ABSTRACT

Conjugate Hierarchical Models for Spatial Data: An Application of an Optimal Selection Procedure

John Jacob McBride

Mentor: Thomas L. Bratcher, Ph.D.

The theory of generalized linear models provides a unifying class of statistical distributions that can be used to model both discrete and continuous events. In this dissertation we present a new conjugate hierarchical Bayesian generalized linear model that can be used to model counts of occurrences in the presence of spatial correlation. We assume that the counts are taken from geographic regions or *areal units* (zip codes, counties, etc.) and that the conditional distributions of these counts for each area are distributed as Poisson having unknown rates or relative risks. We incorporate the spatial association of the counts through a *neighborhood* structure which is based on the arrangement of the areal units. Having defined the neighborhood structure we then model this spatial association with a *conditionally autoregressive* (CAR) model as developed by Besag (1974). Once the spatial model has been created we adapt a subset selection procedure created by Bratcher and Bhalla (1974) to select the areal unit(s) having the highest relative risks.

Conjugate Hierarchical Models for Spatial Data:
An Application of an Optimal Selection Procedure

by

John Jacob McBride

A Dissertation

Approved by Department of Statistical Science

Jack Tubbs

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by Dissertation Committee

Tom L Bratcher

Thomas L. Bratcher, Ph.D.

Dennis A. Johnston, Ph.D.

Dean M. Young

Dean M. Young, Ph.D.

Joseph D. White, Ph.D.

John W. Seaman, Jr.

John W. Seaman, Jr., Ph.D.

Jeanne Hill

Jeanne Hill, Ph.D.

James D Stamey

James D. Stamey, Ph.D.

Accepted by the Graduate School
May 2006

Larry Lyon

J. Larry Lyon, Ph.D., Dean

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As I celebrate the end of this long journey I am humbled by certain awareness; the completion of this dissertation was made possible only through the gifts of our maker. Included in those gifts is the gift of a wonderful family, the gift of the amazing faculty and staff at Baylor University - namely my advisor Dr. Tom Bratcher but also my committee members: Dr. Jack Tubbs, Dr. John Seaman, Dr. Dean Young, Dr. James Stamey, Dr. Dennis Johnston, and Dr. Jeanne Hill. Above all, I am grateful for the gift of Lindsay. All of you in many ways and more provided me the strength and encouragement to chase down those elusive three letters that now follow my name. For those that I have not mentioned by name - there are just too many to list - please know that you were also paramount to my success. Perhaps one day I too will lend an unbiased ear or a strong shoulder for you to lean on.

Words, sounds, speech, men, memory, thoughts,
fears and emotions--time--all related... all made from one... all made in one.
Blessed be his name.

John Coltrane, *A Love Supreme*

CHAPTER ONE

Introduction

Statistical methods for spatial data have steadily gained attention over the past few years as a result of simulation-based computing procedures such as Markov Chain Monte Carlo (MCMC) and technological advances in geographic information systems (GIS). These technological advances have motivated some to author comprehensive texts which address the theoretical aspects and the computing techniques associated with the different applications of spatial models. See, for example, Banerjee, Carlin, and Gelfand (2004), Lawson (2003), Cressie (1993). While many statisticians are interested in predicting or *kriging* unobservable quantities among some specified spatial domain others such as statistical epidemiologists and public health officials are interested in accurately modeling not only the spread of infectious diseases but the disease risk of non-infectious diseases and ultimately producing a disease map.

The concept of disease mapping dates back centuries to an early example by Dr. John Snow (Snow (1854)) who mapped the addresses of cholera victims in relation to the locations of water supplies. Snow used this particular disease map to identify putative sources of disease outbreak. This is just one application of a disease map used in public health, Lawson (2004) and Waller and Gotway (2003) give an overview of the wide range of uses of disease maps used for studying the geographical distribution of disease. However, the two typical uses are those used to assess the need for geographical variation in health resource allocation and those useful in research studies pertaining to the relationship between disease incidence/prevalence and explanatory variables. The first case is intended to produce a map ‘clean’ of any random noise. This means the map delineates elevated risk. The latter is sometimes called ‘ecological analysis’ and can be regarded as spatial regression. In this analysis the focus is on the relationship between disease incidences and explanatory covariates, usually at an aggregated spatial level, in order to assess specific hypotheses. Of course, these hypotheses are addressed visually through some disease map. In this work we concentrate mostly on modeling issues and give only a brief discussion of the cartographic issues pertaining to the representation of geographic information. However, for the reader who wishes to investigate topics relating to symbolic representation, display methods of intensity, or color scheme see Lawson (2001) or Pickle and Hermann (1995).

To begin, we assume that our data arise from a geographic region which can be divided into smaller areas such as census tracts, counties, precincts, etc. and that we have the available aggregate counts for these geographic *areal units*. Thus, we have lost all information at the individual level. Waller (2003) considers the trade-off between statistical stability of risk estimates and geographic precision. The detection of locally elevated risk requires geographically small units; however, these smaller regions result in rate estimates based on smaller samples. In the case of a rare disease the rates computed for these less populated units are most often unstable.

While there are numerous methods suggested in the literature that address the previous dilemma, we focus mainly on methods which make use of hierarchical Bayesian models. Clayton and Kaldor (1987, 1989) were some of the first to incorporate Bayesian modeling techniques in the area of ecological analysis. Others include Clayton and Bernardinelli, 1992; and Mollié 1996). Statisticians for the most part assume that the aggregate counts for the areal units are distributed as Poisson with unknown relative risks. The usual method is to model the logarithm of the relative risks with a hierarchical generalized linear model consisting of both local and regional covariates as well as a random effects term for each areal unit corresponding to *unstructured heterogeneity* (Lawson, 2003). To account for the spatial correlation, sometimes called *structured heterogeneity* (Lawson, 2003), of the areal units many will include an additional random effects term for each areal unit. It is usually assumed that the collection of full conditional distributions for the spatial components defines a Marko random field (MRF). A commonly used MRF model is the intrinsic Gaussian autoregression prior considered by Besag (1974) and Besag, York, and Mollié (1991).

We take a much different approach in that we assign a conjugate prior to the logarithm of the rates and model the prior means with a hierarchical generalized linear spatial model. We have essentially taken the conjugate hierarchical generalized linear model presented by Albert (1988) and added a spatial component. In doing so, we have gained the ability to directly quantify the overdispersion which is usually present in areal unit data. The reader will find a discussion of both the frequentist and the traditional Bayesian methods used for spatial regression and disease mapping in Chapter 4. Furthermore, we apply our unique conjugate model to an original data set constructed by this author, the Waco Police Department, and the Center for Geographic Applied Spatial Research at Baylor University, wherein the ideas and concepts used in disease mapping are transferred to mapping call rates of habitat burglaries. Instead of aggregate counts of disease we consider aggregate counts of 911 calls classified as habitat burglaries. In addition, we also provide the reader with a procedure for selecting the beat(s) having the highest relative crime risk.

Clearly the emphasis of this dissertation is the material presented in Chapter 4; however, the supplementary material useful for the understanding of Chapter 4 is thoroughly developed in the preceding chapters. Chapter 2 gives an overview of the first theme of this dissertation, subset selection. We give an in-depth discussion of a subset selection procedure useful in determining a ‘best’ parameter among several populations. We actually apply this selection procedure to home run hitting data in Chapter 3. In that same chapter we present the second theme: conjugate hierarchical generalized linear models. Chapter 3 consists of a complete formulation of the conjugate hierarchical generalized linear model proposed by Albert (1988), along with derivations of posteriors and marginal distributions, as well a discussion of the relationship between prior distributions and the relative posterior sensitivity. We conclude this chapter with an example of a quadratic conjugate hierarchical generalized linear model used to model Sammy Sosa’s career home run hitting data.

CHAPTER TWO

An Optimal Selection Procedure

2.1 Introduction

A decision maker is often faced with the task of selecting among several processes or populations the one which will produce the greatest yield or highest rate. Similarly, one may wish to select the smallest yield or lowest rate. For example, an experimenter might be interested in determining which production technique gives the lowest percentage of defects; a crime analyst might consider which reporting district has the highest rate of violent crimes; a baseball fan would inquire about the best home run hitter of the 20th century. In any case a selection must be made with less than certain information. There are of course various procedures for selecting a subset to contain the 'best' parameter. In Section 2.2 we will give a brief overview of several subset selection procedures appearing the literature. Then in Section 2.3 we will review a Bayes solution to the selection procedure with respect to a constant loss function. Finally, we will utilize this Bayesian selection procedure in a simulation study and assess the probability of correct selection and expected size in which the populations generating the samples are Poisson.

2.2 Literature Review

The concept of subset selection began as early as 1957 with a paper describing a statistic which arises in ranking and selection (Gupta, 1957). Shanti Gupta had already been working in the area of decision theory called ranking before he and Milton Sobel developed the concept of subset selection. Together these authors provide the literature with a variety of classical mechanisms for subset selection of location and shape parameters in both discrete and continuous distributions. In fact, Roger Berger (1980) and Thomas Santner (1995), two of Gupta's former students have also contributed to the area of subset selection. However, to our knowledge before Bratcher and Bhalla (1974) there are no published works that contain a Bayesian subset selection procedure. R.P. Bland, who was working in the area of ranking and multiple comparisons, had an unpublished manuscript in which he gave a detailed description of a Bayes' solution to the selection problem that utilized a linear loss function. This is not to be confused with Bratcher and Bhalla (1974) who used a constant loss function to derive their selection procedure. You may find an example of their method in Stamey, Bratcher, and Young (2004) who applied the selection procedure to Poisson rates subject to misclassification. Currently there are still very few Bayesian subset selections procedures found in the literature. For an alternative to Bratcher and Bhalla see Gupta and Yang (1985), Deely and Berger (1988), or Schulter, Deely, and Nicholson (1997). The works by Deely are unique in that the experimenter must predetermine the size of the subset. Furthermore, Deely provides a different selection procedure based on the posterior predictive distribution rather than the usual posterior distribution; see Schulter et al (1997).

2.3 Decision Theoretic

Bratcher and Bhalla (1974) derive a decision theoretic approach to partitioning *m* parameters into two sets. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)$ be the parameter vector of interest, for instance a collection of Poisson rates or binomial proportions. There are $2^m - 1$ subsets (excluding the null set) of the *m* parameters which may be selected as the superior set, *S*.

Each possible superior set corresponds to a composite of actions generated from *m* two decision problems of the form $d_+^i : \theta_i \in S$ or $d_-^i : \theta_i \in S^C$, *i* = 1, 2, ... , *m*. Bratcher and Bhalla (1974) assume the following constant loss functions:

$$L_+^i(\theta) = \begin{cases} 0 \text{ if } \theta_i = \theta_{\max} \\ c_1 \text{ if } \theta_i \neq \theta_{\max} \end{cases} \text{ and } L_-^i(\theta) = \begin{cases} c_2 \text{ if } \theta_i = \theta_{\max} \\ 0 \text{ if } \theta_i \neq \theta_{\max} \end{cases} \quad i = 1, 2, \ldots , m,$$

where L_+^i is the loss function for decision d_+^i and L_-^i is the loss function for decision d_-^i. Then the total loss incurred for selecting a subset of size *N* is given by

$$\begin{aligned} L_S(\theta) &= (N-1)c_1, \text{ if } \theta_{\max} \in S \\ &= Nc_1 + c_2, \text{ if } \theta_{\max} \notin S \end{aligned}$$

where $\theta_{\max} = \max\{\theta_1, \theta_2, \ldots, \theta_m\}$ and c_1 and c_2 are constants. We should note that c_2 is greater than c_1 since it represents the loss of the more serious error of not selecting $\theta_{\max}$. In the selection process one does not need c_1 and c_2 but only the ratio $c = c_2/c_1$. Clearly the action of whether or not to include θ_i in the subset *S* should in some way depend on the two losses that may result. From the Bayes decision criterion we will include θ_i in *S* if the expected loss of inclusion is less than that of exclusion that is, $E\left(L_+^i(\underline{\theta}) \mid \underline{x}\right) \leq E\left(L_-^i(\underline{\theta}) \mid \underline{x}\right)$ where $\underline{x}$ represents the data or a vector of sufficient statistics for $\underline{\theta}$. Writing the expectations as a function of $\Pr(\theta_i = \theta_{\max} \mid \underline{x})$ gives the following

$$\begin{aligned} E\left(L_+^i(\underline{\theta}) \mid \underline{x}\right) &= 0 \cdot \Pr(\theta_i = \theta_{\max} \mid \underline{x}) + c_1 \cdot \Pr(\theta_i \neq \theta_{\max} \mid \underline{x}) \\ &= c_1\left[1 - \Pr(\theta_i = \theta_{\max} \mid \underline{x}\right] \end{aligned}$$

$$\begin{aligned} E\left(L_+^i(\underline{\theta}) \mid \underline{x}\right) &= 0 \cdot \Pr\left(\theta_i \neq \theta_{\max} \mid \underline{x}\right) + c_2 \cdot \Pr\left(\theta_i = \theta_{\max} \mid \underline{x}\right) \\ &= c_2 \Pr\left(\theta_i = \theta_{\max} \mid \underline{x}\right). \end{aligned}$$

The decision to include θ_i in *S* can now be rewritten as

$$c_1\left[1-\Pr\left(\theta_i=\theta_{\max}\mid \underline{x}\right)\right]\le c_2\cdot\Pr\left(\theta_i=\theta_{\max}\mid \underline{x}\right)$$

or

$$\Pr\left(\theta_i=\theta_{\max}\mid \underline{x}\right)\ge 1\big/\left(c+1\right). \tag{2.1}$$

It is clear from equation (2.1) that the decision to place θ_i in the superior set is not just dependent on the sample information but also the penalty constant *c*. In the next section we investigate the relationship between *c*, sample size, the probability of correct selection and expected size.

2.4 Sample Size Determination Study

2.4.1 Preliminaries

Until now there have been no known sample size determination studies used to calculate the probability of correct selection and expected size given the assumption of a constant loss function as in the decision theoretic presented by Bratcher et al (1974). For this sample size determination study we consider only the Poisson model but our method can be easily transferred to the binomial or other discrete models. Therefore, we assume the data $x_{i1},\ldots,x_{in}$ for the i^{th} population can be modeled as Poisson with rate λ_i. Assuming *n* is the common sample size, we may summarize the sample information with the sufficient statistics $t_i=\sum_{j=1}^{n}x_{ij}$. Thus, our decision criterion in Equation (2.1) will be based on the posterior of λ_i given the totals t_i. In the absence of prior information it is customary to assign independent non-informative priors to the Poisson rates. For the first case we assume $\pi\left(\lambda_i\right)\propto c$, which leads to the posteriors being distributed as independent gamma$\left(t_i+1,n\right)$. The assumption of a flat uniform prior reduces the complexity of the computations used for calculating the probability of correct selection and expected size since we can make use of functions available in the statistical package R. Alternatively, we could assume a hierarchical Poisson model which would greatly increase the computing time. Depending on how 'informative' the prior structure, a third stage hierarchy may provide "borrowing strength" across the parameters and reduce the posterior variability. For the second case we assume conjugacy and assign independent gamma$\left(\alpha,\beta\right)$ to the Poisson rates. To complete the hierarchy we add the following distributions for the hyper priors:

$$\alpha \sim \exp\left(1\right)$$

$$\beta \sim \text{gamma}\left(.001,.001\right)$$

The resulting posterior and full conditional distributions are analytically intractable making the free statistical software WinBUGS an ideal candidate to aid in calculating the probability of correct selection and expected size. WinBUGS will use an adaptive rejection sampling procedure to simulate values from the posterior. Furthermore, we will use the 'R2WinBUGS' package created by Sturtz, Ligges, and Gelman (2004) to call WinBUGS from R. The next subsection gives tables as well as explanations that summarize the results of our simulation study.

2.4.2 Results

Appendix A gives a complete annotated version of the computer program used to generate the following tables; however, we will now provide the reader with a brief outline of the steps used in the simulation.

1. Using the triangle distribution with endpoints (a, b) and mode c generate sets of size five and ten. Each of these sets will represent a set of Poisson rates. A large number of these sets – typically greater than 10,000 - must be used for the Monte Carlo simulation.

2. Sampling from Poisson distributions with the rates proportional to those from step one generate the count totals.

3. For each population calculate the *probability inclusion criterion PIC* $\equiv \Pr\left(\lambda_i = \lambda_{\max} \mid t_1, \ldots, t_m\right)$. If the *PIC* exceeds $1/(c+1)$ we take action d_i^+

4. Calculate the percentage of times that the superior set actually contained $\lambda_{\max}$. This is the probability of correct selection.

5. Calculate the relative frequency in which action d_i^+ is taken. This is the expected size.

In this study the parameter specification for the triangle distribution used to simulate the Poisson rate is motivated by two examples. The first appears in Suissa and Salmi (1989) whereby the physicians were interested in assessing the best treatment – placebo, radiotherapy, chemotherapy, or both - for Hodgkin’s disease. The second appears in Kvam and Miller (2002) in which the experimenters were concerned with selecting the largest pump failure rate. In Suissa et al. (1989) the clinicians recorded the number of observed leukaemias for patients that were given one of four different treatments. Here the experimenter would be interested in determining the sample size required for selecting a subset of treatments given a specified probability of correct selection. As for Kvam et al. (2002) the failure data was from 10 pump systems in the Farley-1 nuclear power plant. Tables 1-2 give both approximations of the probability of correct selection and expected size when only five populations (leukemia treatments) were studied whereas Tables 3-4 give the results for 10 populations (pumps). Table 1 and Table 3 refer to the case of the uniform prior and Table 2 and Table 4 refer to the case of a hierarchical model.

Table 1. Uniform prior with $m = 5$

	Probability of Correct Selection					Expected Size				
size \ penalty	c = 4	c = 5.67	c = 9	c = 11.5	c = 19	c = 4	c = 5.67	c = 9	c = 11.5	c = 19
n = 5	0.778	0.8352	0.89	0.9095	0.9443	1.6883	1.9571	2.3028	2.4753	2.8234
n = 10	0.844	0.8832	0.9227	0.9386	0.9603	1.588	1.7969	2.058	2.1927	2.461
n = 15	0.8744	0.9076	0.9407	0.9531	0.9714	1.5291	1.7027	1.9293	2.0432	2.2702
n = 20	0.8881	0.9176	0.9459	0.9568	0.9736	1.4768	1.6272	1.8261	1.9313	2.1321
n = 25	0.9098	0.9346	0.9568	0.9659	0.9794	1.4533	1.5952	1.7659	1.8562	2.0274
n = 30	0.9185	0.9436	0.9648	0.9721	0.984	1.4122	1.546	1.7021	1.7831	1.9512
n = 45	0.9296	0.948	0.9656	0.974	0.9849	1.3582	1.4614	1.5932	1.6687	1.8086

Table 2. Hierarchical model with $m = 5$

	Probability of Correct Selection					Expected Size				
size \ penalty	c = 4	c = 5.67	c = 9	c = 11.5	c = 19	c = 4	c = 5.67	c = 9	c = 11.5	c = 19
n = 5	0.7824	0.8358	0.8949	0.9176	0.9485	1.6973	1.9846	2.3474	2.5394	2.8926
n = 10	0.8442	0.8861	0.9241	0.9402	0.9651	1.5981	1.8138	2.0926	2.2342	2.5169
n = 15	0.8704	0.9063	0.9386	0.9505	0.9706	1.5357	1.7176	1.9496	2.0704	2.2935
n = 20	0.8949	0.9255	0.9522	0.9625	0.9767	1.489	1.6492	1.8567	1.9539	2.16
n = 25	0.9113	0.9337	0.9569	0.9675	0.9807	1.4534	1.5926	1.7752	1.8653	2.0417
n = 30	0.918	0.9406	0.9643	0.9732	0.9836	1.4333	1.5652	1.7357	1.8185	1.9835
n = 45	0.9348	0.9522	0.9706	0.9779	0.9867	1.363	1.4645	1.5953	1.6614	1.7996

Table 3. Uniform prior with $m = 10$

	Probability of Correct Selection					Expected Size				
size \ penalty	c = 9	c = 11.5	c = 19	c = 24	c = 39	c = 9	c = 11.5	c = 19	c = 24	c = 39
n = 5	0.7845	0.8206	0.8803	0.9001	0.9346	2.6971	3.0179	3.6945	3.9946	4.6161
n = 10	0.8499	0.877	0.9206	0.9362	0.959	2.438	2.6829	3.1813	3.424	3.8802
n = 15	0.8814	0.9042	0.9374	0.9515	0.9684	2.2744	2.4748	2.8842	3.0753	3.4641
n = 20	0.9051	0.9238	0.9511	0.9607	0.9753	2.157	2.3326	2.6882	2.8517	3.2048
n = 25	0.9194	0.9351	0.9601	0.9684	0.9795	2.0774	2.2378	2.5608	2.7098	3.0024
n = 30	0.9299	0.9454	0.9658	0.9727	0.9823	2.0189	2.1667	2.4539	2.5773	2.8475
n = 45	0.9437	0.9544	0.9724	0.9795	0.9875	1.8521	1.9656	2.2066	2.3154	2.5321

Table 4. Hierarchical model with $m = 10$

	Probability of Correct Selection					Expected Size				
size \ penalty	c = 9	c = 11.5	c = 19	c = 24	c = 39	c = 9	c = 11.5	c = 19	c = 24	c = 39
n = 5	0.7907	0.8308	0.8914	0.9132	0.9437	2.7666	3.1183	3.8327	4.1597	4.8079
n = 10	0.8516	0.8786	0.9217	0.9365	0.9571	2.4851	2.7364	3.2517	3.491	4.1205
n = 15	0.8829	0.9049	0.9415	0.9538	0.9708	2.2996	2.5098	2.9363	3.1343	3.5315
n = 20	0.9083	0.9255	0.9537	0.9647	0.9772	2.1863	2.3701	2.7332	2.9032	3.2411
n = 25	0.9173	0.9322	0.9552	0.9645	0.9767	2.1049	2.2643	2.5935	2.7372	3.0533
n = 30	0.9297	0.9446	0.9665	0.9716	0.9817	2.0024	2.1494	2.4553	2.5946	2.8641
n = 45	0.9453	0.9565	0.9745	0.9803	0.9873	1.8822	1.9974	2.2362	2.3397	2.5559

The hierarchical model did not noticeably out perform the model with a non-informative prior; the purported “borrowing strength” is not prevalent, at least not in these results. For that matter since computing time is significantly decreased when using the non-informative prior it is suggested to the reader not to use the hierarchical conjugate model to determine the sample. We now return to the example from Suissa and Salmi (1989). If the loss for not selecting $\theta_{\max}$ is four times the loss for selecting $\theta_i \neq \theta_{\max}$, i.e. $c = 4$, and the probability of correct selection is 85% then we would require a sample size of approximately 10 and the expected size is 1.58. However, if the penalty constant is increased to nine then we would require a sample of size five, but our expected size is now 2.30.

CHAPTER THREE

Hierarchical Generalized Linear Models and Subset Selection

3.1 Introduction

Originally introduced by Nelder and Wedderburn (1972), the *generalized linear model* (GLM) provides an extension to ordinary regression analysis by allowing the response variable to be non-Gaussian. In the classical linear model we typically specify the error term as a Gaussian random variable whereas in the case of the generalized linear model we model the responses $Y_1, \ldots, Y_n$ directly and assume the means $\mu_1, \ldots, \mu_n$ satisfy some specific *p*-dimensional function $g\left(\mu_i\right) = x^T \beta$. Clearly the GLM provides a unifying class of statistical models that generalizes classical linear models. Gelfand and Ghosh (2000) comment the GLM "avoid having to select a single transformation of the data to achieve the possibly conflicting objectives of normality, linearity, and homogeneity of variance." Since their inception GLM's have been used in a wide range of applications including but not limited to analysis of multicategory data (Leonard and Novick (1986)), dynamic or state space extensions of non-normal time series and longitudinal data, discrete time survival data, and non-Gaussian spatial processes (Best, Ickstadt, and Wolpert, 2000; or Banerjee, Carlin, and Gelfand (2004)). Moreover, the GLM is widely used in Poisson regression, which we provide as an example in the final section of this chapter.

In the first three sections of this chapter we provide the reader with key references to the development of both the Bayesian and classical GLM, as well as a comprehensive overview of the mathematical components of both the Bayesian and classical GLM. For the case of the Bayesian GLM as presented in Section 3.4, we discuss the types of prior distributions that are typically used in the GLM. Continuing in the Bayesian framework, Section 3.5 is devoted to a special type of Bayesian GLM which we will refer to as a *conjugate generalized linear model*. Finally, in Section 3.6 we will construct several Bayesian generalized linear models for home run hitters and then apply the previously developed subset selection procedure to determine the hitter with the highest home run hitting rate.

3.2 Literature Review

Brad Carlin once said that “perhaps the single most important contribution of statistics to the field of scientific inquiry is the general linear model” (Carlin and Louis (2000)). Perhaps this would explain the well-developed literature pertaining to the GLM. For an introductory exposition on the classical GLM the reader is referred to the texts by McCullagh and Nelder (1989); Fahrmeir and Tutz (1991); and McCulloch and Searle (2001). Whereas the above-mentioned texts provide an adequate collection of estimation and hypothesis testing procedures for various parameters in the GLM setting, the SAS help file gives a detailed discussion on how to obtain various statistics and model checking diagnostics for the GLM. As is the case of classical linear models, several authors have extended the GLM to include for latent variables (random effects) in which case we have generalized linear mixed models (GLMM’s). Breslow and Clayton (1993) laid the framework for the concept of GLMMs while Zhao, Staudenmayer, Caoull, and Wand (2004) have even developed Bayesian generalized linear mixed models, which are ultimately a special case of what are commonly referred to as *hierarchical generalized linear models*. For an introductory text that gives a development of the hierarchical generalized linear model the reader is referred to Gelman, Carlin, Stern, and Rubin (2004). For an advanced discussion of the hierarchical generalized linear model see Mallick, Dey, and Ghosh (2000). Besides these texts the reader will find a discussion of the hierarchical generalized linear model in West (1985) and Albert (1988). Both of these examine some of the theoretical and computational issues pertaining to the hierarchical generalized linear model, but for a further discussion of the various proposed priors and methods for their implementation see Ibrahim and Laud (1991), Dellaportas and Smith (1993), and Ghosh, Natarajan, Stroud, and Carlin (1998).

3.3 Classical Generalized Linear Models

The Generalized Linear Model (GLM) is characterized by three components: the random component associated with the response variable Y_i, a systematic component related to the explanatory variables used in the predictor function, and a link function that specifies the function of $E(Y)$. In the formulation of a GLM it is tacitly assumed that the underlying sampling distribution of the response variable Y_i is a member of the exponential family and that conditioned on θ_i the responses Y_i are independent. Generally, a member of the exponential family has a density function of the form

$$f\left(y_i \mid \theta_i, \phi\right) = \exp\left\{a^{-1}\left(\phi_i\right)\left[y_i \cdot \theta_i - b\left(\theta_i\right)\right] + c\left(y_i, \phi_i\right)\right\}, \tag{3.1}$$

where the θ_i are unknown, but the $a\left(\phi_i\right) > 0$ are known. The parameters θ_i and ϕ_i in (3.1) are commonly referred to as the canonical and dispersion (scale) parameters respectively. It has been shown (McCullagh and Nelder (1972)) that the mean and variance of a random variable having density (3.1) are related to its canonical and shape parameters by $\mu_i = E(Y_i \mid \theta_i) = b'(\theta_i)$ and $Var(Y_i \mid \theta_i) = b''(\theta_i)a(\phi_i)$. These results can be easily derived by making use of the fact that

$$E\left[\frac{\delta \log f(Y_i \mid \theta_i)}{\delta \theta_i}\right] = 0$$

and

$$E\left[\frac{\partial^2 \log f(Y_i \mid \theta_i)}{\partial \theta_i^2}\right] + E\left[\frac{\delta \log f(Y_i \mid \theta_i)}{\delta \theta_i}\right]^2 = 0$$

Important special cases of the exponential family include the binomial distributions with success probabilities $\pi_i = \exp(\theta_i)/\left[1+\exp(\theta_i)\right]$, $a(\phi_i) = 1$ and the Poisson distributions with rates $\lambda_i = \exp(\theta_i)$ and scale parameter $a(\phi_i) = 1$

As mentioned, the two other components of the GLM are the systematic component and the link function. The systematic component of a GLM relates a vector $(\eta_1, \ldots, \eta_n)$ to the explanatory variables through a linear model. Let x_i be a known $p \times 1$ vector of regression coefficients and β a vector of unknown regression parameters. Then each component of $\eta = (\eta_1, \ldots, \eta_n)$ is $\eta_i = x_i^T \beta, \quad i = 1, \ldots, n.$

Finally, the link function is what ‘links’ the systematic component to the random components through $\eta_i = g(\mu_i)$ where g is a known monotonic differentiable function. It follows that the function g links the mean, $E(Y_i \mid \theta_i)$ to the explanatory variables through the formula $g(\mu_i) = x_i^T \beta, \quad i = 1, \ldots, n.$ If $g(\mu) = \mu$, as in the case of ordinary regression with normally distributed Y_i, then we say g is the identity link and $\eta_i = \mu_i$. Moreover, in the case where the canonical parameter is equal to the systematic component i.e. $\theta_i = g(\mu_i) = x_i^T \beta$, we say that g is a *canonical link*. Examples of their use are found in Poisson Loglinear models and Binomial Logit models. Alternatively, since the mean is a function of the canonical parameter; recall, $E(Y_i \mid \theta_i) = b'(\theta_i)$, some authors such as Gelfand and Ghosh (2000) find it more convenient to generalize (3.1) by expressing the canonical parameters as some function h of the inner product $x_i^T \beta$.

That is,

$$f(y_i \mid \theta_i) = \exp\left\{a^{-1}(\phi_i)\left[y_i h\left(x_i^T \beta\right) - b\left(h\left(x_i^T \beta\right)\right)\right]\right\} c(y_i, \phi_i) \ , \tag{3.2}$$

where h is a strictly increasing, and a sufficiently smooth function. Now having expressed the likelihood in terms of the covariates β we can find estimates of these parameters and in turn estimate the means for certain levels.

The classical estimation procedure for GLMs is maximum likelihood where the dispersion parameters ϕ_i are assumed known and the design matrix $X = (x_1, \cdots, x_N)$ has rank p. The likelihood function is

$$L(\beta) \propto \exp\left[\sum_{i=1}^{n} a^{-1}(\phi_i)\left\{y_i h\left(x_i^T \beta\right) - b\left(h\left(x_i^T \beta\right)\right)\right\}\right]. \tag{3.3}$$

Taking the partial of equation (3.3) with respect to the vector β gives the score vector

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} a^{-1}(\phi_i)\left\{y_i b'\left(h\left(x_i^T \beta\right)\right)\right\} h'\left(x_i^T \beta\right) x_i, \tag{3.4}$$

and the Fisher information matrix is

$$I(\beta) = -E\left[\frac{\partial^2 \log L}{\partial \beta \partial \beta^T}\right] = X^T D V(\beta) \Delta^2(\beta) X, \tag{3.5}$$

where $D = Diag\left(a^{-1}(\phi_1),\cdots,a^{-1}(\phi_n)\right), V(\beta) = Diag\left(b''\left(h\left(x_1^{\ T}\beta\right)\right),\cdots,b''\left(h\left(x_n^{\ T}\beta\right)\right)\right)$,

and $\Delta(\beta) = Diag\left(h\left(x_1^{\ T}\beta\right),\cdots,h\left(x_n^{\ T}\beta\right)\right)$.

Typically the maximum likelihood estimate $\hat{\beta}$ is found by using some iterative procedure such as Newton-Raphson or Fisher Scoring and then a goodness-of-fit statistic for the model (Nelder et al. (1972); Agresti (2002)) is computed. Furthermore, Lehmann (1998) gives regularity conditions for which $\hat{\beta}$ is asymptotically $N\left(\beta, n^{-1}I^{-1}(\beta)\right)$, which in turn provides the basis for most test statistics and confidence intervals.

3.4 Bayesian Generalized Linear Models

In the Bayesian paradigm a model having likelihood of the form (3.2) would require a prior for the unknown regression parameters. Albert (1988) in his conjugate GLM assigns non-informative priors to the regression and scale parameters. Gelfand and Ghosh (2000) mention that a commonly used choice for β is the multivariate normal; that is, $\beta \sim N(\beta_0, \Sigma)$, where β_0 and $\sum$ are known. Assuming a multivariate normal for β and taking the product of the likelihood and prior leads to a posterior of the form

$$\pi(\beta \mid y) \propto \exp\left\{\sum_{i=1}^{n} a^{-1}(\phi_i)\left[y_i h\left(x_i^{\ T}\beta\right) - b\left(h\left(x_i^{\ T}\beta\right)\right)\right] - \frac{1}{2}(\beta-\beta_0)^T \Sigma^{-1}(\beta-\beta_0)\right\}. \quad (3.7)$$

The normalizing constant for (3.7) would need to be found using numerical integration thus making the posterior analytically intractable. In fact, there is no closed form expression for either the posterior mean or variance. However, we can use a numerical integration technique such as importance sampling to calculate these numbers or we can use a Markov Chain Monte Carlo (MCMC) method such as the Metropolis-Hastings algorithm or the Gibbs sampler to generate samples from the posteriors. In the complete or partial absence of prior information the experimenter may choose to use a noninformative prior, thus making the posterior distribution proportional to the likelihood. Under these circumstances any Bayesian analysis will be similar to a likelihood analysis. However, the use of uniform priors could result in an improper posterior (see e.g. Ibrahim and Laud (1991)).

To complete the hierarchy of the GLM one would need to assign a prior distribution for the unknown covariance matrix. It is mentioned in Gelfand and Ghosh (2000) that one option is to use an inverse Wishart distribution for the unknown covariance matrix Σ, symbolically $\Sigma \sim IW\left(\Psi,\nu\right)$. Specifically, the prior on Σ would have the functional form, $\pi\left(\Sigma\right) \propto \exp\left\{-\frac{1}{2}tr\left(\Psi\,\Sigma^{-1}\right)\right\}\left|\Sigma\right|^{-\frac{1}{2}\nu}$. Taking the product of the likelihood (3.7) and the prior for Σ gives a posterior

$$\pi\left(\beta,\Sigma\,|\,y\right) \propto \exp\left\{\sum_{i=1}^{n} a^{-1}(\phi_i)\left[y_i h\left(x_i^T\beta\right)-b\left(h\left(x_i^T\beta\right)\right)\right]\right\}$$
$$\times\left|\Sigma\right|^{-\frac{\nu+1}{2}}\exp\left[-\frac{1}{2}tr\left\{\left[(\beta-\beta_0)(\beta-\beta_0)^T+\Psi\right]\Sigma^{-1}\right\}\right]. \qquad (3.9)$$

There is no closed expression for the normalizing constant which makes (3.9) analytically intractable. Any posterior analysis will need to be handled through numerical integration. Gelfand and Smith (1990) have shown that Gibbs sampling proves to be a useful technique for generating samples from (3.9). To implement this procedure one would require the full conditional distributions, i.e. $\pi(\Sigma\,|\,\beta,y)$ and $\pi(\beta\,|\,\Sigma,y)$. The full conditional for Σ can be easily derived by mere inspection of (3.9); that is, $\Sigma\,|\,\beta,y \sim IW\left(\Psi+(\beta-\beta_0)(\beta-\beta_0)',\nu+1\right)$. However, $\pi(\beta\,|\,\Sigma,y)$ turns out to be a nonstandard distribution and is known only up to a multiplicative constant where

$$\pi(\beta \mid \Sigma, y) \propto \exp\left\{ \sum_{i=1}^{n} a^{-1}(\phi_i)\left[y_i h\left(x_i^T \beta\right) - b\left(h\left(x_i^T \beta\right)\right)\right] - (\beta - \beta_0)' \Sigma^{-1} (\beta - \beta_0) \right\}.$$

In this case we would need to use a procedure such as adaptive rejective sampling to assist the Gibbs sampler. The reader should be reminded that the canonical parameter appearing in the hierarchy is a parametric function of the unknown regressors and is not assigned a prior distribution. Albert (1988), on the other hand, takes a much different approach to the Bayesian GLM by assigning independent conjugate priors to the canonical parameters and in turn models the prior means. In Section 3.5 we give an overview of Albert's conjugate GLM.

3.5 Conjugate Generalized Linear Models

Again, suppose that conditioned on θ_i the random variables Y_i are independent belonging to the exponential density (3.1). Instead of modeling the canonical parameters as $\theta_i = h\left(x_i^T \beta\right)$ first stage, Albert (1988) considers them independent with conjugate density

$$\pi(\theta_i \mid m_i, \lambda_i) = \exp\left\{ \lambda_i \left[m_i \theta_i - b(\theta_i) + k(m_i, \lambda_i) \right] \right\}. \tag{3.10}$$

Assuming regularity conditions hold, $0 = \frac{\partial}{\partial \theta_i} \int \pi(\theta_i \mid m_i, \lambda_i)\, d\theta_i = \int \frac{\partial}{\partial \theta_i} \pi(\theta_i \mid m_i, \lambda_i)\, d\theta_i$,

or $\int \left(\lambda_i m_i - \lambda_i b'(\theta_i)\right) \pi(\theta_i \mid m_i, \lambda_i)\, d\theta_i = 0$. Thus the hyperparameter m_i is the prior mean of the sampling mean μ_i, i.e., $m_i = E\left[b'(\theta_i)\right] = E[\mu_i]$. Returning to (3.10), the hyperparameter λ_i is a precision parameter that reflects the strength of information regarding the prior means m_i and $k(m_i, \lambda_i)$ is a normalizing constant. Albert (1989) notes that as λ_i approaches infinity the prior density (3.10) becomes concentrated about m_i. Now, to obtain the posterior distribution we take the product of the likelihood (3.1) and the conjugate density (3.10), in which case the conditional distributions of θ_i given hyperparameters λ_i and m_i are independent having posterior density

$$\pi(\theta_i \mid y_i, m_i, \lambda_i) = \exp\left\{(\lambda_i + \phi_i)\left[m_i(y)\theta_i - b(\theta_i) + k\left(m_i(y), \lambda_i + \phi_i\right)\right]\right\}, \qquad (3.11)$$

where the posterior mean of μ_i is $m_i(y) = (y_i\phi_i + m_i\lambda_i)/(\phi_i + \lambda_i)$ and $y = (y_1, \ldots, y_n)$. Furthermore, if we combine the likelihood (3.1) with the prior density (3.10) and integrate over the range Θ we obtain the marginal or unconditional density for y_i:

$$\begin{aligned} f(y_i \mid m_i, \lambda_i) &= \int_{\Theta} f(y_i \mid \theta_i)\pi(\theta_i \mid m_i, \lambda_i)\, d\theta_i \\ &= \frac{c(y_i, \phi_i)k(m_i, \lambda_i)}{k\left(m_i(y_i), \lambda_i + \phi_i\right)}. \end{aligned} \qquad (3.12)$$

To make this model in fact a GLM we assume that the set of prior means $\{m_i\}$ satisfy the model $g(m_i) = x_i^T\beta$. Unlike Gelfand and Ghosh (2000) who use relatively informative priors for their Bayesian GLM, Albert (1988) considers a noninfomative prior for β and $\lambda = \lambda_i$ for all i in the form $\pi(\lambda, \beta) \propto (1+\lambda)^{-2}(\lambda > 0)$, implying that β is uniform. Christiansen and Morris (1997) and Albert (1989) suggest that the advantage of such an exponential mixture that contains a scale parameter is the ability to handle extra variability or overdispersion. We will explore the use of this model in the next section.

3.6 A Bayesian Analysis of Home Run Hitters

Unquestionably, three of the most prolific home run hitters of the 21st century are all-stars Mark McGuire, Sammy Sosa, and Barry Bonds. It is worthwhile to compare these three hitters as they are similar in physical size, age, and career span. In his paper, “A Bayesian Analysis of a Poisson Random Effects Model for Home Run Hitters,” Albert (1992) develops three models used in ranking the true home rates or “true” rates of 12 of the greatest home run hitter’s pre-1992. Before we begin the development of any hierarchical generalized linear model suggested in the work, it is worth mentioning that there exist several examples in the literature of Bayesian Poisson regression. In an example taken from Lindley (1965), El-Sayyad (1973) used a Bayesian GLM to model the counts of triplets born in Norway between the years 1911 and 1940. The counts were assumed to be distributed Poisson (λ_i) with the usual loglinear link expressed in terms of the interval of time at which the counts were recorded, i.e. $\log(\lambda_i) = \beta_o + \beta_1 time_i$. A Jeffreys’ prior was used for the unknown regression parameters. In another example appearing in Gelman et al. (2004), the authors use a hierarchical Poisson regression to analyze police stops in New York City.

3.6.1 Poisson Sampling Model

We consider the number of home runs hit per season, which we will denote z_i, as being distributed binomial with parameters t_i (the total number of at-bats for season *i*) and *p* being the probability of hitting a home run during a give plate appearance. Since *p* is small, we can approximate the distribution of z_i by a Poisson distribution with mean $t_i\gamma$, where γ is the player’s true home run rate over his career. If y_i represents the proportion of home runs hit in t_i at-bats for a particular player, then the sampling density of y_i is given by

$$f\left(y_i \mid \gamma\right)=\frac{e^{-t_i\gamma}\left(t_i\gamma\right)^{t_iy_i}}{\left(t_iy_i\right)!}, \qquad t_iy_i=0,1,2,\ldots \tag{3.13}$$

If the observed proportions $y_1,\ldots,y_N$ for each player are assumed independent, then the likelihood is given by

$$L\left(\lambda \mid y\right)\propto e^{-\gamma\sum t_i}\gamma^{\sum t_iy_i} \tag{3.14}$$

If γ is assigned the Jaynes prior, $\pi\left(\gamma\right)=\gamma^{-1}$, then the resulting posterior is of the gamma form with shape parameter $\alpha=\sum t_iy_i$ and scale parameter $\varphi=\sum t_i$. Bratcher and Stamey (2000) mention that the Jaynes prior can be considered the limiting form of a non-informative proper conjugate gamma prior, i.e. $\gamma\sim\text{Gamma}\left(0,0\right)$. The resulting posterior mean and standard deviation are $E\left[\gamma \mid y\right]=\sum t_iy_i\big/\sum t_i$, $SD\left[\gamma \mid y\right]=\sqrt{\sum t_iy_i}\big/\sum t_i$, respectively. Note that the posterior mean is just ratio of career home runs to career at-bats.

3.6.2 Poisson Conjugate Model

As was the case in Albert (1992) the Poisson model with a noninformative prior is not a good fit to the home run data pertaining to the these three batters. The model assumes that $E\left[y_i \mid \gamma\right]=var\left[\sqrt{t_i}\,y_i \mid \gamma\right]=\gamma$; however when comparing the sample mean of $\left\{y_i\right\}$ versus the sample variance of $\left\{\sqrt{t_i}\,y_i\right\}$ for each batter it is clear that the variability of the home run data is much higher than predicted by the Poisson model. Table 5 gives a summary.

Table 5. Point Estimates for the Hitters

	Career HR	Career AB	HR/AB	Mean	Variance
McGuire	583	6187	0.09423	0.0938	0.5821
Bonds	703	9098	0.07727	0.0797	0.4016
Sosa	574	8021	0.07156	0.0663	0.57

The column entitled 'Mean' gives the sample mean of $\{y_i\}$, whereas the column entitled 'Variance' gives the sample variance of $\left\{\sqrt{t_i}\, y_i\right\}$. The reader will find that the variance estimate is much higher than the rate estimate, giving the indication that there may be overdispersion. Correcting for this overdispersion we use a mixture distribution having an additional scale parameter as in Albert (1989). In addition we do not assume that the rate parameter in Poisson is constant but allow this rate parameter to differ for each year. Expressing (3.13) as an exponential of the log likelihood we arrive at a density of the form (3.1).

$$
\begin{aligned}
f\left(y_i \mid \gamma_i\right) &= \exp\left\{\log\left(\frac{e^{-t_i\gamma}\left(t_i\gamma_i\right)^{t_i y_i}}{\left(t_i y_i\right)!}\right)\right\} \\
&= \exp\left\{t_i\left[y_i \log t_i\gamma_i - \gamma_i\right] - \log\left(t_i y_i\,!\right)\right\} \\
&= \exp\left\{\phi_i\left[y_i \cdot \theta_i - b\left(\theta_i\right)\right] - c\left(y_i, \phi_i\right)\right\},
\end{aligned}
$$

where the canonical parameter θ_i is equal to $\log\gamma_i$, the cumulant function is $b\left(\theta_i\right) = \exp\left(\theta_i\right)$, and the scale parameter is t_i. According to (3.10) the prior distribution for θ_i is $\pi\left(\theta_i \mid m, \lambda_i\right) \propto \exp\left\{\lambda_i\left(m\theta_i - e^{\theta_i}\right)\right\}$, relating this to the Poisson rate gives $\gamma_i \mid \lambda_i, m \sim \text{Gamma}\left(\lambda_i m, \lambda_i\right)$. In this setting, *m*, the mean of the marginal distribution, would represent a batter's true rate and λ_i is an additional season-specific parameter, originally suggested by West (1985), that can model the extra variation in the data. The scale parameter is particularly important as it is used to adjust for the aberrant seasons and identify the seasons that are inconsistent with the main body of data. To make this a hierarchical model we assign a uniform distribution to the prior mean and a chi squared distribution (see e.g. West, 1985) to the scale parameter. The hierarchical Poisson regression can be summarized as

$$\begin{aligned} &y_i \mid \gamma_i \sim Poisson\left(t_i \gamma_i\right) \\ &\gamma \mid m, \lambda_i \sim gamma\left(m\lambda_i, \lambda_i\right) \\ &m \sim Uniform\left(.01, .15\right) \\ &\lambda_i \sim \chi^2\left(p\right). \end{aligned}$$

The reader will notice that the degrees of freedom for the chi squared distribution are left unspecified. As previously mentioned, the shape parameter appearing in the conjugate prior is used to downweight the inconsistent home run years thus making the degrees of freedom potentially very influential. Table 6 gives insight to the sensitivity of the posterior analysis of the true rate *m* with respect to this scale parameter.

Table 6. Approximate Posterior Moments of the True Rate

	Chi squared (10)			Chi squared (30)			Chi squared (100)		
	Mean	SD	PIC	Mean	SD	PIC	Mean	SD	PIC
McGuire	0.1143	0.01631	0.71716	0.09996	0.01095	0.88004	0.09469	0.00696	0.9757
Bonds	0.09887	0.01448	0.22348	0.08293	0.00861	0.10948	0.07755	0.00537	0.02424
Sosa	0.08607	0.01464	0.0606	0.07096	0.00877	0.01076	0.06761	0.00549	0.00064

It is evident from Table 6 that as the scale parameter λ increases the posterior standard deviations for *m* decrease and the differences between the probability inclusion criterions (the PICs) increase. This separation among the hitters' rates is due in part to the down weighting of the inconsistent years as seen in the posterior mean of Sammy Sosa's true rate. Sosa's observed rate was found to be .071 (Table 5) but using the hierarchical model produces an estimate of .067. As for the actual selected subset, using a penalty constant $c = 2$ we conclude that McGwire is the 'best' hitter; each subset contains only McGwire. Restricting our attention to the models with a larger scale parameter gives a more robust selection procedure as the choice of a larger penalty constant has little effect.

Although the hierarchical conjugate model is useful in modeling the rates of each hitter it is generally believed that a player's hitting ability increases until about the middle of his career, reaches a peak, and declines towards the end of his career. This idea is well supported by the career of Sammy Sosa (see, Table 7).

Table 7. Sosa's Career Total Home Runs and At-bats

At-bats (AB)	Home runs (HR)	Rate (HR/AB)
183	4	0.021858
532	15	0.028195
316	10	0.031646
262	8	0.030534
598	33	0.055184
426	25	0.058685
564	36	0.06383
498	40	0.080321
642	36	0.056075
643	66	0.102644
625	63	0.1008
604	50	0.082781
577	64	0.110919
556	49	0.088129
517	40	0.077369
478	35	0.073222

It appears as if his career rates can be described with a quadratic having vertex around his 10th season. With this in mind, we provide an example of a more realistic model where the true rate for the i^{th} year satisfies a log-linear model.

3.6.3 Quadratic Linear Model

In the previous model we assumed that a player's true rate was constant; however, in order to adequately represent a player's maturation we will use a quadratic log-linear model to represent the yearly rates. In our hierarchical conjugate model we assume that the true rate for the i^{th} year can be expressed as $\log m_i = \beta_0 + \beta_1 i + \beta_2 i^2$. The proposed hierarchical model is similar to Albert (1992) in that we assign independent uniform distributions to the regression coefficients, but rather than assigning a noninformative prior to the scale parameter we assign it a Chi squared distribution. In summary,

$$\begin{aligned}
&y_i \mid \gamma_i \sim Poisson\left(t_i \gamma_i\right)\\
&\gamma_i \mid m_i, \lambda_i \sim gamma\left(m_i \lambda_i, \lambda_i\right)\\
&\lambda_i \sim \chi^2\left(p\right)\\
&\log\left(m_i\right) = \beta_0 + i\beta_1 + i^2\beta_2\\
&\beta_0, \beta_1 \sim Uniform\left(-1,1\right) \text{ and } \beta_2 \sim Uniform\left(-6,6\right)
\end{aligned}$$

We contrast this model with two other hierarchical models: a conjugate model having no quadratic systematic component, and a hierarchical GLM with a quadratic systematic component and the usual canonical link. The first of the two alternatives is identical to the conjugate model presented in the previous section whereas the latter is similar to both Albert's (1992) quasilikelihood quadratic model and Gelfand & Ghosh's (2000) hierarchical GLM. The hierarchical GLM is as follows:

$$y_i \mid \gamma_i \sim Poisson\left(t_i\gamma_i\right)$$
$$\log\left(\gamma_i\right) = \beta_0 + i\beta_1 + i^2\beta_2$$
$$\pi\left(\beta_0,\beta_1,\beta_2\right) \propto MVN\left(0,1000\cdot \mathrm{I}\right)$$

Using WinBUGS we compute the Deviance Information Criterion (DIC) for each model and find the 'better-fitting' model to be the conjugate hierarchical model with a quadratic systematic component; however, the DIC is sensitive to the degrees of freedom in the Chi squared distribution. The model with the second smallest DIC was the non-conjugate hierarchical quadratic, which means that the hierarchical conjugate model with no quadratic systematic component has the highest DIC. The ability of the conjugate hierarchical model to outperform the non-conjugate hierarchical model is not surprising since the former contains a scale parameter to dampen the effects of the inconsistent years. Furthermore, the magnitude of the scale parameter also reflects the strength of one's prior belief about the means. Now we will use the conjugate hierarchical quadratic model to construct a subset to determine which year was Sosa's 'best'.

Table 8. Probability Inclusion Criterions for Sammy Sosa Years 1998-2001

Year	PIC
1998	0.0168
1999	0.21602
2000	0.35475
2001	0.41917

Comparing these PIC's to the criterion constant $1/(c+1)$ when $c=3$we have a subset containing the years 2000 and 2001. Recall that Sosa's observed home run was highest in 2001 but he hit his most home runs in 1998. One explanation for the PIC being so low in 1998 is that in the previous year his home run hitting rate was only 5%

and the large value of 66 is assumed to be a result of sampling variability. If we increase the penalty constant to $c = 4$ then we would include 1999. In that year he hit at the same rate as in the previous year when he hit his most home runs.

CHAPTER FOUR

Disease Mapping Basics

4.1 Introduction

Cartography, the science of map-making, dates back over 5000 years ago to the ancient Egyptian and Mesopotamian civilizations. These ancient civilizations used detailed sketches of their land for displaying human activity. According to Walter (2000) there are two main areas of cartography: *general* and *thematic*. The intent of general cartography is to provide maps of several geographic phenomena (e.g. a globe). Thematic cartography aims to create maps that display spatial patterns and or spatial relationships between certain phenomena. Thematic cartography, began somewhere around the early 19th century with the first maps being case event maps that displayed the locations of crime, weather related events, and eventually infectious diseases. The maps displaying the locations of cases of infectious diseases were eventually coined *disease maps*.

The disease maps of the 19th century were mostly used for displaying the geographic distribution and spatial patterns of infectious diseases (e.g. yellow fever in New York or cholera in London). Dr. John Snow (1855) used *spot maps* (case event maps) to reveal how cholera was spread through contaminated water sources in London. Generally, the purpose of these spot maps is to reveal any clustering of a disease, unlike the disease maps created in the latter part of the 19th century that are more etiological in purpose and contained information regarding chronic non infectious diseases like cancer. Haviland (1875) is considered to be one of the early pioneers of disease maps for cancer and heart disease. Using the census data for London and Wales Haviland was able to account for population exposure. He incorporated this population exposure by calculating adjusted regional rates of these chronic diseases. In essence he was trying to assess the geographic distribution of what would later be called ‘disease risk’ which is the focus of most of today’s disease atlases. We will revisit the concept of adjusted rates in a later section with a presentation of two standard methods, internal and external.

In Sections 4.2 and 4.3 of this chapter we will give an overview of the types of graphical representations used in disease mapping, provide the reader with a summary of classical and non-parametric techniques used in assessing risk, and furnish the reader with an extensive background of the current Bayesian methods used in disease mapping. Continuing on in Section 4.4 we will compare and contrast some of the current Bayesian spatial models appearing in the literature with a new class of conjugate spatial models via an example pertaining to the elevated risk of lip cancer in Scottish counties. Finally, in Section 4.5 we will extend many of the basic concepts of disease mapping to assess elevated crime rates. Our crime data set consists of counts of habitat burglary 911 calls in the Waco community for the year 2000. We will construct a model using covariates obtained from the United States Census Bureau and use that model in conjunction with the previously presented subset selection procedure to assess which region has the highest rate of habitat burglaries.

4.2 Disease Mapping Basics

Succinctly, the goal of any map is to communicate data to some audience. However, the choices of what types of maps and which types of data to use merit great attention. The reason being is that we use these maps to give insight to the geographic variations of disease risks and to detect not only spatial clustering or non-random trends, but also putative sources of disease risk. Ultimately the choice of what map to use will be determined by the availability of data. In the next section we will briefly review some of the common types of disease maps and their purposes. For a thorough treatment we recommend Waller & Gotway (2004).

4.2.1 Types of Disease Maps

Typically disease maps are regional in that the area corresponding to the study population can be partitioned into finitely many smaller regions usually referred to as *counties* or sometimes *areal units*. The major distinction of disease maps follows from whether or not we have the exact locations of the events. If the exact locations or points are known, as in case-event data, we simply plot the locations using some type of symbol. The default symbol for most software packages is a filled dot. In such situations, the point map is often referred to as a dot map. Clearly these maps would be used to monitor the spread of infectious diseases and can also be useful in identifying potential point sources of disease outbreaks. As mentioned earlier, Dr. John Snow used a point map at the local street level to identify the putative source of cholera. Because these maps fail to take into account population density they can be quite misleading. After all, larger regions tend to more populated. To rectify this problem many epidemiologists use a graduated color map.

Suppose that the attribute values to be mapped are not locations but rather summaries associated with the actual areal units; this is usually the result of medical record confidentiality. Moreover these summaries are often the aggregate counts of disease occurrence for each region. There are several maps used to accommodate this type of data. One of which is the classed symbol map. For this map a symbol corresponding to a class or collection of attribute values is plotted in the center of each region. A slight variation of this map is the graduated symbol map where the same symbol (usually a circle) varies with the attribute value or class of attribute values. It is common for the class sizes associated with the attribute values to be dissimilar in magnitude making it necessary to use a proportional symbol map. In which case, the size of the symbol plotted for each is proportional to the magnitude of the class in which that region's attribute values falls into. These symbol maps although useful in some situations can be useless when the study area contains a plethora of regions; especially when the regions are very different in size and are very spatially related. By spatially related we mean regions that are closer tend to have similar attribute values. In these situations as well as others most would find color to be a better method of visual discernment than symbols. With that in mind it is no surprise that choropleth maps are the most common type of map used for the display of areal data.

In general a choropleth map uses different color and pattern combinations to depict different values of the attribute variable corresponding to each region. There are both classed and un-classed choropleth maps. The classed choropleth maps assign to each region a possibly non-unique color corresponding to one of finitely many non-overlapping intervals that are associated with a set of attribute values. Alternatively, for the unclassed choropleth maps each region is assigned a unique color among a continuum and no two regions share precisely the same color. The following figure is a classed choropleth map of standard morbidity ratios for the Scottish lip cancer data appearing in Clayton and Kaldor (1987).

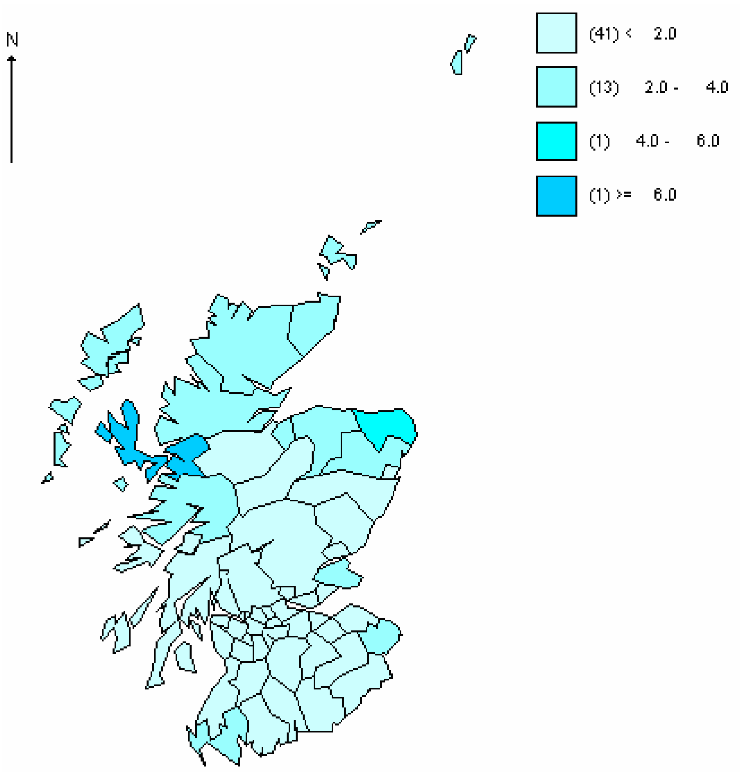

Figure1. Choropleth map of the Scottish lip cancer. The map gives the standard morbidity ratios (*SMR* = observed /expected) for the 56 counties of Scotland during the years 1975-1980. The darker shade of blue indicates a higher incidence rate. Note the higher incidence rate among the northern coastline. The higher incidence is attributable to the higher percentage of agriculture workers, namely fisherman.

Like most maps these choropleth maps certainly have their critics; Tukey once offered the advice, “Pray” (Tukey 1988, p.116). In fact many cartographers and epidemiologist find choropleth maps relatively crude. Particularly because when you represent a region with only one color you may fail to capture the changing disease risk, especially when the disease varies continuously over space. But there is a trade off between statistical stability of disease risk estimates and geographic resolution. We address these in Section 4.2.3.

For illustrative purposes (Figure 1) we chose to map the standardized morbidity/mortality ratios (SMR). This ratio of observed to expected purportedly assesses the geographic distribution of disease risk by indicating regions with elevated risks. In the next section we present the reader with alternative quantities used in disease mapping that are also useful in assessing disease risk.

4.2.2 Disease Risk

Waller and Gotway (2004) define term *disease risk* as the probability of a person contracting the disease within a specified time period. They go on to say that "disease risk is a dynamic, unobservable quantity that can be modified by characteristics such as age, gender, occupation, and diet" (Walter and Gotway 2004, p.9). These characteristics are called risk factors. The identification of risk factors is a central role in statistical epidemiology and is briefly addressed in Section 4.3 but for now we focus on what quantities to actually map. Having defined risk in terms of probability we are motivated to use statistical methods in the construction of choropleth maps. That being said, primitive choropleth maps were often maps displaying counts and proportions. As mentioned earlier a common belief is that larger urban regions tend to be more densely populated, therefore counts fail to capture the notion of elevated risk. Instead epidemiologists began to map rates and proportions. It must be noted that the term rate used in the epidemiological nomenclature refers to the ratio of the number of occurrences of a particular event (e.g. incidence of disease or mortality) per unit time; whereas, a statistical rate is the usual ratio of the number of occurrences to the number of people at risk. Of course one might be interested in other parameters besides average disease risk. For example, one could be interested in comparing the risks between individuals with and without a certain exposure to a disease or even different exposure levels. These *relative risks* are typically measured with risk ratios or risk differences. Incidentally, the statistical techniques used to estimate relative risks are similar to those used to estimate disease risk. We now proceed to one of the earliest methods used to estimate disease risk.

Assume that for each areal unit $i = 1, \ldots, n$ we have an associated count of the number of occurrences of a particular event denoted by Y_i as well as the total number of people at risk N_i. We could assume the data follows a binomial probability model, that is assume $Y_i \mid p_i \overset{ind.}{\sim} Bin\left(N_i, p_i\right)$ where p_i represents the probability of contracting the disease in area *i*. It is common knowledge that most non-infectious diseases affect people of certain ages and genders disproportionately. In order to make rates from different areal units comparable we would need to remove the effects of known risk factors by adjusting the rates accordingly. A very simple, yet straightforward way to correct for a known risk factor is to stratify. We could divide each areal unit into stratums $j = 1, \ldots, J$ so that we have $n \times J$ total units. We denote the associated disease risk for area *i* and stratum *j* as p_{ij} so that our data is now of the form

$$Y_{ij} \mid p_{ij} \overset{ind.}{\sim} Bin\left(N_{ij}, p_{ij}\right). \tag{4.1}$$

The classical estimation procedure for estimating the p_{ij}'s is maximum likelihood with the estimators taking on the form $\hat{p}_{ij} = Y_{ij} / N_{ij}$. This estimation procedure works well in the limiting case but for a rare disease the data is usually too sparse to get stable estimates of the p_{ij}'s. This leads to the proportionality assumption

$$\frac{p_{ij}}{\left(1 - p_{ij}\right)} = \theta_i \times \frac{p_j}{\left(1 - p_j\right)}, \tag{4.2}$$

so that the effect of being in area i is a product of each of the strata-specific *reference odds* $p_j/(1-p_j)$ and the common *odds ratio*, θ_i, for that area. The advantage of such an assumption (4.2) is that we have reduced the number quantities to estimate per area from J to 1. However, Wakefield, Best, and Waller (2000) mention that the proportionality assumption is strong and must be checked. They recommend plotting $\hat{p}_{ij}/(1-\hat{p}_{ij})$ versus $\hat{p}_j/(1-\hat{p}_j)$ for each area i to assess this proportionality.

A variety of parameters may be estimated in the binomial model (Wakefield, Best, and Waller, 2000). For example, the reference odds may be estimated simultaneously with the common odds ratios (e.g. Clayton 1996). Alternatively we can fix the reference odds via a reference set (external standardization) or use the overall odds for the study region (internal standardization), that is,

$$\frac{\hat{p}_j}{(1-\hat{p}_j)} = \frac{\sum_i Y_{ij}}{\sum_i (N_{ij} - Y_{ij})}.$$

Now that the $\hat{p}_j$ are treated as known quantities, the MLEs of the common odds ratios θ_i may be estimated via the logistic regression model

$$\text{logit}\, p_{ij} = \log\theta_i + \hat{\gamma}_j, \tag{4.3}$$

where the $\hat{\gamma}_j = \log\{\hat{p}_j/(1-\hat{p}_j)\}$ are known offsets. In model (4.3) the $\hat{p}_j$ are treated as known quantities thus (4.3) does not recognize the uncertainty in γ_j, which according to Wakefield, Best, and Waller (2000) this may be a problem if these quantities are not estimated from extensive data. In ecological regression and hypothesis generation studies we may wish to assume a GLM on the log of the common odds ratio i.e.

$$\log\theta_i = X_i^T\beta + \alpha\,, \tag{4.4}$$

where β is a $k\times 1$vector of regression coefficients. Breslow and Day (1987, Chapter 4) caution using internal standardization with known offsets (Equation (4.3)) since the *a* priori estimation of γ_j may remove some of the effect of the exposure X_i. For example, individuals of a certain race may tend to live in areas with large values of X_i.

Admittedly the structure of this binomial model (4.1)-(4.4) is seemingly elegant and readily analyzable in several statistical packages, but it suffers from one major flaw: its inability to adequately handle overdisperion in the data (i.e. $\operatorname{var}\left(Y_{ij}\right) > N_{ij}\,p_{ij}\left(1-p_{ij}\right)$. This overdispersion may have both spatial (*structured heterogeneity*) and non-spatial (*unstructured heterogeneity*) components and may arise from unmeasured risk factors. Also, compared to other potential models the binomial formulation is not computationally convenient since the aggregation of counts over the stratums ($Y_i = \sum_j Y_{ij}$) is not hardly recognizable. There are alternatives to the binomial model. Recall from Chapter 3 and Albert (1992) that in the case where p_{ij} is small (rare disease) we may approximate the binomial distribution (4.1) by the Poisson distribution$Y_{ij} \sim \text{Poisson}\left(N_{ij}\times p_{ij}\right)$. Typically we assume that the disease risk associated with area *i* and stratum *j* is proportional the disease risk over stratum *j*. Specifically,

$$p_{ij} = \theta_i \times p_j\,, \tag{4.5}$$

where θ_i corresponds to the relative risk of disease in area *i* with respect to the reference rate p_j in each stratum. Wakefield, Best, and Waller (2000) point out that a great advantage of the Poisson approximation is that, when combined with the proportionality assumption (4.5), we may collapse over strata to obtain

$$Y_i \sim \text{Poisson}\left(E_i \times \theta_i\right), \qquad (4.6)$$

where $Y_i = \sum_j Y_{ij}$ and $E_i = \sum_j N_{ij} p_j$ is the *expected number* of cases in area *i* with strata-specific reference rates p_j. Banerjee, Carlin, and Gelfand (2004, Chapter 5) refer to the use of reference rates in the calculation of the expected number of cases as *external standardization*. Alternatively, if no set reference rates are available Banerjee, Carlin, and Gelfand (2004) recommend computing the expected number of cases by using an overall rate computed from the data, i.e. $E_i = N_i \bar{r} \equiv N_i \left(\frac{\sum_i y_i}{\sum_i N_i} \right)$. They refer to this process as *internal standardization*. It is easily shown that the standard maximum likelihood estimate of the relative risk, θ_i under the Poisson assumptions on Y_i is the SMR, $\hat{\theta}_i = Y_i / E_i$. Clayton and Kaldor (1987) give a very technical discussion of why taken together, $\left\{\hat{\theta}_i, i = 1, \dots, n\right\}$ are not necessarily the best estimates of $\left\{\theta_i\right\}$. However, for a thorough but non-technical discussion of the drawbacks to using SMR as a measure of relative risk we refer the reader to Lawson (2003, chap. 1) or Waller and Gotway (2004, chap. 4). We give a few of those drawbacks. The most alarming consequence of using SMRS is that zero SMRs do not distinguish variation in the expected counts. Also we point out that SMRs being ratio-based estimators are notoriously unstable since the variability in the estimated rates depends on population size. Consequently, the rates corresponding to larger areal units will be better estimated than the rates corresponding to the smaller areal units, and this may obscure spatial patterns. Similarly, rates based on small populations tend to be artificially elevated due to the relatively small computed expected count appearing in the denominator. These inflated rates do not necessarily reflect an increase in relative risk but a lack of data. Waller and Gotway (2004) refer to this as the *small number problem*. There have been several methods proposed to overcome the small number problem. One solution to the small number problem is to aggregate the counts of smaller regions; however, we lose resolution and give up geographic information. Another solution is to use a comparative map, one that compares each rate or in the case of a probability map, the p-value associated with each rate to a common measure (see e.g., Choynawski, 1959). Finally we could use a technique similar to both scatter plot smoothing found in regression and weighted moving average methods used in time series called *spatial smoothing*.

4.2.3 Spatial Smoothing

There are an abundance of spatial smoothers appearing in the literature. Although some are relatively informal or nonparametric while others are just the antithesis, they share the same goal: produce more stable risk estimates than the usual MLEs or SMRs. According to Waller and Gotway (2004, p.87), “the basic idea is to ‘borrow’ information from neighboring regions to produce a better (i.e. more stable and ‘less noisy’ estimate) of the risk associated with each region and thus separate out the ‘signal’ (i.e. spatial pattern) from the noise.” The nonparametric smoothing techniques can be categorized as either interpolation methods or kernel regression. However, “when substantive hypotheses and/or greater amounts of prior information are available” (Lawson, Browne, and Rodeiro 2003, p.6), it may be appropriate to employ a model-based approach to estimating local relative risks. We will discuss the use of a model based procedure for spatial smoothing later in this section, but for now we consider a nonparametric approach called the *locally weighted average procedure*.

As its name suggests the locally weighted average procedure considers the observed counts of the *neighboring* regions to estimate local relative risks. Borrowing notation from Waller and Gotway (2004, chap. 4) the relative risk estimate for the i^{th} area takes on the general form $\tilde{r}_i = \sum_j w_{ij} r_j \Big/ \sum_j w_{ij}$. In the preceding statement r_j can be either the SMR or percentage of occurrence $\left(y_j / N_j\right)$ of the j^{th} neighbor of area *i* while the weights take on the form

$$w_{ij} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ share the same boundary} \\ 0 \text{ otherwise} \end{cases}.$$

In matrix form $\left\{w_{ij}\right\}$ constitute a *spatial proximity matrix* or *adjacency matrix*. Rather than using a common boundary as a neighborhood criterion we could have used the distance as measured by some metric (Euclidean, taxi cab, absolute value) between the centroids of each area. That is,

$$w_{ij} = \begin{cases} 1 \text{ if } d_{ij} < \varepsilon \\ 0 \text{ otherwise} \end{cases},$$

where d_{ij} represents the distance between the centroid of area *i* and the centroid of area *j*. The latter approach to smoothing is often called disk smoothing. The locally weighted average procedure presented requires no distributional assumptions for the data. If we are willing to assume some sort of parametric distribution then we may use a type of nonparametric regression procedure that utilizes kernel weights.

If $Y_1, \ldots Y_n$ are data collected from the probability distributions $f\left(y_i \mid \theta_i\right)$, then Brillinger (1990) provides a general method for deriving estimators by maximizing a weighted log likelihood function. That is, a set of estimators for $\{\theta_i\}$ is found by maximizing

$$L(\theta) = \sum_j w_{ij} f\left(y_j \mid \theta_j\right). \tag{4.6}$$

If the data are normally distributed with unknown means θ_i or distributed as Poisson with unknown rates θ_i then the estimates found from maximizing (4.6) take on the same form as the locally weighted averages, disregarding the actual weights. This speaks to the theoretical cohesion that is gained by using the above estimation procedure. As hinted above, what is dissimilar about the two smoothing procedures is the specification of the weights. In this semi-parametric procedure we use a kernel function in conjunction with the spatial locations of the centroids to aid in the calculation of the weights. Thus we define the weights to be

$$w_{ij} = \mathrm{kern}\left(\frac{s_i - s_j}{b}\right),$$

where the kernel function, $\mathrm{kern}(\cdot)$, is a bivariate probability density function that is symmetric about the origin and integrates to 1 over the domain. The parameter *b*, called the *bandwith*, controls the amount of smoothing.

We conclude this subsection with a somewhat theoretical justification for the use of Bayesian methodology in disease mapping. Also, we mention that the current literature is saturated with applications of Bayesian methods in public health data. Andrew Lawson (Lawson, 2001; Lawson, Biggeri, Böhning, Lesaffre, Viel, Bertollini, 2003) and Elliot, Wakefield, Best, and Briggs (2001) were some of the first to publish texts consisting of a compilation of articles that articulate the advantages of Bayesian models in disease maps while Lawson and Gotway (2004) soon followed with a self-contained introductory text.

We begin our motivation by assuming the data $Y_i \mid \theta_i$ are distributed as Poisson with means $E_i\theta_i$, where the unknown parameter θ_i represents the risk of disease for the i^{th} area. It should be noted that the concept of conditional independence implies that any spatial correlation observed in the data is a function of the unknown risks θ_i.

Our development is similar to Marshall (1991) but we use expected cases instead of population size to construct our estimates of relative risk. We define the prior mean and variance of θ_i to be $E\left[\theta_i\right] = m_{\theta_i}$ and $Var\left(\theta_i\right) = \sigma^2_{\theta_i}$. If $r_i \equiv Y_i / E_i$ then the expected rate is equivalent to the prior mean, that is $E\left[r_i \mid \theta_i\right] = \theta_i$ and the prior variance is proportional to the expected disease incidence/prevalence, i.e. $Var\left(r_i \mid \theta_i\right) = \theta_i / E_i$. Thus the prior mean and prior variance of the rates r_i are $E_r\left[r_i\right] = m_{\theta_i}$ and $Var_r\left(r_i\right) = \sigma^2_{\theta_i} + m_{\theta_i} / E_i$. According to Marshall (1991) the best linear Bayes estimator of θ_i derived from minimizing expected total squared-error loss is

$$\hat{\theta}_i = C_i r_i + \left(1 - C_i\right) m_{\theta_i}. \tag{4.7}$$

The constant C_i in Equation 4.7, which is called the *shrinkage factor*, is the ratio of prior variance to the data variance or $C_i = \sigma^2_{\theta_i} \Big/ \left(\sigma^2_{\theta_i} + m_{\theta_i} / E_i\right)$. When the data is sparse, i.e. E_i is negligible and $C_i \rightarrow 0$. If the prior variance is diminutive, the Bayes estimator converges to the prior mean. Alternatively, in the absence of prior information or when the prior variance for the relative risks is considerable the shrinkage factor approaches one, i.e. $C_i \rightarrow 1$ and the Bayes estimator shrinks towards the observed rate.

Before the advent of WinBUGS most investigators would have taken an *empirical Bayes* (EB) approach like that of Clayton and Kaldor (1987) where the prior mean and variance of the relative risks are estimated from the data and substituted back in the likelihood in order to derive the posterior distribution of the risks. To avoid over-specification Marshall (1991) assumed a global prior mean and global prior variance and used method of moments to find estimates for these parameters. Caution is warranted because the prior variance estimate could be negative. Count data often has excess variability not accounted for by the Poisson model. We account for this overdispersion by defining an additional structure for the *regional relative risks*. This in turn requires a prior structure for the mean and variance of the regional relative risks. Clayton and Kaldor (1987) used various procedures like maximum likelihood estimation and the EM algorithm to estimate these hyperparameters. The investigator could of course assign hyperpriors to these hyperparameters, consequently taking on a fully-Bayesian approach. The reader will find that the additional structure comprising the Bayesian framework offers a richer framework for modeling covariate effects and spatial correlation. Due its dominance in the literature we save discussion of the fully Bayesian approach used in estimating relative risks and disease mapping for the chapter.

CHAPTER FIVE

Hierarchical Bayesian Models for Disease Mapping

5.1 Introduction

Our discussion of hierarchical Bayesian modeling procedures used in disease mapping begins by assuming that the aggregate count for each areal unit is distributed as Poisson with unknown relative risk θ_i, that is $Y_i \mid \theta_i \sim \text{Poisson}\left(E_i\theta_i\right)$. At this point it is customary to either specify a joint distribution for the relative risks, similarly a joint distribution on some function of the relative risks, or we can assume that the relative risks satisfy some generalized linear model consisting of both fixed and random effects. Clayton and Kaldor (1987) were the first to specify a joint distribution of the relative risks. They used empirical Bayes (EB) to estimate the hyperparameters from the marginal distributions of the aggregate counts. There are a number of possibilities for the joint distribution. Tsutakawa, 1985 assumed a multivariate normal distribution for the logits of disease risk. Clayton and Kaldor (1987) pioneered the use of conjugacy. The use of the conjugate gamma distribution for the relative risks results in the data being distributed as negative binomial. This model provides straightforward estimates of the shape and scale parameters. Despite the simplicity of using a conjugate prior, Clayton and Kaldor (1987) preferred a multivariate log-normal distribution for the relative risks. The reason being is that the latter has the ability to incorporate spatial correlation. In fact, they even modeled the log relative risks with the *conditional autoregressive* (CAR) model originally suggested by Besag (1974). Wakefield, Best, Waller (2000) note the empirical Bayes methods suffer from a number of limitations. In particular, the estimate of relative risk fail to reflect the uncertainty associated with the various hyperparameters. Consequently, these estimates are over-precise. This would prompt Besag, York, and Mollié (1991) to later extend these CAR models to a fully Bayesian setting using Markov chain Monte Carlo algorithms.

5.2 Second Stage Hierarchical Generalized LinearModels for Disease Mapping

We begin with the fully Bayesian approach to spatial smoothing presented in seminal papers by Besag, York, and Mollié (1991) and later Clayton and Bernardinelli (1992). Intending to model the extra-variability usually present in areal unit data the authors utilized a second stage model incorporating both known covariates and random effects. Rather than explicitly assigning a distribution to the log relative risks, Besag et al. (1991) assumed that the log relative risks satisfied a hierarchical Bayesian generalized linear spatial model. Considering the first stage model on the data, $Y_i \mid \theta_i \sim \text{Poisson}\left(E_i\theta_i\right)$ where θ_i is the unknown relative risk for the i^{th} areal unit we have shown in Section 3.2 of Chapter 3 that the canonical link for the Poisson model is $\log\left(\theta_i\right)$. Since the relative risks must be positive it seems reasonable to use this canonical link in the formulation of the hierarchical generalized linear spatial model. In the works of Besag et al. (1995), Mollié (1996), Wakefield, Best, and Waller (2003) as well as numerous others (for a list of references see Ghosh, Natarajan, Waller, and Kim, 1999) the log relative risks are modeled as the sum of two random components and the inner product of a $k \times 1$ vector of explanatory variables X_i^T with a $k \times 1$ vector or regression coefficients β. Symbolically we have

$$\log\left(\theta_i\right) = X_i^T \beta + U_i + V_i, \tag{5.1}$$

where the sum of the last two terms is often referred to as a *convolution Gaussian prior* (Mollié, 1996). The last component in (5.1) is used to model any unstructured heterogeneity among the log-relative risks. In order to model the similarity of these contributions to log relative risks we usually assign an exchangeable Gaussian prior of the form$V \sim MVN\left(0, \sigma_v^2 I\right)$. Most investigators (Waller and Gotway 2004, p 412; Banerjee, Carlin, and Gelfand 2004, p. 164) will assign a somewhat vague inverse gamma hyperprior for the variance. Likewise, if we express the distribution of *V* in terms of precision (i.e. the reciprocal$\tau_v = 1/\sigma_v^2$) we assign a gamma distribution. As for the other random component $U_i$we cannot overemphasize the importance of its prior specification.

There are two common approaches to modeling the spatially structured contribution to the log-relative risks . We may specify the joint distribution of $U = \{U_i\}$, or we can assume that the set of full conditional distributions $U_i \mid U_j = u_j, j \neq i, i = 1, \ldots, N$ define a *Markov random field* (MRF). Besag (1974) reconciles the two approaches by exploiting *Brook's Lemma* (1964); however, it was Geman and Geman (1984) that provide the next critical step that allows us to use a Gibbs sampler to generate from the joint posterior of *U* uniquely determined by the full conditionals. Wakefield, Best, and Waller (2000, pp.110-114) parallel the two approaches for the case when *U* is multivariate normal and we have a Gaussian MRF.

Suppose that the joint distribution of the random spatial components is multivariate normal with zero mean vector and covariance structure$\sigma_u^2 \Sigma$, where Σ is an $N \times N$ correlation matrix. We use the familiar notation$U \sim MVN\left(0_N, \sigma_u^2 \Sigma\right)$. If we let

$Q = \Sigma^{-1}$ and define Q_{ij} to be the (i, j) entry of the matrix *Q* then using properties of the multivariate normal (e.g. Searle, Casella, and McCulloch, 1991) we can derive the set of full conditional distributions

$$U_i \mid U_j = u_j, j \neq i \sim N\left(\sum_{j=1}^{N} w_{ij} u_j, \sigma_u^2 D_{ii} \right), \tag{5.2}$$

Where $w_{ii} = 0$, $w_{ij} = -Q_{ij} / Q_{ii}$, and $D_{ii} = Q_{ii}^{-1}$. Besag (1974) refers to the specification in (5.2) as an *autonormal model*. It is interesting to note that if we set $w_{ij} = 1/(N-1)$ then conditional mean in (5.2) is an average of the $u_j, i \neq j$. Wakefield, Best, and Waller (2000) show that the two approaches are related through the relationship $Q = D^{-1}(I - W)$ where *D* is a diagonal matrix with diagonal elements D_{ii}. Clearly, once the elements of the correlation matrix Σ have been specified, as in the joint formulation, the investigator can produce the elements of the weight matrix *W* and the diagonal matrix *D*. Alternatively, if the investigator specifies *W* and *D* then the correlation matrix for *U* can be easily obtained from elementary matrix computations. It is worth mentioning that "convenient" choices of *W* and *D* can lead to a joint model that is not well defined either because *Q* is singular or Σ is not symmetric (see e.g. Banerjee, Carlin, and Gelfand 2004, pp.79-81). This has lead some to use the joint formulation or point-referenced model of the random components *U* instead of the conditional model approach.

The spatial components in (5.1) model extra-Poisson variability in the log-relative risks that varies "locally"; that is, areal units in close proximity will tend to have similar risks. The spatial dependencies among these risks may be modeled through the off-diagonal terms of the correlation matrix Σ. For example, we can assume that the spatial dependence between two areal units is a function of only the distance, d_{ij}, between the population-averaged centroids of areas *i* and *j*. The underlying stochastic process is said to be *isotropic* since the covariogram (covariance function) between areal units is a function of the length (d_{ij}) of the vector that separates any two units. Raftery and Banfield (1991) suggest a one-parameter exponential function that can be used to calculate the elements of $\sum$. Diggle, Tawn, and Moyeed (1998) discuss using a two-parameter family called the Matérn class (Matérn, 1986) that uses a modified Bessel function for calculating the correlations.

Banerjee, Carlin, and Gelfand (2004, p.162) remark that while such possible joint models for *U* may seem sensible, they turn out to be very difficult to fit even in the isotropic case, due to the number of matrix inversions required. Furthermore, intercentroidal distance may be appropriate when the areal units are of roughly equal size. However, it may make little sense, especially when dealing with very irregular spatial units. As a result, it is customary in most hierarchical analyses of areal unit data to adopt a conditional formulation of *U* that makes use of the same adjacency matrix presented with the locally weighted average spatial smoother back in Section 4.2.3. In fact, from a spatial perspective we would think that the full conditional distribution for U_i should only depend upon the neighbors of its associated area, *i*. Using the notation ∂_i to represent the adopted neighborhood structure (e.g., the one setting $w_{ij} = 1$or 0 depending of whether *i* and *j* are adjacent or not), we specify the full conditional distributions for the random components U_i such that $\pi\left(u_i \mid u_j, i \neq j\right) = \pi\left(u_i \mid u_j, j \in \partial_i\right)$. By this notation we mean that the full conditional for U_i is identical to the conditional distribution of U_i is conditioned only on the values of its neighbors. For the autonormal model in (5.2) this would suggest that the conditional mean of U_i is a linear combination of the spatial components of the neighboring areal units. Moreover, we can make the conditional mean a truly weighted average by using the weights of the previously defined adjacency matrix. That is, if we define the weights as $w_{ij} = 1$ if area *i* is adjacent to area *j*, and $w_{ij} = 0$ otherwise (by convention $w_{ii} = 0$, for all *i*) then the conditional distributions will be of the form $U_i \mid U_j = u_j, i \neq j \sim N\left(\frac{\sum_{j \neq i}^{N} w_{ij} u_j}{\sum_{j \neq i}^{N} w_{ij}}, \frac{D_{ii}}{\tau_u} \right)$. It is convenient to make the conditional variance proportional to the number of neighbors by setting $D_{ii} = 1 \Big/ \sum_{j \neq i}^{N} w_{ij}$ and to use the precision $\tau_u = 1 \big/ \sigma_u^2$. The following is referred to as an *intrinsic Gaussian autorgressive* structure

$$U_i \mid U_j = u_j, i \neq j \sim N\left(\frac{\sum_{j \neq i}^{N} w_{ij} u_j}{\sum_{j \neq i}^{N} w_{ij}}, \frac{1}{\tau_u \sum_{j \neq i}^{N} w_{ij}} \right). \tag{5.3}$$

We give (5.3) the notation $CAR(\tau_u)$.

It has been shown (Besag, 1974) that the set of conditional distributions in (5.3) uniquely defines a corresponding multivariate normal joint distribution; i.e. let $Q = D^{-1}\left(I - W^*\right)$ where element (i, j) of W^* is $1 \big/ (\text{number of neighbors})$ and D is previously defined. However, Waller (2002) states that the choice of weights in (5.3) leads to a singular precision matrix. To see this, note that the i^{th} row of $I - W^*$ sums to zero. Thus, Q has rank $n-1$ and is not invertible. Ultimately this means that the spatial similarity implied by the conditional distributions does not translate directly into a model of spatial correlation. It is also worth mentioning that the priors defined by (5.3) are improper since they only define contrasts between pairs of $U_i s$; however, through the inclusion of the data the posteriors will be proper (Banerjee, Carlin, and Gelfand 2004). This fact, along with the ability to use the Gibbs sampler, is what compels most investigators to simply ignore the improper nature of the CAR model. Also, to allow identifiability of an intercept in $X_i^T \beta$, one adds the constraint $\sum_{i=1}^{N} u_i = 0$. This constraint is easily imposed by recentering each sampled vector U about its own mean following each Gibbs iteration. Besag and Kooperberg (1995), Cressie (1993, pp.407-408, 410-423), and Banerjee, Carlin, and Gelfand (2004, pp. 163-165) provide detailed discussion of conditional autoregressive structures.

To make (5.1) a legitimate hierarchical Bayesian generalized linear spatial model we need to assign a third-stage prior to the precision parameter, and complete the hierarchy of the CAR model (5.3). In addition, we also need to assign a prior distribution to the vector of unknown regression coefficients. As in most hierarchical generalized linear models we assign an arbitrarily vague prior (improper uniform or normal with large variance) to the regression coefficients. However, we cannot simply follow suit and assign an arbitrarily vague prior to the precision parameter; after all, the precison parameter controls the amount of extra-variability allocated to the spatial component. Ghosh, Natarajan, Waller, and Kim (1999) discuss restrictions on parameters for these hyperpriors to ensure proper posteriors. However, a more important concern than proper posteriors is a 'fair' assignment of τ_u to avoid overemphasis on the role of the global or local risks. A question that is complicated by the conditional nature of τ_u.

We do not attempt to address the concern of prior suitability but instead propose a new model that makes use of conjugacy. A conjugate hierarchical spatial model will help remove the influence τ_u has in controlling spatial similarity by introducing a dispersion parameter. Furthermore, by including a dispersion parameter related to the relative risks we may be able to better quantify the extra-variability present in areal unit data. Ultimately we would be able to assess the regions with abnormal risk. This is done through analysis of the scale parameter.

5.3 Conjugate Hierarchical Generalized Linear Models for Disease Mapping

In this section we extend the conjugate approach to the hierarchical generalized linear model given by Albert (1988) to allow for the spatial correlation consistent in the areal unit data used in disease mapping. Just as before we assume that the aggregate count associated with each areal unit is distributed as Poisson with expected case E_i and local relative risk θ_i. It is easily shown that the Poisson distribution is a member of the exponential family. Rewriting the Poisson likelihood using the natural logarithm function gives

$$\begin{aligned} f\left(y_i \mid \theta_i\right) &= \exp\left\{\log\left(\frac{e^{-E_i\theta_i}\left(E_i\theta_i\right)^{y_i}}{\left(y_i\right)!}\right)\right\} \\ &= \exp\left\{y_i \log\theta_i - E_i\theta_i + y_i \log E_i - \log y_i\,!\right\} \\ &= \exp\left\{y_i \cdot \theta_i^* - b\left(\theta_i^*\right) - c\left(y_i, E_i\right)\right\}, \end{aligned}$$

where the canonical parameter θ_i^* is equal to $\log\theta_i$, the cumulant function is $b\left(\theta_i^*\right) = E_i e^{\theta_i}$. According to Albert (1988) the conjugate prior for the canonical parameter (log relative risk) is $\pi\left(\theta_i^* \mid m_i, \lambda_i\right) \propto \exp\left\{\lambda_i\left(m_i\theta_i^* - E_i e^{\theta_i^*}\right)\right\}$, which means that the relative risks are distributed as $\theta_i \mid \lambda_i, m_i \sim \text{Gamma}\left(\lambda_i m_i, E_i\lambda_i\right)$. A gamma parameterized in this way implies that $E\left[Y_i\right] = E_{\theta_i}\left\{E\left[Y_i \mid \theta_i\right]\right\} = m_i$, but more importantly implies that the prior mean is inversely proportional to the expected number of cases, i.e. $E\left[\theta_i\right] = m_i/E_i$. In lieu of this parameterization it makes more sense to parameterize the gamma in such a way that the first two prior moments are independent of the expected number of cases. Furthermore, the parameters should be specified in a manner that facilitates the use of a GLM to model the marginal relative risks. One proposed parameterization given by Clayton and Kaldor (1987) is a $\text{Gamma}\left(\lambda_i m_i, \lambda_i\right)$, the other due to Christiansen and Morris (1997), is a $\text{Gamma}\left(\lambda_i, \lambda_i/m_i\right)$. For each parameterization the mean is equal to m_i but the variances differ slightly. The first parameterization has a variance that is proportional to the prior mean, whereas the latter has a variance that is a quadratic function of the prior mean. Because of convergence reasons we choose to align our work with that of Christiansen and Morris (1997). Recall that the scale parameter λ_i helps measure the extra-variability in the Poisson model by making the variance of the relative risk inversely proportional to λ_i. Thus larger values of λ_i will shrink the corresponding relative risk towards its prior mean m_i. As in Albert (1988) and Clayton and Kaldor (1987) we may incorporate fixed effects through a log linear model on the prior mean; that is,

$$\log m_i = X_i^T\beta, \tag{5.4}$$

where X_i^T is a $k \times 1$ vector of explanatory variables and β is a $k \times 1$ vector of regression coefficients. Equation (5.4) is strikingly similar to (5.1) except that we have now placed the GLM on the local prior means and not the relative risks themselves. Clearly, (5.4) does not directly account for any spatial correlation that may exist among the relative risks. We may correct for the spatial correlation by adding the the spatial random effects U_i to the GLM giving

$$\log m_i = X_i^T \beta + V_i + U_i, \tag{5.5}$$

where U_i is assigned a CAR prior. There is good reason to believe that the scale parameter included in the gamma prior accounts for much of the heterogeneity among the risks; however, it does not preclude the use of an additional latent variable. We now turn our attention to an example that highlights the similarities of (5.5) and (5.1).

5.4 *Example: Scottish Lip Cancer Data*

We motivate a comparison of the Poisson log-relative risk model (5.1) and the conjugate Poisson gamma model (5.4) by using a data set originally constructed by Clayton and Kaldor (1987). The data consisted of observed and expected counts of lip cancer registered in the 56 Scottish counties during the years 1975-1981. As previously mentioned the authors made various distributional assumptions for the local relative risks and then used EB methods for their posterior calculations. Banerjee, Carlin, and Gelfand (2004, p.167) later analyzed the data using a hierarchical generalized linear spatial model of the form (5.1), which also employed a CAR model for the spatial random effects. Recall that Figure 1 back in Chapter 4 displays the choropleth map for the crude estimates of relative risk. One county level covariate X_i, the percentage of the population engaged in agricultural, fishing or forestry (AFF) is available thus making (5.1) equivalent to

$$\log\theta_i = \beta_0 + \beta_1 X_i \cdot 10^{-1} + V_i + U_i \tag{5.6}$$

where V_i is assigned an exchangeable normal prior with mean zero and U_i is assigned the usual CAR model (5.3). The regression coefficients in (5.6) are given diffuse normal priors and the two precision parameters relating to the random components are assigned gamma distributions. The shape and scale parameters for the spatial precision parameter are .01 whereas the shape and scale parameters for the precision parameter relating to the unstructured heterogeneity are .001. These priors came from Best et al. (1985) and Bernardo, Berger, Dawid and Smith (1999). The hierarchy regarding Equation (5.6) can be summarized as

$$\begin{aligned}
&y_i \mid \theta_i \sim \text{Poisson}\left(E_i\theta_i\right)\\
&\beta_0, \beta_1 \overset{iid}{\sim} N\left(0, 1.0E^{-5}\right)\\
&V_i \mid \tau_v \overset{iid}{\sim} N\left(0, \tau_v\right)\\
&U_i \mid \tau_u \sim CAR(\tau_u)\\
&\tau_v \sim \text{Gamma}\left(1.0E^{-3}, 1.0E^{-3}\right)\\
&\tau_u \sim \text{Gamma}\left(1.0E^{-1}, 1.0E^{-1}\right)
\end{aligned}$$

Similarly, (5.5) can be rewritten as

$$\log m_i = \beta_0 + \beta_1 X_i \cdot 10^{-1}{}_i + V_i + U_i \tag{5.7}$$

where V_i is assigned an exchangeable Gaussian prior and U_i is assigned the usual CAR model (5.3). The regression coefficients in (5.7) are given diffuse normal priors and the precision parameter relating to the random component U_i is assigned a gamma distribution. The shape and scale parameters for the gamma distribution are both .01.

The hierarchy regarding the conjugate hierarchical model in Equation 5.7 can be summarized as

$$
\begin{aligned}
& y_i \mid \theta_i \sim \text{Poisson}\left(E_i\theta_i\right) \\
& \theta_i \mid m_i, \lambda_i \sim \text{Gamma}\left(\lambda_i, \lambda_i / m_i\right) \\
& \lambda_i \sim \chi^2\left(\rho\right) \\
& \beta_0, \beta_1 \overset{iid}{\sim} N\left(0, 1.0E^{-5}\right) \\
& V_i \mid \tau_v \overset{iid}{\sim} N\left(0, \tau_v\right) \\
& U_i \mid \tau_u \sim CAR(\tau_u) \\
& \tau_v \sim \text{Gamma}\left(1.0E^{-3}, 1.0E^{-3}\right) \\
& \tau_u \sim \text{Gamma}\left(1.0E^{-1}, 1.0E^{-1}\right)
\end{aligned}
$$

Model fitting is carried out using MCMC simulation methods implemented in the WinBUGS software. Because the models used in a mapping context may exhibit high correlations between model parameters, necessarily leading to highly autocorrelated samples, we use separate chains with different initial values for each model. Specifically, 75,000 values collected from three different chains (excluding the 4000 burn-in values) are used in the calculations of the relative risk posterior distribution $\theta_i \mid y_i$. We check the convergence by examining the line graphs provided in WinBUGS. Table 9 displays estimates of the posterior means for the local relative risk of each county under the three competing models. The column entitled *IV* is the second-stage hierarchical model suggested by Banerjee, Carlin, and Gelfand (2004), while the three remaining columns entitled *I*, II and *III* refer to the conjugate hierarchical model increasing in degrees of freedom. The data given in Table 9 are for the 56 counties arranged in descending order of incidence as measured by SMRs, which vary between 0 and 642. There is a noticeable

Table 9. Lip Cancer Incidence In Scotland by county: SMRs and Bayesian Estimates of Relative Risk

County	Observed	Expected	I	II	III	IV	SMR
1	9	1.4	496.4	479.2	475.7	474.7	642.8571
2	39	8.7	432	432.7	433.6	434.4	448.2759
3	11	3	334.2	333.1	330.8	332.2	366.6667
4	9	2.5	300.8	291.2	289.9	289.7	360
5	15	4.3	322.2	321.1	321.6	320.9	348.8372
6	8	2.4	378.1	386.3	384	383.6	333.3333
7	26	8.1	299.5	295.6	296.3	295.5	320.9877
8	7	2.3	280.2	281.1	281.4	281.8	304.3478
9	6	2	228.7	221.3	221.2	220.7	300
10	20	6.6	294.6	292.4	292.9	292.2	303.0303
11	13	4.4	275.3	280.5	279	280.9	295.4546
12	5	1.8	315.2	329.5	326.7	331.9	277.7778
13	3	1.1	255.2	263.3	263.1	263.3	272.7273
14	8	3.3	211.6	201.7	201.7	199.9	242.4242
15	17	7.8	190.4	181.7	181	180.6	217.9487
16	9	4.6	207.7	208.2	208.3	208.8	195.6522
17	2	1.1	200.2	204.8	207.4	206.1	181.8182
18	7	4.2	125.9	118.5	117.2	117.4	166.6667
19	9	5.5	179.4	191.1	193.4	194.4	163.6364
20	7	4.4	137.8	136.8	137.1	137.4	159.0909
21	16	10.5	144.7	139.9	139.3	139.1	152.381
22	31	22.7	144	145.3	145.5	145.4	136.5639
23	11	8.8	121.1	118.4	118.4	118.2	125
24	7	5.6	95.64	85.75	84.34	83.73	125
25	19	15.5	117.2	117.7	117.9	118.4	122.5807
26	15	12.5	106.5	102.4	101.3	101.2	120
27	7	6	98.05	94.58	93.85	93.31	116.6667
28	10	9	104.6	103.7	104.2	103.8	111.1111
29	16	14.4	118.1	120.7	121.2	121.9	111.1111
30	11	10.2	98.04	91	89.47	89.09	107.8431
31	5	4.8	88.86	85.35	84.89	84.69	104.1667
32	3	2.9	139.4	143.3	143.4	143.1	103.4483
33	7	7	98.73	96.46	96.18	95.98	100
34	8	8.5	86.75	79.63	78.56	77.77	94.11765
35	11	12.3	85.66	84.78	84.44	84.44	89.43089
36	9	10.1	77.9	75.5	75.05	75.5	89.10891
37	11	12.7	88.31	87.6	87.53	87.69	86.61417
38	8	9.4	68.98	62.06	60.96	60.41	85.10638
39	6	7.2	98.23	98.88	99.68	99.12	83.33333
40	4	5.3	61.94	57.64	56.99	56.66	75.4717
41	10	18.8	54.24	53.18	52.84	53.06	53.19149
42	8	15.8	66.68	72.54	72.83	73.34	50.63291
43	2	4.3	89.43	93.57	93.41	93.17	46.51163
44	6	14.6	45.72	46.09	45.83	46.17	41.09589

Table 9 continued.

45	19	50.7	41.1	42.7	43	43.15	37.47535
46	3	8.2	54.5	57.21	57.48	57.54	36.58537
47	2	5.6	47.16	46.89	46.37	46.55	35.71429
48	3	9.3	42.87	43.33	42.81	42.94	32.25807
49	28	88.7	33.85	34.93	35.11	35.27	31.56708
50	6	19.6	44.76	50.67	51.44	51.66	30.61225
51	1	3.4	49.08	48.48	47.81	47.68	29.41177
52	1	3.6	46.01	45.05	44.8	44.37	27.77778
53	1	5.7	39.45	39.89	39.9	39.72	17.54386
54	1	7	37.96	40.33	39.79	40.15	14.28571
55	0	4.2	64.1	81.88	83.37	83.99	0
56	0	1.8	71.56	75.56	75.47	75.95	0

decrease in variability for the Bayesian relative risk estimates as compared to the crude SMRs. The risk estimates computed using the non-conjugate hierarchical model range from 35 to about 475; using the conjugate model with 5 degrees of freedom associated with the chi-squared prior the estimates range from 33 to 496; using the conjugate model with 25 degrees of freedom the estimates range from 35 to 479; and using the conjugate model with 50 degrees of freedom the estimates range from 35 to 475. In general, models *III* and *IV* provide similar estimates of relative risk, especially when the expected count is small for a county adjacent to other low-risk counties. For example, the relative risks for county 24, which is adjacent to counties 27, 30, 31, 44, 47, 48, 55, and 56 (all of which are considered to be low-risk) are very similar; whereas the relative risk estimate computed using model *I* (conjugate with smaller degrees of freedom for the scale parameter) is much closer to the crude SMR. We can attribute this to the degree of belief that we have in our GLM. As mentioned back in Chapter 3 the scale parameter appearing in the conjugate gamma distribution controls the amount of faith one places in the GLM. Since the mean of a chi-squared distribution is identical to its degrees of freedom, larger degrees of freedom translate to a greater belief in the GLM. Consequently, the posterior estimates corresponding to these unstable areas are shrunk towards a localized neighborhood prior mean. Further examples of this occurrence are prevalent in counties 30, 33, 36, and 38. Also, the reader will find that in most cases when the observed count is less than the expected number of cases the relative risk estimates computed using model *I* are closest to the SMRs. It turns out that this is not the case when the observed count exceeds the expected number.

It is clear from Table 9 that the conjugate model offers a way to quantify one's own prior belief regarding the amount of overdispersion present in the areal units. By increasing the associated degrees of freedom of the hyperprior for the dispersion parameter we essentially place more belief in the specified GLM; thus relying on the CAR model to capture the spatial correlation among the neighboring areas. Alternatively, when the belief in the GLM is minimal the relative risk estimates provided by the conjugate hierarchical model are very similar to weighted average of the crude SMRs corresponding to the surrounding areas. Of course, these results are completely derived from this dataset. Ultimately we would want to embark on a simulation study like that appearing in Kafadar (1994) where she compared the EB methods of Clayton and Kaldor (1987) with several other non-parametric spatial smoothers.

5.5 Example: Waco Crime Data

In this final section we will illustrate how the previously discussed disease mapping techniques can be applied to mapping relative crime risks. We will use an original data set constructed by the authors and the Waco, Texas Police Department. The data set consists of aggregate counts of 911 calls pertaining to habitat burglaries collected during the year 2000 for each police beat. To clarify, a beat is a unit of area very similar to a census track. Since the United States Census Bureau employs a different system of units (e.g. tracks, blocks, etc.), we will use a GIS software package called ArcGIS to configure the two different units of measure with regard to census covariate information. That is, we will determine which of the census blocks correspond to what beats and aggregate the appropriate covariate information. This process is known as *dissolving*. In addition to the census covariate information, we have also recorded the number of houses in each beat which will be used to calculate the number of expected calls per beat via internal standardization (see Banerjee, Carlin, and Gelfand 2004, p.161). Figure 2 is a choropleth map displaying the raw unsmoothed call rates for habitat burglaries in Waco during the year 2000. Note the high concentration in the central part of the community.

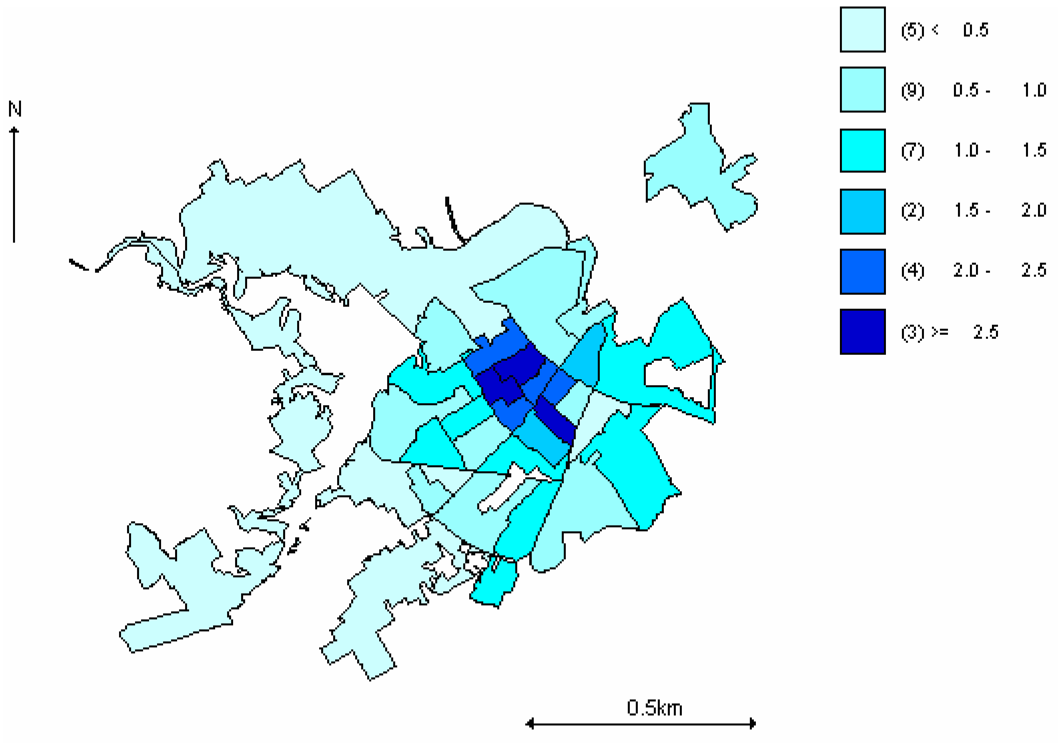

Figure 2. Choropleth Map of the Unsmoothed Call Rates for Waco in the Year 2000

The model that we will use is nearly identical to the conjugate hierarchical model presented in Section 5.4 e.g. Equation (5.7). That is, we assume that the prior mean for each beat satisfies the following generalized linear spatial model

$$\log m_i = \beta_0 + \beta_1 X_i \cdot 100^{-1}{}_i + V_i + U_i. \tag{5.8}$$

We use the same previously defined hierarchical structure

$$\begin{aligned}
&y_i \mid \theta_i \sim \text{Poisson}\left(E_i\theta_i\right)\\
&\beta_0, \beta_1 \overset{iid}{\sim} N\left(0, 1.0E^{-5}\right)\\
&V_i \mid \tau_v \overset{iid}{\sim} N\left(0, \tau_v\right)\\
&U_i \mid \tau_u \sim CAR(\tau_u)\\
&\tau_v \sim \text{Gamma}\left(1.0E^{-3}, 1.0E^{-3}\right)\\
&\tau_u \sim \text{Gamma}\left(1.0E^{-1}, 1.0E^{-1}\right)
\end{aligned}$$

where $E_i = N_i\tilde{r}$, $\tilde{r}$ refers to the number of 911 calls for habitat burglaries per habitat in the Waco community and N_i is the number of habitats in beat i . The covariate X_i in Equation (5.8) is the number of African Americans that reside in beat *i* during the year 2000 as recorded by the U.S. Census Bureau. Using WinBUGS, we estimate the relative call rate for each police beat. A technical aside: the convergence of the posterior distributions of the crime rates was relatively quick compared to those of the Scottish Lip Cancer. In fact, the length of the Markov chain required for convergence was only one third the size. Furthermore, the correlation between samples from the posterior distribution was insignificant after only 15,000 values. The results are summarized in Table 10.

Table 10. Bayesian and Classical Estimates of Relative Rates for Buglary Calls

Beat	Observed	Expected	Num.of Houses	SMR	Bayes df=10	Bayes df=50
1	57	41	1245	138.8	137.1	137.4
2	48	28	836	174.1	170.5	169.9
3	55	58	1751	95.3	97.44	97.33
4	50	71	2161	70.2	71.45	71.21
5	57	25	766	225.7	222.2	222.5
6	95	33	1004	286.9	282.9	283
7	83	35	1052	239.3	234.4	234.3
8	117	43	1308	271.2	267.9	267.8
9	71	35	1068	201.6	197.2	197.2
10	16	64	1983	24.8	28.69	28.72
11	10	4	133	227.8	203	206
12	19	20	596	96.7	99.68	100.8
13	38	78	2356	48.9	52.07	52.61
14	92	78	2373	117.6	117.1	117.4
15	15	28	854	53.3	54.98	55.13
16	47	41	1245	114.5	113.1	113
17	64	42	1268	153	151	151.6
18	53	21	642	250.4	241.6	241.7
19	64	67	2018	96.2	96.34	95.95
20	42	36	1081	117.8	114.7	114.4
21	52	74	2241	70.4	71.56	71.78
22	73	54	1630	135.8	134.9	134.8
23	100	98	2960	102.4	102.8	102.8
24	71	86	2597	82.9	83.14	83.15
25	67	79	2753	84.4	74.02	74.21
26	47	39	1173	121.5	117.4	117
27	16	63	1908	25.5	29.33	29.77
28	26	43	1308	60.3	59.82	59.58
29	22	89	2698	24.7	26.88	26.76
30	33	116	3509	28.5	30.07	29.99

The last two columns in Table 10 are the means for the posterior distributions of the rates as computed from the conjugate hierarchical Bayesian spatial model. The reader will notice that for the beats experiencing a significant call rate $(\text{i.e., observed/expected} > 1)$, the Bayesian rate estimates $\left(E[\theta \mid y]\right)$ are generally smaller than the MLEs. Again, this is attributable to spatially weighted average mechanism employed by the Bayesian model. For example, beat 11 had an expected count of fours calls however experienced 10. The corresponding MLE for the rate is 2.27 but the Bayesian estimates are only 2.03 and 2.06. Figure 3 is a choropleth map of the smoothed rates using a conjugate hierarchical spatial model with 50 degrees of freedom on the scale parameter.

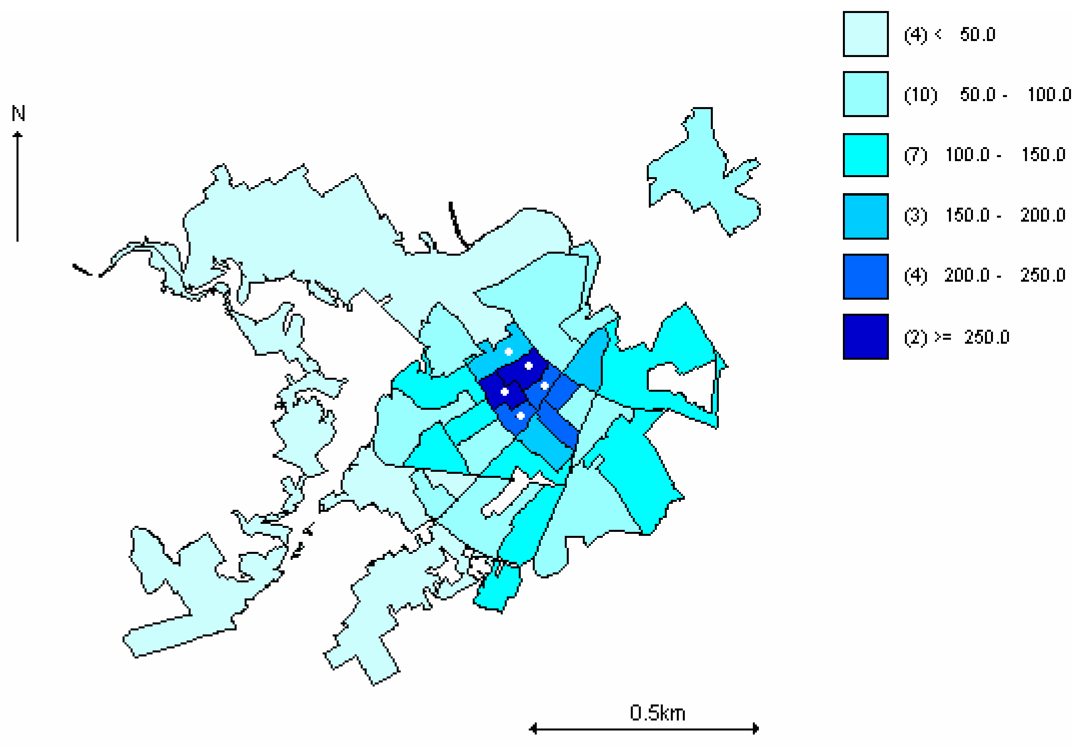

Figure 3. Choropleth Map of Smoothed Habitat Burglary 911 Call Rates

The beats having the white dots in the center will be used in the application of the subset selection procedure presented in Chapter Two. Starting from the most southern beat and moving clockwise they are seven, eight, nine, six, and five respectively. Beat nine has the lowest estimated call rate whereas beats eight and six have the highest estimated call rate. The calculated $PIC \equiv \Pr\left(\theta_i = \theta_{\max} \mid y_5, \ldots, y_9\right)$ for the 5 beats are displayed in Table 11.

Table 11. Probability Inclusion Criterions for the Five Beats

Beat	PIC
5	0.031
6	0.611
7	0.049
8	0.306
9	0.002

Based on this table the two PICs that exceed 1/5 are six and eight, thus our selected subset would consist of the rates corresponding to those two beats. In fact a loss penalty of $c = 33$ would be required to include beat five. The PICs were calculated via Monte Carlo integration.

CHAPTER SIX

Conclusion

The hierarchical conjugate model presented throughout this work provides an alternative to the customary generalized linear model. As seen it may be used in a variety of settings including statistical epidemiology and disease mapping. Matter of fact, it may be used in nearly any situation where one has overdispersed count data. The hierarchical conjugate model provides the same flexibility as the standard Bayesian GLM; however, it allows for extra-variability sometimes present in count data. The addition of a scale parameter used for this extra-variability comes at the expense of computing time and convergence. Convergence for risk mapping can be rather slow but this is also the case for the standard Bayesian GLM found in the literature. Therefore, the presented model is recommended at this point in time. That is, until an adequate simulation can be conducted to compare the conjugate hierarchical generalized linear spatial model to several other spatial smoothers found in the literature. Once the efficacy of this model is established we would like to extend this model to the area of forecasting or prediction.

At the beginning of this work the author had this situation in mind. We envisage a situation where data has been observed at several counties or in general areal units but for some reason has not been observed at other regions. Perhaps sampling was too difficult or in fact too dangerous. With the partial information collected we would ultimately like to predict the current state at these unobserved regions, whether it be the actual number of cases or the disease risk. This is most commonly referred to in the time series arena as a *state space model.* We believe that concept of this model provides a natural extension to the spatial arena. The major difference of course lies in the type of autocorrelation. It is customary to assume that the autocorrelation among the random variables is a function of time. We could assume that the regions are correlated over time but we would also assume correlation as a function of distance.

The concept of spatial prediction is not new. In fact it is referred to as *kriging* in the literature. The interest lies in extending *kriging* to lattice data. You see it is customary to assume that the spatial domain is continuous over $\Re^2$. This assumption along with a few others, namely isotropy, allow for a nice correlation structure for the domain. Obviously, this is not the case for disease mapping. In disease mapping we often use a nearest neighbors approach. Recall that Chapter 5 discusses this in great detail. Thus, it can be said that future work would somehow make use of the conjugate hierarchical model for prediction in the disease mapping setting.

APPENDICES

APPENDIX A

Computer Programs for Chapter 2

A.1.1 Outline for computer programs used in simulation in Chapter 2

The following gives a detailed outline of the computer program used in the simulation in chapter 2.

Part I – Generating the Poisson counts

1. Generate a $10,000\times m$ matrix consisting 10,000*m values from a Uniform(0,1). Call this matrix u. The value m refers to the number of parameters whereas 10,000 is the number of simulation iterations.

2. Create another $10,000\times m$ matrix in which the entries are found by setting the corresponding entries of u equal to the inverse cdf of a triangle distribution having endpoints (.2, 4) with mode 2 and solving. Each row now represents a set Poisson rates.

3. Order each row from greatest to least so that largest Poisson rate for each row is located in the first column. This will be helpful in calculating the probability of correct selection. This matrix will be called *lammat*.

4. Using the matrix of Poisson rates generate a matrix of count totals and call this matrix *countmat*. Each row of *countmat* will represent a sample of total counts having rates proportional to those from the corresponding row of *lammat*.

Part II – Calculating the Probability of Correct Selection and Expected Size

Case 1: Independent uniform priors

1. When the number of parameters is small (at most 5) we can use standard functions inside R, namely *dgamma*, *pgamma*, and *integrate*, to calculate the *probability inclusion criterion* $\equiv \Pr\left(\lambda_i = \lambda_{\max} \mid t_1,\ldots,t_m\right)$ where $t_1,\ldots t_m$ represent the counts for some row in *countmat*. For the case when $m > 5$ use Monte Carlo integration to calculate the probability inclusion criterions. It should be noted that no loops are required to calculate these probabilities but instead make use of the 'apply' function, which takes as arguments an array of values, a margin (in this case we apply the function over the rows), and an internal or user-defined function.

2. Create a matrix called *princludemax* in which each entry for some row will represent the probability inclusion criterions for the corresponding superior set. Recall that each

row in *rlambda* represents a set of *m* parameters, thus each row of *princludemax* should sum to one.

3. Since the first element of each row in *rlambda* was the maximum of the parameter set, the first entry of each row in *princludemax* will represent the probability inclusion criterion for the largest parameter in the corresponding *rlambda* row.

4. The probability of correct selection is the percentage of times that the probability inclusion criterions in the first column of *princludemax* exceeded the *criterion constant* $1/_{c+1}$, where $c \geq m-1$. Similarly, the expected size is the relative frequency of **all** the probabilities in *princludemax* that exceed the criterion constant.

Case II: Three Stage Hierarchy

The code for computing the probability of correct selection in the case of a hierarchical model only differs to the uniform case in the first step. The reader will use a package called 'R2WinBUGS' to generate samples from the posteriors. See Sturtz, Ligges, and Gelman (2004) for a discussion on how to call WinBUGS from R. The following steps provide an outline for constructing *princludemax* as in the previous case

1. Install and load the package "R2WinBUGS" in R and create a directory in your machine to store a text file consisting of the BUGS model statement. The initial values will be supplied to WinBUGS via an R list. Of course the data values are the actual rows of the matrix *countmat*.

2. Define a function in R called *winbug* that will take as an argument a row from the matrix *countmat* and specify the objects to be used in the internal function *bugs*. The function *bugs* requires a data object, a list of initial values, a vector of parameters for the inference, the number of chains, the number of iterations, the location of the model file. However, storing the various parameter estimates and the burnin specification are optional. A sample *bugs* specification could appear as *bugs(data,inits,model.file=" ",parameters,n.chains., n.iter,bugs.directory=" ")*

3. Recall that all iterations in the simulation process produce a set of exactly *m* total counts. For one iteration, to calculate the probability inclusion criterions for the *m* parameters we sample 9,000 posterior values from each of the *m* posteriors using three chains of length 4000 (1000 burn in values) and construct a matrix *lam.sim*. The probability inclusion criterion for one of the *m* candidates, say *k*, is found by calculating the relative frequency in which the k^{th} column was a maximum. These values will form the rows of the matrix *princludemax*.

A.2.2 Code used for the simulation study presented in Chapter 2

The following code was used to calculate the probability of correct selection and expected size for the case when the number of populations was five and a flat uniform prior was assumed.

(a,b) is the interval and c is the mode. n refers to the sample size and m is the number of populations

```
inposlambda<-function(a,b,c,n,m){
    assign("Global.res",a,b,env=.GlobalEnv)
    assign("Global.res",c,n,env=.GlobalEnv)
    assign("Global.res",m,env=.GlobalEnv)
    #####This code generates parameters from a triangle distribution with mode=c
    u<-matrix(runif(10000*m),ncol=m)
    u<-t(apply(u,1,sort))
    u<-t(apply(u,1,rev))
    rlambda<-function(x){
        pmode<-(c-a)/(b-a)
        ifelse(x<pmode,a+sqrt((c-a)*(b-a)*x),b-sqrt((x-pmode)*(c-b)*(b-a)+(c-
        b)^2))}
    lammat<-rlambda(u)
    countmat<-matrix(rpois(10000*m,n*lammat),ncol=m)
    posmax<-function(x){return(which(x==max(x),arr.ind=TRUE))}
        psone<-function(y){
        assign("Global.res",y,env=.GlobalEnv)
            funone<-
            function(x){return(pgamma(x,y[2]+1,n)*pgamma(x,y[3]+1,n)*dga
            mma(x,y[1]+1,n)*pgamma(x,y[4]+1,n)*pgamma(x,y[5]+1,n))}
            integrate(funone,0,Inf)$value}
        pstwo<-function(y){
        assign("Global.res",y,env=.GlobalEnv)
            funtwo<-
            function(x){return(pgamma(x,y[1]+1,n)*pgamma(x,y[3]+1,n)*dga
            mma(x,y[2]+1,n)*pgamma(x,y[4]+1,n)*pgamma(x,y[5]+1,n))}
            integrate(funtwo,0,Inf,stop.on.error = FALSE)$value}
        psthree<-function(y){
        assign("Global.res",y,env=.GlobalEnv)
            funthree<-
            function(x){return(pgamma(x,y[2]+1,n)*pgamma(x,y[1]+1,n)*dga
            mma(x,y[3]+1,n)*pgamma(x,y[4]+1,n)*pgamma(x,y[5]+1,n))}
            integrate(funthree,0,Inf,stop.on.error = FALSE)$value}
        psfour<-function(y){
        assign("Global.res",y,env=.GlobalEnv)
            funfour<-
            function(x){return(pgamma(x,y[1]+1,n)*pgamma(x,y[2]+1,n)*pga
            mma(x,y[3]+1,n)*dgamma(x,y[4]+1,n)*pgamma(x,y[5]+1,n))}
            integrate(funfour,0,Inf,stop.on.error = FALSE)$value}
        psfive<-function(y){
        assign("Global.res",y,env=.GlobalEnv)
            funfive<-
            function(x){return(pgamma(x,y[1]+1,n)*pgamma(x,y[2]+1,n)*pga
            mma(x,y[3]+1,n)*pgamma(x,y[4]+1,n)*dgamma(x,y[5]+1,n))}
            integrate(funfive,0,Inf,stop.on.error = FALSE)$value}

princludemax<-apply(countmat,1,psone)
prcs<-c(rep(0,4))
for (i in 1:4) prcs[i]<-length(princludemax[princludemax>(1/(i+4))])/10000
princludetwo<-apply(countmat,1,pstwo)
princludethree<-apply(countmat,1,psthree)
princludefour<-apply(countmat,1,psfour)
princludefive<-apply(countmat,1,psfive)
set<-c(princludemax,princludetwo,princludethree,princludefour,princludefive)
expectedsize<-rep(0,4)
for (i in 1:4) expectedsize[i]<-length(set[set>(1/(i+4))])/10000
return(cat("Uniform Prior","Sample size =",n,"# Parameters =",m,"PrCS
=",prcs,"Expected Size =",expectedsize))
}
```

The following code is used to calculate the probability of correct selection and expected size when the number of parameters exceeds five.

```
simposlambda<-function(a,b,c,n,m){
    assign("Global.res",a,b,env=.GlobalEnv)
    assign("Global.res",c,n,env=.GlobalEnv)
    assign("Global.res",m,env=.GlobalEnv)
    #####This code generates parameters from a triangle distribution with mode=c
    u<-matrix(runif(10000*m),ncol=m)
    u<-t(apply(u,1,sort))
    u<-t(apply(u,1,rev))
    rlambda<-function(x){
        pmode<-(c-a)/(b-a)
        ifelse(x<pmode,a+sqrt((c-a)*(b-a)*x),b-sqrt((x-pmode)*(c-b)*(b-a)+(c-
        b)^2))}
    lammat<-rlambda(u)
    countmat<-matrix(rpois(10000*m,n*lammat),ncol=m)
    posmax<-function(x){return(which(x==max(x),arr.ind=TRUE))}
        generate<-function(x){
        posample<-rgamma(m*25000,x+1,n)
        y<-matrix(posample,ncol=m,byrow=T)
        test<-unlist(c(apply(y,1,posmax),1:m))
        postprobtab<-(table(test)-1)/25000
        postprob<-rep(0,m)
        for (i in 1:m) postprob[i]<-postprobtab[[i]]
        return(postprob)}
set<-apply(countmat,1,generate)
set<-t(set)
princludemax<-set[,1]
prcs<-rep(0,4)
for (i in 1:4) prcs[i]<-length(princludemax[princludemax>(1/(i+(m-1)))])/10000
expectedsize<-rep(0,4)
for (i in 1:4) expectedsize[i]<-length(set[set>(1/(i+(m-1)))])/10000
return(cat("Uniform prior","Sample size =",n,"# Parameters =",m,"PrCS
=",prcs,"Expected Size =",expectedsize))
}
```

The following code was used to calculate the probability of correct selection and expected size for the hierarchical model.

```
#####(a,b) is the interval and c is the mode.  n refers to the sample size and m is the
number of populations
hiposlambda<-function(a,b,c,n,m){
    assign("Global.res",a,b,env=.GlobalEnv)
    assign("Global.res",c,n,env=.GlobalEnv)
    #####This code generates parameters from a triangle distribution with mode=c
    u<-matrix(runif(10000*m),ncol=m)
    u<-t(apply(u,1,sort))
    u<-t(apply(u,1,rev))  #ensures that the largest parameter is in the first column
    #####Inverse CDF of Triangle Distribution
    rlambda<-function(x){
        pmode<-(c-a)/(b-a)
        ifelse(x<pmode,a+sqrt((c-a)*(b-a)*x),b-sqrt((x-pmode)*(c-b)*(b-a)+(c-
        b)^2))}
    lammat<-rlambda(u)
    ####Generates counts from the Poisson having means 'lammat'
    countmat<-cbind(matrix(rpois(10000*m,n*lammat),ncol=m),1:10000)
    posmax<-function(x){return(which(x==max(x),arr.ind=TRUE))}

####This code calls WinBUGS to compute posterior samples from hierarchical
models##############
#####The WinBUGS model specification "simulation.txt" is stored in the folder
"c:/John/WinBUGS"#####
#####The simulated values are saved to matrix, 'counts.sim'.  We are only interested in
parameter 'mu' ###

winbug<-function(x){
    z<-x[c(1:m)]
    data <- list ("n", "z","m")
    inits1 <- list(lambda=rep(1,m),alpha=1,beta=1)
    inits2 <- list(lambda=rep(1,m),alpha=1,beta=1)
```

```
        inits3 <- list(lambda=rep(1,m),alpha=1,beta=1)
        inits<-list(inits1,inits2,inits3)
        parameters <- c("mu")
         counts.sim<- bugs (data, inits, parameters,
        "c:/John/R2WinBUGS/simulation.txt", n.chains=3,
        n.iter=4000,n.burnin=1000,DIC=F,bin=3000)$sims.matrix[,c(1:m)]
        test<-unlist(c(apply(counts.sim,1,posmax),1:m))
        postprobtab<-(table(test)-1)/9000
        postprob<-rep(0,m)
        for (i in 1:m) postprob[i]<-postprobtab[[i]]
        #print(c(postprob,x[m+1]))
        return(postprob)}

set<-apply(countmat,1,winbug)
set<-t(set)
princludemax<-set[,1]
prcs<-rep(0,4)
        #####This loop caculates pr. of correct selection for different c=4,5,6,7
        for (i in 1:4) prcs[i]<-length(princludemax[princludemax>(1/(i+(m-1)))])/10000
        expectedsize<-rep(0,4)
        for (i in 1:4) expectedsize[i]<-length(set[set>(1/(i+(m-1)))])/10000
        return(cat("Hierarchical Model","Sample size=",n,"# Parameters=",m,"PrCS
        =",prcs,"Expected Size=",expectedsize))
}
```

The following WinBUGS model specification should be stored in a file located in the current working directory in R.

```
model
{
        for( i in 1 : m )
        {
        z[i]~ dpois(mu[i])
        mu[i] <- lambda[i]*n
        lambda[i]~dgamma(alpha, beta)
        }
alpha~dexp(1)
beta~dgamma(.001, .001)
}
```

APPENDIX B

Instructions for Importing Adjacency Matrices and Maps into WinBUGS

*These instructions were taken from the personal page of Ms. Yue Cui, graduate student at University of Minnesota. See http://www.biostat.umn.edu/~yuecui/convert.r

1. Create a .cgm file from Arcview

- Open your Arcview shape file in Arcview, for example, a.shp
- Click "Export" in File menu
- Choose from List Files of Type "CGM Clear Text" and save .cgm file(e.g. a.cgm). This .cgm file is a text file and can be opened with any word processor like Notepad. Save this in the current R work directory.

2. Convert the .cgm file into a .txt file in Splus format readible by WINBUGS

- Copy and paste the following text. This script creates a function in R called ‘convert’ which creates a .txt file containing polygon boundaries

```
convert<-function(cgmfile)
  {
#outfile <- "test.txt"
 outfile<-paste(cgmfile,".txt",sep="")

#This one works
#fortest<-scan("test.cgm",what=list(name=""),sep="\n")
 fortest<-scan(paste(cgmfile,".cgm",sep=""),what=list(name=""),sep="\n")

fortest<-fortest$name
fortest<-fortest[grep("VIS",fortest)]
fortest<-as.matrix(fortest)

totpolyn<-length(fortest[grep("POLYGON_SET",fortest)])

polyn<-0
count<-0;
indicator<-0;
#First tried rep(0,1000), but there is a polygon with 167 rows, so we get
#167*3*2=1002 coordinates in one polygon. Error occurs with NA output
coord<-rep(0,5000)

write(paste("map:",totpolyn,"\n"),outfile)
for (i in 1:totpolyn) {
    write(paste( i, paste("grid",i,sep="")),outfile,append=T)
  }


for (i in 1:length(fortest)){
  for(j in (1:nchar(fortest[i]))){
      letter<-substring(fortest[i],j,j)
      #At first try to use AsciiToInt,stupid
    #if (AsciiToInt(letter)<=AsciiToInt("9")&
    #    AsciiToInt(letter)>=AsciiToInt("0"))
    if(letter<="9"&letter>="0"){
                 if (indicator==0) {count<-count+1;
                              indicator<- 1
                            }
                   coord[count]<-coord[count]*10+
                      type.convert(letter)
                      }
    #add the following else if statement because if and only when
    #a POLYGON_SET is encounted, polyn is increased by 1
    else if (letter=="P"){ polyn<-polyn+1;indicator<-0}

    #a CLOSEVIS is encounted, output the coordinates set, this is
    #done independently with increase of polyn since a POLYGON_SET may
    #consist of several small polygons
    else if(letter=="C"){
     coordmat<-cbind(rep(paste("grid",polyn,sep=""),count/2),
                coord[2*(1:(count/2))-1],
                coord[2*(1:(count/2))])

     if (polyn == 1) { write("",outfile,append=T)}
     else           { write(c(NA,NA,NA),outfile,append=T)}
     write(t(coordmat),outfile,append=T,ncol=3)
     #polyn<-polyn+1
     #the above statement is not right because 2 CLOSEVIS may lie in one
     #same polygon, but we output a polygon coordinates set whenever a
     #CLOSEVIS is encountered, although it may have the same label,
     #denoted by "grid&polyn" with previous or next coordinate sets.

     count<-0
     indicator<-0
     coord<-rep(0,5000)
     #if(grep("\r",substring(fortest[i],j+1,nchar(fortest[i]))))
     #break
    }

    else      indicator<-0
  }
}
write("END",outfile,append=T)
}
```

- Under the command line, type in convert("filename"), then ‘filename.txt’ will be generated and saved in the working directory. Here "filename" doesn't have .cgm in it. e.g, when you have a.cgm in folder, you should submit convert("a") instead of convert("a.cgm")

3. Read .txt file into WinBUGS

- Open .txt file in WINBUGS using open in File menu, remember to choose Text[.txt] or Text[Dos encoding][.txt] from File of type option. Now WinBUGS will pull up a text editing window showing the .txt file.

- Click Map menu and choose Import Splus, if the .txt file is in correct Splus format, then a Save as window will come out, and you can now save it as a .map file ready for mapping, e.g. ‘a.map’. If the .txt has an error in it, WinINBUGS will beep.
- Now close WinBUGS and restart it.
  Click map in WinBUGS window and choose Adjacency Tool, in the pop-out window, choose the .map you just created from the dropdown list of maps, click on adj map and an Adjacency Map window will come out showing the map same as the one you exported out of Arcview. Now you can click adj matrix icon in ‘Adjacency Tool’ to get an adjacency matrix for this map.

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis*, New Jersey: John Wiley & Sons, Inc.

Albert, J. (1988), "Computational Methods Using a Bayesian Hieararchical Generalized Linear Model," *Journal of the American Statistical Associaton*, 92, 916-925.

Albert, J. (1992), "A Bayesian Analysis of a Poisson Random-Effects Model for Home Run Hitters," *The American Statistician*, 46, 246-253.

Banerjee, S., Carlin, B., and Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/CRC Press.

Bechhofer, R., Santner, T., and Goldsman, D. (1995), *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*, New York: John Wiley and Sons, Inc.

Berger, R (1980), "Minimax Subset Selection for the Multinomial Distribution," *Journal of Statistical Planning and Inference*, 4, 391-402.

Berger, J. and Deely, J. (1988), "A Bayesian Approach to Ranking and Selection of Related Means with Alternatives to Analysis of Variance Methodology," *Journal of the American Statistical Association*, 83, 364-373.

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society*, Ser. B, 36, 192-236.

Besag, J., York, J., and Mollíe, A. (1991), "Bayesian Image Restoration with Two Applications in Spatial Statistics," *Annal of the Institute of Statistics and Mathematics*, 46, 1-59.

Besag, J., and Kooperberg, C. (1995), "On Conditional and Intrinsic Autoregressions," *Biometrika*, 82, 733-746.

Best, N., Ickstadt, K., and Wolpert, R. (2000), "Spatial Regression for Health and Exposure Data Measured at Disparate Resolutions," *Journal of the American Statistical Association*, 95, 1076-1088.

Bratcher, T., and Bhalla, P. (1974), "On the Properties of an Optimal Selection Procedure," *Communications in Statistics*, 3, 1974, 191-196.

Breslow, N., and Clayton, D. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9-25.

Breslow, N., and Day, N. (1987), "Statistical Methods in Cancer Research," in *The Analysis of Cohort Studies* (Vol. 2, 82), Lyon: IARC Publications.

Brillinger, D. (1990), "Spatio-Temporal Modelling of Spatially Aggregated Binary Data," *Survey Methodology*, 16, 255-269.

Carlin, B., and Louis, T. (2000), *Bayes and Empirical Bayes Methods for Data Analysis,* Boca Raton, FL: Chapman & Hall/CRC Press.

Choynowski, M. (1959), "Maps Based on Probabilities," *Journal of the American Statistical Association*, 54, 385-388.

Christiansen, C., and Morris, C. (1997), "Hierarchical Poisson Regression Modeling," *Journal of the American Statistical Association*, 92, 618-632.

Clayton, D., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Desease Mapping," *Biometrics*, 43, 671-682.

Clayton, D., and Bernardinelli, L. (1992), "Bayesian Methods for Disease Mapping," in *Geographical and Environmental Epidemiology*, eds. P. Elliot, J. Cuziak, D. English, and R. Stern, Oxford: Oxford University Press, pp. 205-220.

Clayton, D. (1996), "Genaralised Linear Mixed Models," in *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. Spiegelhalter, London: Chapman and Hall, pp. 279-301.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley.

Dellaportas, P., and Smith, A. (1993), "Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling," *Applied Statistics*, 42, 443-459.

Diggle, P., Tawn, J., and Moyeed, R. (1998), "Model-based Geostatistics," *Applied Statistics*, 47, 299-350.

Elliot, P., Wakefield, J., Best, N., and Briggs, D. (2001), *Spatial Epidemiology*, Oxford: Oxford University Press.

Fahrmeir, L., and Tutz, G. (1994), Multivariate Statistical Modeling Based on Generalized Linear Models, New York: Springer-Verlag.

Gelman, A., Carlin, J., Stern, H., and Rubin,D. (2004), *Bayesian Data Analysis*, New York: Chapman & Hall.

Ghosh, M., and Gelfand, A. (2000), "Generalized Linear Models: A Bayesian View," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. Dey, S. Ghosh, B. Mallick, New York: Marcel Dekker, pp. 3-21.

Ghosh, M., Natarajan, K., Stroud, T., and Carlin, B. (1998), "Generalized Linear Models for Small-Area Estimation," *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., Natarajan, K., Waller, L., and Kim, D. (1999), "Hierarchical GLMs for the Analysis of Spatial Data: An Applicaton to Disease Mapping," *Journal of Statistical Planning and Inference*, 75, 305-318.

Gupta, S., and Sobel, M. (1957), "On a Statistic Which Arises in Ranking and Selection Problems," *Annals of Mathematical Statistics*, 28, 957-967.

Gupta, S., and Yang, H. (1985), "Bayes Subset Selection Procedures for the Best Population," *Journal of Statistical Planning and Inference*, 12, 213-233.

Haviland, A. (1975), *The Geographical Distribution of Diseases in Great Britain*, London: Smith Elder.

Ibrahim, J.G., and Laud, P.W. (1991), "On Bayesian Analysis of General Linear Models Using Jeffreys's Prior," *Journal of the American Statistical Association*, 86, 981-986.

Kafadar, K. (1997), "Geographical Trends in Prostate Cancer Mortality: An Application of Spatial Smoothers and the Need Adjustment," *Annals of Epidemiology*, 7, 35-45.

Kvam, P., and Miller, J. (2002), "Common Cause Failure Prediction Using Data Mapping," *Reliability Engineering and System Safety*, 76, 273 – 278.

Lawson, A., Browne, W., and Rodeiro, C. (2003), *Disease Mapping with WinBUGS and MLwiN*, New York: John Wiley & Sons, Ltd.

Lehmann, E. (1999), *Elements of Large Sample Theory*, New York: Springer-Verlag.

Leonard, T., and Novick, M.R. (1986), "Bayesian Full Rank Marginalization for Two-Way Contingency Tables," *Journal of Educational Statistics*, 11, 33-56.

Marshall, R. (1991), "Mapping Disease and Mortality Rates Using Empirical Bayes Estimators," *Applied Statistics*, 40, 283-294.

Matérn, B. (1986), *Spatial Variation* (2nd ed.), Berlin: Springer.

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models*. Chapman and Hall, London.

McCulloch, C., and Searle, S. (2001), *Generalized, Linear, and Mixed Models*, New York: John Wiley & Sons.

Mollie, A. (1996), “Bayesian Mapping of Disease,” in *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S. Richardson, and D. Spiegelhalter, London: Chapman and Hall, pp. 359-379.

Nelder, J.A., and Wedderburn, R.W.M. (1972), “Generalized Linear Models,” *Journal of the Royal Statistical Association*, Ser. A, 135, 370-384.

Raferty, A., and Banfield, J. (1991), “Stopping the Gibbs Sampler, the Use of Morphology, and Other Issues in Spatial Statistics,” *Annals of the Institute of Statistical Mathematics*, 43, 32-43.

Schulter, P., Deely, J., and Nicholson, A. (1997), “Ranking and Selecting Motor Vehicle Accident Sites by Using a Hierarchical Bayesian Model,” *The Statistician*, 46, 293-316.

Searle, S., Casella, G., and McCulloch, C. (1991), *Variance Components,* London: Wiley.

Snow, J. (1855), *On the Mode of Communicatoin of Cholera* (2nd ed.), London: Churchill.

Stamey, J., Bratcher, T., and Young, D. (2004), “Parameter Subset Selection and Multiple Comparisons of Poisson Rate Parameters with Misclassification,” *Computational Statistics and Data Analysis*, 45, 467-479.

Sturtz, S., Ligges, U., and Gelman, A. (2005), “R2WinBUGS: A Package for Running R from WinBUGS,” *Journal of Statistical Software*, 12, 1-17.

Suissa, S., and Salmi, R. (1989), “Unidirectional Multiple Comparisons of Poisson Rates,” *Statistics in Medicine*, 8, 757-764.

Tsutakaw, R, (1985), “Estimation of Cancer Mortality Rates: A Bayesian Analysis of Small Frequencies,” *Biometrics*, 41, 69-79.

Tukey, J. (1988), “Statistical Mapping: What Should Not Be Mapped,” in *Collected Works of John W. Tukey,* pp.109-121. Belmont, CA: Wadsworth.

Wakefield, J., Best, N., and Waller, L. (2000), “Bayesian Approaches to Disease Mapping,” in *Spatial Epidemiology: Methods and Applications*, eds. Elliott, P., Wakefield, J., Best, N., and Briggs, D., Oxford: Oxford University Press, pp.106-127.

Waller, L. (2002), “Hierarchical Models for Disease Mapping,” *Encyclopedia of Environmetrics* (Vol. 1), Chichester: John Wiley & Sons, pp. 1004-1007.

Waller, L., and Gotway, C., (2004), *Applied Spatial Statistics for Public Health Data*, New Jersey: Wiley & Sons Inc.

Walter, S. (2000), “Disease Mapping: A Historical Perspective,” in *Spatial Epidemiology: Methods and Applications*, eds. Elliott, P., Wakefield, J., Best, N., and Briggs, D., Oxford: Oxford University Press, pp 223-239.

West, M. (1985), “Generalized Linear Models: Scale Parameters, Outlier Accommodation and Prior Distributions,” in *Bayesian Statistics*, 2, eds. Bernardo, J., Degroot, M., Lindley, D., and Smith, A., Oxford: Oxford University Press, pp 531-557.

Wolpert, R., and Ickstadt, K. (1998), “Poisson/Gamma Random Field Models for Spatial Statistics,” *Biometrika*, 85, 254-267.