53]*theta.n)/eta.2)*s2.s

theta.n[6] <- rnorm(1, mu.t, sqrt(s2.s))

#### theta 7-9

x.sqsum <- sum(x[(t.0+10):(n+1)]^2)

for(j in 7:9){

s2.s <- 1/(x.sqsum/s.2 + (1/eta.2)*sum(n.m[j,3:53]))

mu.t <- ((sum(d[(t.0+9):n, j]*x[(t.0+10):(n+1)]))/s.2

      + sum(n.m[j,3:53]*theta.n)/eta.2)*s2.s

theta.n[j] <- rnorm(1, mu.t, sqrt(s2.s))

}

PHI[s,] <- x

PHI.2[s,] <- c(t.2, s.2, phi)

PHI.4[s,] <- theta.n
```

```
}


#Predictions 2003 - 2006

start.i = 5000

end.i = S

N=n

pred.1 <- matrix(nrow=3, ncol=51)

pred.2 <- matrix(nrow=3, ncol=51)

pred.3 <- matrix(nrow=3, ncol=51)

pred.4 <- matrix(nrow=3, ncol=51)

for(i in 1:51){

pred.1[1, i]<-quantile(PHI.4[start.i:end.i, i]

      *PHI.2[start.i:end.i, 3]^1*PHI[start.i:end.i, N+1],.025)

pred.1[2, i]<-mean(PHI.4[start.i:end.i, i]

      *PHI.2[start.i:end.i, 3]^1*PHI[start.i:end.i, N+1])

pred.1[3, i]<-quantile(PHI.4[start.i:end.i, i]

      *PHI.2[start.i:end.i, 3]^1*PHI[start.i:end.i, N+1],.975)

pred.2[1, i]<-quantile(PHI.4[start.i:end.i, i]

      *PHI.2[start.i:end.i, 3]^2*PHI[start.i:end.i, N+1],.025)

pred.2[2, i]<-mean(PHI.4[start.i:end.i, i]

      *PHI.2[start.i:end.i, 3]^2*PHI[start.i:end.i, N+1])

pred.2[3, i]<-quantile(PHI.4[start.i:end.i, i]

      *PHI.2[start.i:end.i, 3]^2*PHI[start.i:end.i, N+1],.975)
```

```
pred.3[1, i]<-quantile(PHI.4[start.i:end.i, i]
      *PHI.2[start.i:end.i, 3]^3*PHI[start.i:end.i, N+1],.025)
pred.3[2, i]<-mean(PHI.4[start.i:end.i, i]
      *PHI.2[start.i:end.i, 3]^3*PHI[start.i:end.i, N+1])
pred.3[3, i]<-quantile(PHI.4[start.i:end.i, i]
      *PHI.2[start.i:end.i, 3]^3*PHI[start.i:end.i, N+1],.975)
pred.4[1, i]<-quantile(PHI.4[start.i:end.i, i]
      *PHI.2[start.i:end.i, 3]^4*PHI[start.i:end.i, N+1],.025)
pred.4[2, i]<-mean(PHI.4[start.i:end.i, i]
      *PHI.2[start.i:end.i, 3]^4*PHI[start.i:end.i, N+1])
pred.4[3, i]<-quantile(PHI.4[start.i:end.i, i]
      *PHI.2[start.i:end.i, 3]^4*PHI[start.i:end.i, N+1],.975)
}
```

Bibliography

[1] American Cancer Society. (2006) *Cancer Facts & Figures 2006.* Available at http://www.cancer.org/acs/groups/content/@nho/documents/document/caff2006pwsecuredpdf.pdf.

[2] American Cancer Society. (2009) *Cancer Statistics, 2009.* Available at http://caonline.amcancersoc.org/cgi/content/full/59/4/225.

[3] American Cancer Society. (2010) *What are the key statistics about lung cancer?* Available at http://www.cancer.org/Cancer/LungCancer-Non-SmallCell/DetailedGuide/non-small-cell-lung-cancer-key-statistics.

[4] Centers for Disease Control and Prevention. (2010) *Breast Cancer Statistics.* Available at http://www.cdc.gov/cancer/breast/statistics/.

[5] Ghosh, K. and Tiwari, R.C. (2007) Prediction of U.S. Cancer Mortality Counts Using Semiparametric Bayesian Techniques, *Journal of the American Statistical Association*, **102**(477), 7–15.

[6] Kim H.-J., Fay M. P., Feuer E. J. and Midthune, D. N. (2000) Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, **19**, pp 335–351.

[7] Lawson, A. B. (2009). *Bayesian Disease Mapping, Hierarchical Modeling in Spatial Epidemiology*. New York: CRC Press.

[8] National Cancer Institute. (2010) *Overview of the SEER Program.* Available at `http://seer.cancer.gov/about/overview.html`.

[9] National Cancer Institute. (2010) *SEER Stat Fact Sheets: Lung & Bronchus.* Available at `http://seer.cancer.gov/statfacts/html/lungb.html`.

[10] Pickle, L. W., Hao, Y., Jemal A., Zou Z., Tiwari R. C., Ward E., Hachey M., Howe H. L. and Feuer E. J. (2007) A New Method of Estimating United States and State-level Cancer Incidence Counts for the Current Calendar Year. *CA Cancer J Clin*, **57**, pp.30–42.

[11] Pickle, Linda W, Eric J. Feuer, B.K. Edwards. (2003) US Predicted Cancer Incidence, 1999: Complete Maps by County and State From Spatial Projection Models, *NCI Cancer Surveillance Monograph Series, Number 5*. Bethesda, MD: National Cancer Institute, NIH Publication No. 03-5435.

[12] Rue, H. and L. Held. (2005) *Gaussian Markov Random Fields.* Chapman & Hall/CRC, Boca Raton, FL, 2005.

[13] Tadeusz, D., and T. Hakulinen. (2000) Comparison of Different Approaches to Incidence Prediction Based on Simple Interpolation Techniques. *Statistics in Medicine*, **19**, pp 1741–1752.

[14] Tiwari R. C., Ghosh, K., Jemal A., Hachey M., Ward E., Thun M.J. and Feuer E.J. (2004) A New Method of Predicting US and State-Level Cancer Mortality Counts for the Current Calendar Year, *CA: A Cancer Journal for Clinicians*, **54**(1), 30–40.

VITA

Graduate College

University of Nevada, Las Vegas

Michelle Hamlyn

Degrees:

Bachelor of Arts in Mathematics, 2005

Wittenberg University, Springfield, Ohio

Thesis Title: A Comparison of Spatio-Temporal Prediction Methods of Cancer Incidence in the U.S.

Thesis Examination Committee:

Chairperson, Kaushik Ghosh, Ph. D.

Committee Member, Sandra Catlin, Ph. D.

Committee Member, Anton Westveld, Ph. D.

Graduate Faculty Representative, Sheniz Moonie, Ph. D.