Bayesian Logistic Regression with Spatial Correlation: An Application to Tennessee River Pollution

by

William M. Marjerison, Jr.

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

by

December 2006

APPROVED:

Professor Balgobin Nandram, Thesis Advisor

Professor Bogdan Vernescu, Head of Department

Abstract

We analyze data (length, weight and location) from a study done by the Army Corps of Engineers along the Tennessee River basin in the summer of 1980. The purpose is to predict the probability that a hypothetical channel catfish at a location studied is toxic and contains 5 ppm or more DDT in its filet. We incorporate spatial information and treate it separetely from other covariates. Ultimately, we want to predict the probability that a catfish from the unobserved location is toxic.

In a preliminary analysis, we examine the data for observed locations using frequentist logistic regression, Bayesian logistic regression, and Bayesian logistic regression with random effects. Later we develop a parsimonious extension of Bayesian logistic regression and the corresponding Gibbs sampler for that model to increase computational feasibility and reduce model parameters. Furthermore, we develop a Bayesian model to impute data for locations where catfish were not observed. A comparison is made between results obtained fitting the model to only observed data and data with missing values imputed. Lastly, a complete model is presented which imputes data for missing locations and calculates the probability that a catfish from the unobserved location is toxic at once.

We conclude that length and weight of the fish have negligible effect on toxicity. Toxicity of these catfish are mostly explained by location and spatial effects. In particular, the probability that a catfish is toxic decreases as one moves further downstream from the source of pollution.

Acknowledgements

This thesis draws heavily on past courses and work experience. More importantly, it draws on many people without whom this work would not have been complete.

First and foremost, I would like to thank my thesis advisor, Balgobin Nandram. He is a source of enlightenment for me in the technical aspects of the thesis, and has mentored me countless times. I learned much more from him than just Bayesian statistical methodology. I sincerely thank him for all that he has given me.

It feels strange that the time I spent helping out the family's software business is invaluable. I thank my family too for the unconditional love I continue to receive from them.

I also thank faculty and staff of the Department of Mathematical Sciences with whom I have interacted over the years. They all have had an impact on this work in some way.

Finally, I am greatly indebted to Dr. Jai Won Choi, National Center for Health Statistics, for partial financial support and constant encouragement and praise.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Description

In this paper, we present a sensible yet parsimonious approach to modeling binary probabilities of success of events occuring in contiguous regions of space. In particular, we look at models of the form

$$y_{ij} \sim Ber(p(\underline{x_{ij}}, s_i)) \tag{1.1}$$

in which the jth binary observation at location i has a probability of success p which depends on the covariates as well as on a function of location (i.e., distance from some source) s_i. Jank and Kannan (2005) discusses marketing applications of such models in the frequentist settings. Cressie (1993) is a general exposition on frequentist spatial statistics, whereas Banerjee, Carlin, Gelfand (2004) focuses on recent methods in Bayesian spatial statistics.

Our focus, however, is on spatial structures that are equally-spaced, linear, and ordered, the most important example of which are rivers with equally-spaced observation points. The models then take the form

$$y_{ij} \sim Ber(p(\underline{x_{ij}}, i)). \tag{1.2}$$

We propose a parsimonious model to handle this situation, present Markov Chain Monte Carlo algorithms to infer model parameters, and offer a logically sound imputation method for missing observations. To motivate our approach to this class of problems, we will apply the methodology to studying pollution in the Tennessee River basin. Ultimately, we wish to estimate the proportion of fish at a given location along the river that contains 5 ppm or more of DDT and is hence unsafe for consumption according to law.

In this chapter, we describe the geography of the Tennessee River and the extent of industrial pollution in the Tennessee River basin. We conclude with a rudimentary Bayesian logistic regression model to study the effects of length, weight, and location of channel catfish on whether or not the DDT content of the catfish is greater than 5 ppm.

1.2 Preliminary Analysis of Data

Data for Catfish consists of whether or not it contained 5 ppm or more of DDT in its filet (y_{ij}), its length in centimeters, its weight in grams, and the location indicator (an integer from between 0 and 13, which correspond to locations TRM275 and TRM340, respectively).

As the first step, we consider frequentist logistic regression on the given information. Here we begin by looking at the model

$$y_{ij} \sim Ber(logit^{-1}(\underline{x_{ij}}'\underline{\beta})) \tag{1.3}$$

where $\underline{x_{ij}}$ contains 1 (corresponding to the intercept term), centered length, and centered weight. For a detailed exposition on frequentist logistic regresion, see Lemeshow and Hosmer (1989).

Parameter	Coef	Std Coef	Z	P value
Constant	1.53	0.326	4.70	0.000
Length	0.291	0.370	0.78	0.432
Weight	0.386	0.437	0.88	0.377

Table 1.1: Logistic Regression on Centered Length, Weight

Method	Chi-Square	DF	P value
Pearson	67.3	69	0.537
Deviance	67.1	69	0.542
Hosmer-Lemeshow	11.4	8	0.178

Table 1.2: Goodness of Fit Test for Regression on Centered Length and Centered Weight

We see that only the constant term is significant at the 5 percent level. All three goodness-of-fit tests suggest that the model is a reasonable description of the data. Next, we consider introducing location as a covariate. We use the centered values of distance from location 0 in units of 5 miles, as these values are equivalent to the location indices.

Again, only the constant term is significant at the 5 percent level, with all three goodness-of-fit tests supporting this conclusion.

Next, we consider simple models in the Bayesian paradigm. We assume basic familiarity with Bayesian methods as presented in Box and Tiao (1973).

Parameter	Coef	Std Coef	Z	P value
Constant	1.58	0.338	4.67	0.000
Length	0.350	0.394	0.89	0.375
Weight	0.197	0.492	0.40	0.689
Location	-0.434	0.361	-1.20	0.229

Table 1.3: Logistic Regression on Centered Length, Weight, and Location

Method	Chi-Square	DF	P value
Pearson	69.1	68	0.438
Deviance	65.6	68	0.559
Hosmer-Lemeshow	4.86	8	0.773

Table 1.4: Goodness of Fit Test for Regression on Centered Length, Centered Weight, and Centered Location

The Bayesian equivalent of the previous frequentist models is

$$\begin{aligned} y_{ij}|\underline{\beta} &\sim Ber(logit^{-1}(\underline{x_{ij}}'\underline{\beta})) \\ \underline{\beta} &\sim N(\underline{\theta}, \Sigma) \end{aligned} \tag{1.4}$$

where

$$logit^{-1}(\underline{x_{ij}}'\underline{\beta}) = \frac{\exp \underline{x_{ij}}'\underline{\beta}}{1 + \exp \underline{x_{ij}}'\underline{\beta}} \tag{1.5}$$

and the coefficient vector $\underline{\beta}$ has a normal prior distribution with mean $\underline{\theta}$ given by the frequentist maximum likelihood estimate and the covariance matrix Σ given as 100 times the frequentist estimate for the covariance matrix, 100 being the blow-up factor making the prior distribution reasonably diffuse and noninformative.

Under this model, the posterior distribution of $\underline{\beta}$ is

$$
\begin{aligned}
\pi(\underline{\beta}|\underline{y}) &\propto N(\underline{\theta}, \Sigma) \\
\times \prod_{ij} (logit^{-1}(\underline{x_{ij}}'\underline{\beta}))^{y_{ij}} &(1 - logit^{-1}(\underline{x_{ij}}'\underline{\beta}))^{1-y_{ij}} \qquad (1.6)
\end{aligned}
$$

To sample from this posterior distribution, we construct an independence chain Metropolis sampler with a multivariate student's-t proposal density q with tunable degrees of freedom κ:

$$
q(\underline{\beta}|\underline{y}) \propto (1 + \kappa^{-1}(\underline{\beta} - \underline{\beta_0})'\Sigma^{-1}(\underline{\beta} - \underline{\beta_0}))^{-\frac{\kappa+p}{2}}, \qquad (1.7)
$$

κ for this problem must be less than 2 so that the jump probabilities are between 0.25 and 0.50. See Gelman, Roberts and Gilks (1995) for reasons why jumping probabilities of Metropolis samplers ought to fall in this range. We run a simulation with 101000 iterations, discarding the first 1000 iterations and then taking every 100 iterations afterwards.

Parameter	Mean	Std. Dev.	Num. SE	0.95 Cred. Int.	
Constant	-0.872	2.82	0.0887	-6.29	4.64
Length	0.110	0.089	0.00300	-0.0620	0.277
Weight	0.001	0.00177	0.000061	0.00467	0.00223
Location	-0.246	0.0890	0.00225	-0.448	-0.0866

Table 1.5: Bayesian Logistic Regression with Length, Weight, and Location. Posterior Estimates.

Here, the 95 percent credible intervals determine which components of $\underline{\beta}$ are relevant to the model. Weight and location are significant because their respective 95 percent credible intervals do not contain zero. The Metropolis sampler converges, albeit slowly. Below, we have the empirical autocorrelations of each parameter for lag 1 through 20.

Lag	Constant ACF	Length ACF	Weight ACF	Location ACF
1	0.0217	0.0440	0.0323	-0.0364
2	0.0206	-0.0115	0.0051	-0.0633
3	-0.0277	-0.0262	0.0173	0.0279
4	0.0003	-0.0292	-0.0440	-0.0239
5	-0.0211	0.0100	0.0448	-0.0355
6	0.0262	0.0338	-0.0289	-0.0233
7	-0.0042	0.0002	0.0060	-0.0110
8	-0.0046	0.0008	-0.0253	-0.0196
9	-0.0441	0.0084	-0.0129	0.0027
10	-0.0134	-0.0543	-0.0330	-0.0094
11	-0.0326	-0.0359	-0.0276	0.0037
12	-0.0336	-0.0317	-0.0071	-0.0691
13	-0.0298	-0.0355	-0.0105	-0.0043
14	0.0306	0.0148	-0.0247	-0.0308
15	-0.0279	-0.0140	0.0170	0.0478
16	0.0025	-0.0184	0.004	-0.0337
17	0.0333	-0.0304	-0.0972	0.0034
18	0.0355	0.0366	-0.0157	0.0179
19	0.0274	0.01955	-0.0129	-0.0250
20	0.0140	-0.0173	-0.0578	0.0212

Table 1.6: Bayesian Logistic Regression with Length, Weight, and Location. Empirical Autocorrelations

To conclude this section, we consider a Bayesian logistic regression model with uncorrelated spatial effects:

$$\begin{aligned} y_{ij}|\underline{\beta}, \underline{\nu_i}, \sigma^2 &\sim Ber(logit^{-1}(\underline{x_{ij}}'\underline{\beta} + \underline{\nu_i})) \\ \underline{\beta} &\sim N(\underline{\theta}, \Sigma) \\ \underline{\nu_i}|\sigma^2 &\overset{iid}{\sim} N(\underline{\theta}, 100\Sigma) \\ \underline{\sigma^2} &\sim \frac{1}{(1+\sigma^2)^2} \end{aligned} \tag{1.8}$$

where the expectation of y_{ij} depends on the spatial random effect ν_i of location i as well as the covariates. The prior distribution of σ^2, the variance of the random effects, is proper but does not have finite moments of any order. For other choices of prior distributions for σ^2, consult Gelman (2006).

The result of a Metropolis sampler simulation with 101000 iterations, 1000 burn-in terms, and sampling of every 100 terms thereafter is given below:

Parameter	Mean	Std. Dev.	Num. SE	0.95 Cred. Int.	
Constant	-0.963	2.70	0.0945	-6.16	4.30
Length	0.112	0.0871	0.00316	-0.0550	0.294
Weight	0.00105	0.00175	0.000058	0.00464	0.00220
Location	-0.245	0.0915	0.00344	-0.415	-0.0821
Variance	0.0180	0.0665	0.00224	0.000358	0.179

Table 1.7: Bayesian Logistic Regression with Random Effects. Length, Weight, Location, and Variance. Posterior Estimates.

As with the previous, simpler Bayesian model, only weight and location are significant because their respective 95 percent credible intervals do not contain zero. We now also have the variance parameter σ^2 whose simulated values are not useful on their own.

We monitor convergence of the Metropolis sampler by looking at the sample path autocorrelations of each parameter. Again, convergence is slow because the autocorrelations do not wash-out quickly as the lag increases.

1.3 Justification for AR(1) Spatial Effects

Below, we discuss the auto-correlations and cross-correlations of average catfish length and weight, respectively, for each location. Shumway and Stoffer (2006) define and provide interpretations of these concepts in the time series context.

The cross-covariance function of two series x_t and y_t of lag h is defined as

Lag	Const. ACF	L ACF	W ACF	Loc ACF	Var ACF
1	-0.0062	0.0014	-0.0218	0.0313	0.0913
2	0.0016	0.0173	0.0057	0.0421	-0.0106
3	-0.0112	-0.0255	-0.053	0.054	0.0235
4	-0.0253	-0.0034	0.0211	0.0014	0.0136
5	0.0601	0.0539	0.0085	0.1109	-0.0227
6	0.0697	0.0561	0.0247	0.0267	-0.0203
7	0.0068	0.011	0.0412	0	0.0245
8	0.022	-0.0146	-0.0571	0.0289	0.0786
9	-0.0154	-0.0118	0.0244	0.0127	0.0116
10	-0.0167	-0.0052	-0.0222	0.0205	0.0049
11	-0.0454	-0.014	0.0062	-0.0446	0.0206
12	-0.0116	-0.0144	0.017	-0.084	0.0594
13	0.0075	0.0418	0.0334	0.0624	-0.0115
14	-0.0102	-0.0309	0.0142	0.006	-0.0085
15	0.0446	0.0193	-0.0289	0.0016	-0.0144
16	0.05	0.0673	0.0398	0.0185	-0.0286
17	-0.032	-0.0438	-0.0001	-0.0098	-0.0071
18	0.0074	0.0298	0.0238	0.012	-0.0314
19	-0.0126	0.025	0.0104	0.0305	-0.0134
20	0.007	0.0026	0.0058	-0.0282	-0.0014

Table 1.8: Uncorrelated Spatial Effects Model. Length, Weight, Location, and Variance. Empirical Autocorrelations

$$\gamma_{xy}(h) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)]. \tag{1.9}$$

The cross-correlation function of x_t and y_t of lag h is

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_{xx}(0)\gamma_{yy}(0)}} \tag{1.10}$$

and the auto-correlation function of lag h is

$$\rho_x(h) = \rho_{xx}(h). \tag{1.11}$$

The cross-correlation function also satisfieds

$$\rho_{xy}(h) = \rho_{yx}(-h). \tag{1.12}$$

To estimate such quantities, replace expecations with sample means.

Catfish data for TRM 340 (location 14) was ignored because no catfish was caught at TRM 335 (location 13). However, we are ultimately interested in analysis involving locations 13 and 14. They are dealt in detail in later chapters.

Both autocorrelations peak at $lag = 1$, suggesting that unobserved random spatial effects may be adequately described as an AR(1) process. Nonetheless, the cross-correlation of lag 0 is fairly close to 1 and suggests that the average length and average weight convey similar information.

1.4 Overview of Thesis

The purpose of the thesis is to predict the probability that a catfish is legally toxic - containing 5 ppm or more DDT. In chapter 2, we discuss the geography of the Tennessee river, the history of DDT polution along the river, and chemical properties of DDT. In chapter 3, we introduce positive correlation amongst the spatial effects. For each location, we calculate the probability a catfish is toxic, but we do not account for missing observations. Chapter 4 covers the imputation procedure to be used, and we apply it naively to predict probabilities. In chapter 5, we combine

Lag	Length AC	Weight AC	Length Weight CC
-6	NA	NA	-0.152
-5	NA	NA	-0.281
-4	NA	NA	-0.387
-3	NA	NA	-0.509
-2	NA	NA	-0.132
-1	NA	NA	0.293
0	1	1	0.797
1	0.469	0.349	0.372
2	0.040	0.019	0.213
3	-0.247	-0.108	0.098
4	-0.437	-0.073	-0.018
5	-0.402	-0.218	-0.175
6	-0.151	-0.290	-0.152

Table 1.9: Auto-correlations and cross-correlation of average catfish length and weight at each location

the models developed in chapters 3 and 4 into a single model which performs both imputation of catfish data and prediction of probabilities. We end in chapter 6 by summarizing our findings and indicating further issues of interest.

Chapter 2

Tennessee River Basin and Pollution

2.1 Geography of the Tennessee River

The Tennessee River is the largest tributary of the Ohio River. It is approximately 650 miles (1,046 km) long and is located in the southeastern United States in the Tennessee Valley. The river was once popularly known as the Cherokee River, among other names.

The Tennessee River is formed at the confluence of the Holston and French Broad Rivers on the east side of Knoxville, Tennessee. From Knoxville, it flows southwest through East Tennessee toward Chattanooga before crossing into Alabama. It loops through northern Alabama and eventually forms a small part of the state's border with Mississippi, before returning to Tennessee. At this point, it defines the boundary between Tennessee's other two regions: Middle and West Tennessee. The Tennessee-Tombigbee Waterway, a U.S. Army Corps of Engineers project providing navigation on the Tombigbee River and a link to the Port of Mobile, enters Tennessee near the Tennessee-Alabama-Mississippi boundary. This waterway reduces the navigation distance from Tennessee, north Alabama, and northern Mississippi to the Gulf of Mexico by hundreds of miles. The final part of the Tennessee's run is in Kentucky, where it separates the Jackson Purchase from the rest of the state. It then flows into the Ohio River at Paducah, Kentucky. It is one of a very few rivers in the United States which leave a state and then re-enter it; the Cumberland River is another such river.

The river has been dammed numerous times, primarily by Tennessee Valley Authority (TVA) projects. The placement of TVA's Kentucky Dam on the Tennessee River and the Corps' Barkley Dam on the Cumberland River directly led to the creation of Land Between the Lakes. A navigation canal located at Grand Rivers, Kentucky links Kentucky Lake and Lake Barkley. The canal allows for a shorter trip for river traffic going from the Tennessee to most of the Ohio River, and for traffic going down the Cumberland River toward the Mississippi.

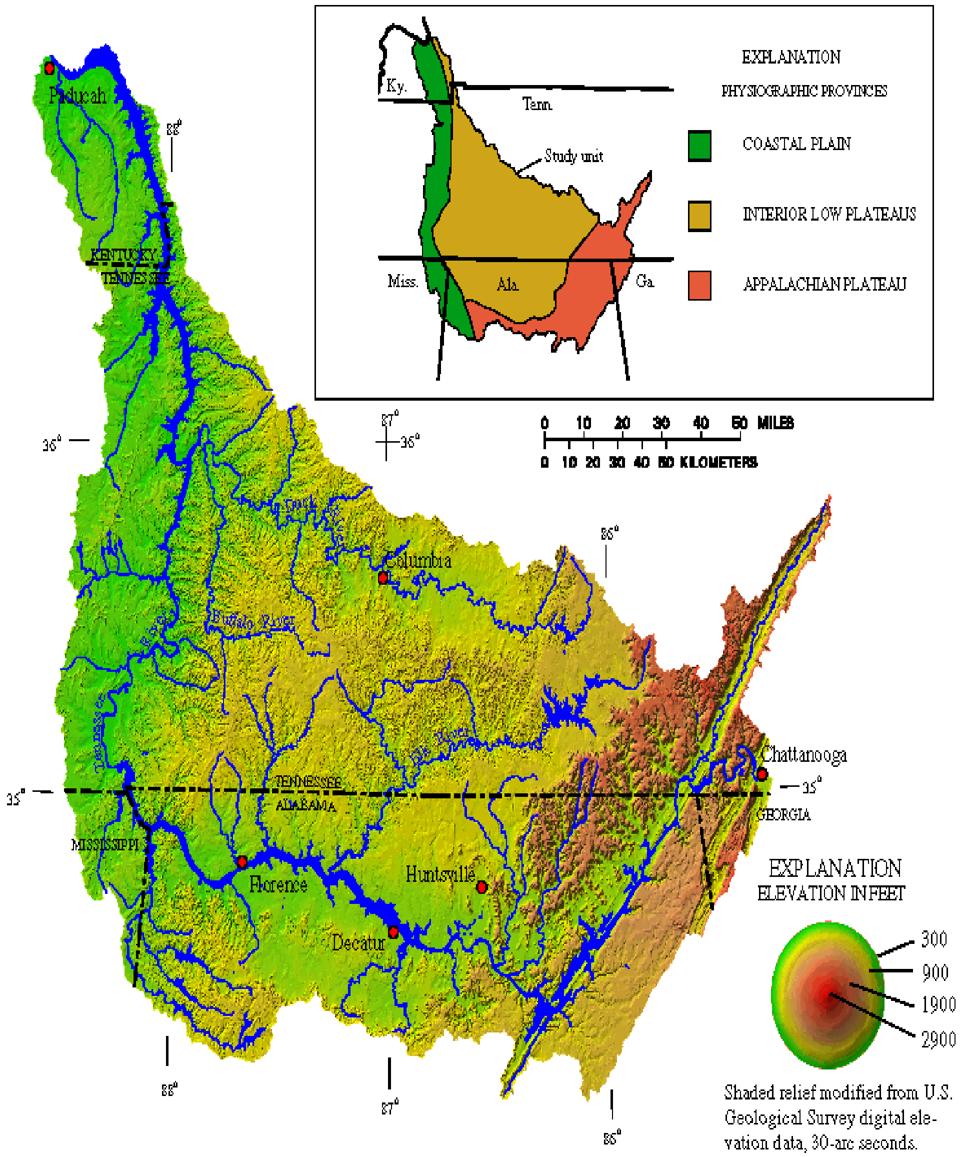

Figure 2.1: Map of Tennessee River (US Army Corps of Engineers)

2.2 Chemistry of the Tennessee River

Between 1947 and 1970, the Olin Corpororation manufactured dichloro-diphenyl-tricholoroethane (DDT) within the Redstone Arsenal and released waste water into the Huntsville Spring Branch of the river. Fish living in the branch were contaminated with DDT over the years by an estimated 408.8 tons of contaminants. Following public concern, the State of Alabama, the U.S. Environmental Protection Agency (EPA) and Olin Corporation entered a Consent Decree (CD) on May 31, 1983 to reduce DDT content of fillets of channel catfish, largemouth bass, and smallmouth buffallo fish to below 5 parts per million.

The deadline for achieving the performance standard is December 31, 2002, for channel catfish, and Dec. 31, 2007, for smallmouth buffalo. The review pannel for assesing performance consists of EPA, TVA, U.S. Fish and Wildlife Service, Department of the Army, the State of Alabama, and nonvoting participants from the town of Triana, Alabama and from Olin. As part of the review process, the U.S. Army Corps of Engineers collected fish specimens along the Tennessee River and its three tributaries in the summer of 1980. They recorded the length (in centimeters), weight (in grams), and the DDT concentration (parts per million, ppm) in the fillet of the fish. See appendix A for this data.

2.3 DDT and its Toxicity

DDT was the first modern pesticide and is arguably the best known organic pesticide. It is a highly hydrophobic colorless solid with a weak, chemical odor that is nearly insoluble in water but has a good solubility in most organic solvents, fat, and oils. DDT is also known under the chemical names 1,1,1-trichloro-2,2-bis(p-chlorophenyl)ethane and dichloro-diphenyl-trichloroethane (from which the abbreviation was derived).

DDT was developed as the first of the modern insecticides early in World War II. It was initially used with great effect to combat mosquitoes spreading malaria, typhus, and other insect-borne human diseases among both military and civilian populations, and as an agricultural insecticide. Paul Hermann Muller, a Swiss chemist of Geigy Pharmaceutical, was awarded the Nobel Prize in Physiology or Medicine in 1948 "for his discovery of the high efficiency of DDT as a contact poison against several arthropods."

DDT has potent insecticidal properties; it kills by opening sodium ion channels in insect neurons, causing the neuron to fire spontaneously. This leads to spasms and eventual death. Insects with certain mutations in their sodium channel gene may be resistant to DDT and other similar insecticides.

The 1970s ban in the U.S. took place amid a climate of public mistrust of the scientific and industrial community, following such fiascoes as Agent Orange and use of the hormone diethylstilbestrol (DES). In addition, the placement of the bald eagle on the endangered species list was also a strong factor leading to its being banned in the United States. The overuse of DDT was claimed to be a major factor in the bald eagle population decline - a claim that has fallen into dispute.

DDT is a persistent organic pollutant with a reported half life of between 2-15 years, and is immobile in most soils. Its half life is 56 days in lake water and approximately 28 days in river water. Routes of loss and degradation include runoff, volatilization, photolysis and biodegradation (aerobic and anaerobic). These processes generally occur slowly. Breakdown products in the soil environment are DDE (1,1-dichloro-2,2-bis(p-dichlorodiphenyl)ethylene) and DDD (1,1-dichloro-2,2-bis(p-chlorophenyl)ethane), which are also highly persistent and have similar chemical and physical properties. These products together are known as total DDT.

DDT is an organochlorine. Some organochlorines have been shown to have weak estrogenic activity; that is, they are chemically similar enough to estrogen to trigger hormonal responses in contaminated animals. This hormonal-mimicking activity has been observed when DDT is used in laboratory studies involving mice and rats as test subjects, but available epidemiological evidence does not indicate that these effects have occurred in humans as a result of DDT exposure.

DDT in small quantities has very little effect on birds; its primary metabolite, DDE, has a much greater impact. DDT and DDE have little impact on some other birds, such as the chicken. DDT is highly toxic to aquatic life, including crayfish, daphnids, sea shrimp and many species of fish. DDT may be moderately toxic to some amphibian species, especially in the larval stages. In addition to acute toxic effects, DDT may bioaccumulate significantly in fish and other aquatic species, leading to long-term exposure to high concentrations.

There are no substantial scientific studies which prove that DDT is particularly toxic to humans or other primates, compared to other widely-used pesticides. DDT can be applied directly to clothes and used in soap, with no demonstrated ill effects. There is no convincing evidence that DDT or its metabolite DDE increase human cancer risk. Mainly on the basis of animal data, DDT is classified as a possible carcinogen (class 2B) by the International Agency for Research on Cancer (IARC) and as class B2, reasonably anticipated human carcinogen by the US National Toxicology Program. This group also contains substances such as coffee and gasoline.

Chapter 3

Incorporating Spatial Effects

In this chapter, we develop a model to explain pollution of catfish in the Tennessee River basin that incorporates spatial information, i.e., distance from the the mouth of the river, differently from the other covariates (length and weight). Procedures to sample from the posterior distribution will be explored, along with a discussion of the computational and theoretical technicalities involved.

3.1 Hierarchical Bayesian Spatial Model

We develop a hierarchical Bayesian model to explain whether or not a catfish of certain length and weight found at a certain location along the river will have more than 5 ppm of DDT or less. This model extends the Bayesian logistic random effects model discussed in chapter 1 and introduces correlation between random effects of adjacent regions so that,

$$y_i|\nu_i, \underline{\beta}, n_i \sim Bin(n_i, logit^{-1}(\underline{x_i}'\underline{\beta} + \nu_i)), i = 1, ..., L \tag{3.1}$$

$$\nu_i|\nu_{i-1}, \gamma, \sigma^2 \sim N(\gamma\nu_{i-1}, \sigma^2), i = 1, ..., L, \nu_0|\gamma, \tau \sim N(0, \frac{\sigma^2}{1-\gamma^2}) \tag{3.2}$$

$$\underline{\beta} \sim N(\underline{\theta}, \Sigma) \tag{3.3}$$

$$\gamma, \rho \overset{iid}{\sim} U(0, 1) \tag{3.4}$$

$$\pi(\sigma^2) \propto \frac{1}{(1+\sigma^2)^2}. \tag{3.5}$$

Here, y_i represents the number of toxic catfish observed out of n_i catfish caught at location i, $\underline{x_i}$ is the average of the vector of covariates of fish observed at location i, with $x_{i1} = 1$, x_{i2} = length, x_{i3} = weight, and x_{i4} = location. Given all other parameters, $logit^{-1}(\underline{x_i}'\underline{\beta} + \nu_i)$ is the probability that a catfish at location i is toxic.

Observe that y_i depends on non-spatial covariates and location in different ways. The effects of weight and height on y_i determined by $\underline{\beta}$ through the linear function $\underline{x_i}'\underline{\beta}$. There are however spatial effects $\underline{\nu}$. Such spatial effects are unobserved, and offers some insight into why straightforward logistic regression that treats spatial information on equal grounds with other covariates can be misleading.

The unobserved spatial effects, $\underline{\nu}$, are assumed to arise from an AR(1) process. Spatial correlation γ between ν_i and ν_{i-1} (adjacent regions), is taken to be positive but strictly less than 1 to assure second-order stationarity. These assumptions are equivalent to the spatial correlation falling off with distance, and excludes hypothetical situations under which the effect of pollution fails to die off with distance for rivers of arbitrary length.

As in other Bayesian models discussed so far, the coefficient vector $\underline{\beta}$ has a normal prior distribution with mean $\underline{\theta}$ equal to the frequentist MLE and the covariance matrix Σ given as a blow-up factor (100) times the frequentist estimate for the covariance matrix so that the prior is diffuse.

3.2 Full Posterior Density for Hierarchical Model

To simplify the algebra and computations, we make the transformations

$$\begin{aligned} \phi_i &= \underline{x_i}'\underline{\beta} + \nu_i, i = 0, \ldots, L & (3.6) \\ \tau &= \frac{1}{\sigma^2} & (3.7) \end{aligned}$$

where $\underline{x_0} = \underline{x_1}$ for convenience, as index $i = 0$ corresponds to a location along the river where catfish may actually be observed. The poseterior density of the transformed model, including the Jacobian of transformation, is then

$$\begin{aligned} \pi(\underline{\beta}, \underline{\phi}, \gamma, \tau|\underline{y}) \propto \prod_{i=1}^{L} Bin_{y_i}(n_i, logit^{-1}(\underline{x_i}'\underline{\beta})) \\ \times \prod_{i=1}^{L} \tau^{\frac{1}{2}} \exp(-\frac{\tau}{2}((\phi_i - \underline{x_i}'\underline{\beta}) - (\phi_{i-1} - \underline{x_{i-1}}'\underline{\beta}))^2) \\ \times((1-\gamma^2)\tau)^{\frac{1}{2}} \exp(-\frac{(1-\gamma^2)\tau}{2}((\phi_0 - \underline{x_0}'\underline{\beta})^2)) \\ \times N_{\underline{\beta}}(\underline{\theta}, \Sigma)\frac{1}{(1+\tau)^2}U_\gamma(0,1)U_\rho(0,1) \qquad (3.8) \end{aligned}$$

3.3 Gibbs Sampler for Hierarchical Model

The above hierarchical model leads to the following conditional posterior densities for parameters, all of which are used to construct a Markov chain using the Gibbs sampler technique.

$$\begin{aligned}
&\pi(\underline{\beta}|\underline{\phi},\gamma,\tau,\underline{y}) \\
&= N((\Sigma^{-1}+\tau\sum(\underline{x_i}-\gamma\underline{x_{i-1}})\otimes(\underline{x_i}-\gamma\underline{x_{i-1}})), \\
&(\Sigma^{-1}+\tau\sum(\underline{x_i}-\gamma\underline{x_{i-1}})\otimes(\underline{x_i}-\gamma\underline{x_{i-1}})) \\
&(\Sigma^{-1}\underline{\theta}+\tau\sum(\phi_i-\gamma\phi_{i-1})(\underline{x_i}-\gamma\underline{x_{i-1}})))
\end{aligned} \tag{3.9}$$

$$\begin{aligned}
\pi(\tau|\underline{\beta},\underline{\phi},\gamma,\underline{y}) &= Ga(\frac{L+3}{2},\frac{s^2}{2}), \\
s^2 &= \sum(\nu_i-\gamma\nu_{i-1})^2+(1-\gamma^2)\nu_0^2
\end{aligned} \tag{3.10}$$

$$\pi(\phi_0|\underline{\beta},\underline{\phi_{-0}},\gamma,\underline{y}) = N(\gamma\phi_1,\tau^{-1}) \tag{3.11}$$

$$\begin{aligned}
&\pi(\phi_i|\underline{\beta},\underline{\phi_{-i}},\gamma,\underline{y}) \propto \\
&Bin_{y_i}(n_i, logit^{-1}(\underline{x_i}'\underline{\beta})) \\
&\times N(\underline{x_i}'\underline{\beta}+\frac{\gamma}{1+\gamma^2}(\phi_{i+1}+\phi_{i-1}-\underline{x_{i+1}}'\underline{\beta}-\underline{x_{i-1}}'\underline{\beta},\tau^{-1})), i=1,\ldots,L-1
\end{aligned} \tag{3.12}$$

$$\begin{aligned}
&\pi(\phi_L|\underline{\beta},\underline{\phi_{-L}},\gamma,\underline{y}) \propto \\
&Bin_{y_i}(n_L, logit^{-1}(\underline{x_L}'\underline{\beta}))N(\gamma\phi_{L-1},\tau^{-1})
\end{aligned} \tag{3.13}$$

$$\begin{aligned}
\pi(\gamma|\underline{\beta},\underline{\phi},\tau,\underline{y}) &\propto \sqrt{1-\gamma^2}N_\gamma(\mu,s^2)I_{(0,1)}(\gamma), \\
\mu &= \frac{\sum_{i=1}^L \nu_i\nu_{i-1}}{\sum_{i=1}^L \nu_i^2}, \\
s^2 &= \frac{1}{\tau\sum_{i=1}^L \nu_i^2}
\end{aligned} \tag{3.14}$$

To sample from the conditional posterior densities of γ we use the accept-reject method.

```
SAMPLING $\gamma$
(1) SAMPLE $U \sim U(0,1)$,
    $\gamma \sim N_\gamma(\mu, s^2)I_{(0,1)}(\gamma)$
(2) CALCULATE $prob = \sqrt{1-\gamma^2}$
(3) If $U > prob$
    then RETURN $\gamma$
    else RETURN TO (1)
```

Figure 3.1: Sampling γ from conditional posterior density

To sample from $\phi_i, i = 1, \ldots, L$, we resort to a combination of griddy Gibbs sampling and a transformation of variables. We first consider the transformation

$$r_i = logit^{-1}(\phi_i), 0 < r < 1 \tag{3.15}$$

Upon the transformation, we obtain the conditional posterior density

$$\pi(r_i|\phi_{-i}, \underline{\beta}, \gamma, \tau, \underline{y}) \propto r_i^{-1}(1-r_i)^{-1}Bin_{y_i}(n_L, r_i)N_{logit(r_i)}(\mu_i, \tau^{-1}). \tag{3.16}$$

We discretize the above density using 100 evenly spaced grid points and draw r_i from the discrete distribution just created. Then, $\phi_i = logit^{-1}(r_i)$ is a sample from the conditional posterior distribution of ϕ_i.

3.4 Simulation Results

In this section, we examine a simulation run of the model discussed in this chapter, and examine the scientific conclusions.

We run the Gibbs sampler defined in previous sections with 101000 iterations, 1000 burn-in terms discarded, and sampling every 100 samples after burn-in as to obtain 1000 samples from the full posterior distribution.

In Table 3.1 we examine posterior quantities of interest. We show posterior quantities for $p = logit^{-1}(\phi)$ rather than for ϕ because we are interested in the probabilities. There are two major observations. First, the 95 percent credible intervals of $\underline{\beta}$ all contain zero, so in particular the average length and weight of catfish at each location in the study is irrelevant according to the model. Second, the general trend for the p's is to decrease as we move downstream (larger location number). Locations where only 3 or 4 out of 6 catfish were toxic have 95 percent credible intervals of p containing 0.50.

	Mean	Std. Dev.	Num. Std. Err.	0.95 Cred. Int.
γ	0.633	0.292	0.00922	(0.0510, 0.995)
τ	1.24	0.857	0.0271	(0.300, 3.48)
β_1	2.06	4.31	0.136	(-6.98, 11.4)
β_2	0.249	1.62	0.0512	(-3.03, 3.42)
β_3	1.50	1.53	0.0484	(-1.60, 4.49)
β_4	-0.95	1.03	0.0326	(-2.37, -0.038)
ϕ_0	2.82	1.49	0.0473	(-0.0418, 5.84)
p_1	0.894	0.0977	0.00309	(0.615, 0.995)
p_2	0.972	0.0443	0.00140	(0.845, 0.995)
p_3	0.862	0.0922	0.00291	(0.645, 0.985)
p_4	0.818	0.111	0.00351	(0.545, 0.975)
p_5	0.882	0.0925	0.00293	(0.625, 0.985)
p_6	0.947	0.0743	0.00235	(0.735, 0.995)
p_7	0.822	0.118	0.00374	(0.545, 0.975)
p_8	0.719	0.136	0.00429	(0.415, 0.925)
p_9	0.714	0.145	0.00457	(0.385, 0.935)
p_{10}	0.927	0.0763	0.00241	(0.705, 0.995)
p_{11}	0.575	0.154	0.00488	(0.265, 0.855)
p_{12}	0.691	0.140	0.00443	(0.395, 0.915)

Table 3.1: Gibbs Sampler Posterior Quantities after burn-in

To assess independence of the samples after burn in, we look at the lag 1 autocorrelations. It indicates that there is little autocorrelation between the samples after burn-in. Hence, we may conclude that the samples obtained can be treated as values from the full posterior density.

Parameter	AC
γ	0.00634
τ	-0.0301
β_1	0.0314
β_2	-0.0250
β_3	-0.00429
β_4	0.0421
ϕ_0	-0.000676
p_1	-0.0180
p_2	0.0232
p_3	-0.000324
p_4	0.0529
p_5	-0.0350
p_6	-0.00526
p_7	-0.0270
p_8	0.0183
p_9	0.00811
p_{10}	-0.0139
p_{11}	0.0220
p_{12}	-0.0265

Table 3.2: Gibbs sampling autocorrelation after burn-in

Chapter 4

Imputing Missing Catfish Data

In the study conducted by the Army Corps of Engineers on along the Tennessee River, no catfish were observed at location TRM335. Ideally, we would fit the model developed chapter 2 to all observation locations between TRM270 and TRM340 to determine the effect of missing observations. To accomplish this goal, we develop a simple hierarchical Bayesian model to impute length and weight of $n_{13} = 6$ hypothetical catfish at TRM 335 (location 13).

4.1 Hierarchical Bayesian Model

Let $\underline{z_{ij}}$ denote the vector of the log-transformed length and weight of catfish j at location i: $z_{ij,L}$ is log-transformed length, and $z_{ij,W}$ is log-transformed weight. Furthermore, let $\underline{\mu_i} = (\mu_{i,L}, \mu_{i,W})'$.

We examine the model

$$\underline{z_{ij}}|\underline{\mu_i}, \rho, \sigma_L^2, \sigma_W^2 \sim N(\underline{\mu_i}, \Sigma) \tag{4.1}$$

$$\begin{aligned} \mu_{i,L}|\mu_{i-1,L},\gamma_L,\delta_L^2 &\sim N(\gamma_L\nu_{i-1,L},\delta_L^2), i=1,...,L, \\ \mu_{0,L}|\gamma_L,\delta_L^2 &\sim N(0,\delta_L^2(1-\gamma_L^2)^{-1}) \end{aligned} \tag{4.2}$$

$$\begin{aligned} \mu_{i,W}|\mu_{i-1,W},\gamma_W,\delta_W^2 &\sim N(\gamma_W\nu_{i-1,W},\delta_W^2), i=1,...,L, \\ \mu_{0,W}|\gamma_W,\delta_W^2 &\sim N(0,\delta_W^2(1-\gamma_W^2)^{-1}) \end{aligned} \tag{4.3}$$

$$\delta_L^2 = \frac{1-\epsilon_L}{\epsilon_L}\sigma_L^2 \tag{4.4}$$

$$\delta_W^2 = \frac{1-\epsilon_W}{\epsilon_W}\sigma_W^2 \tag{4.5}$$

$$\sigma_L^2,\sigma_W^2 \overset{iid}{\sim} (1+\sigma^2)^{-2}I_{[0,\infty)}(\sigma^2) \tag{4.6}$$

$$\rho,\epsilon_L,\epsilon_W,\gamma_L,\gamma_W \overset{iid}{\sim} U(0,1) \tag{4.7}$$

where

$$\Sigma = \begin{bmatrix} \sigma_L^2 & \rho\sigma_L\sigma_W \\ \rho\sigma_L\sigma_W & \sigma_W^2 \end{bmatrix} \tag{4.8}$$

There are several points to note. First, $\underline{z_{13,j}}, j=1,\ldots,n_{13}=6$ unobserved parameters but are the same in every other respect as data for other catfish.

4.2 Full Posterior Density for Hierarchical Model

To simplify derivations, we make the following transformation

$$\tau_L = \sigma_L^{-2}, \tau_W = \sigma_W^{-2} \tag{4.9}$$

Then the full posterior density of the model is given by

$$\begin{aligned} \pi(\underline{\mu}, \underline{\gamma}, \underline{\tau}, \rho, \underline{\epsilon}, \underline{z_{13}}|\underline{z_{-13}}) \propto \\ \prod_{i,j} N_{z_{ij}}(\underline{\mu_i}, \Sigma) \\ \times \prod_{i=1}^{L} \delta_L^{-1} \exp(-\frac{1}{2\delta_L^2}(\mu_{i,L} - \gamma_L \mu_{i,L})^2) \\ \times((1-\gamma_L^2)\delta_L^{-2})^{\frac{1}{2}} \exp(-\frac{1-\gamma_L^2}{2\delta_L^2}\mu_{0,L}^2) \\ \times \prod_{i=1}^{L} \delta_W^{-1} \exp(-\frac{1}{2\delta_W^2}(\mu_{i,W} - \gamma_W \mu_{i,W})^2) \\ \times((1-\gamma_W^2)\delta_W^{-2})^{\frac{1}{2}} \exp(-\frac{1-\gamma_W^2}{2\delta_W^2}\mu_{0,W}^2) \\ \times(1+\sigma_L^2)^{-2} I_{[0,\infty)}(\sigma_L^2)(1+\sigma_W^2)^{-2} I_{[0,\infty)}(\sigma_W^2) \\ \times U_\rho(0,1)U_{\gamma_L}(0,1)U_{\gamma_W}(0,1)U_{\epsilon_L}(0,1)U_{\epsilon_W}(0,1) \end{aligned} \tag{4.10}$$

4.3 Gibbs Sampler for Hierarchical Model

We now examine the conditional posterior densities of each parameter to construct a Gibbs sampler for simulation.

$$\begin{aligned}\pi(\gamma_L|\underline{\mu},\gamma_W,\underline{\tau},\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) &\propto \\ \sqrt{1-\gamma_L^2}N_{\gamma_L}(\theta,s^2)I_{(0,1)}(\gamma_L), \\ \theta &= \frac{\sum_{i=1}^{L}\mu_{i,L}\mu_{i-1,L}}{\sum_{i=1}^{L}\mu_{i,L}^2}, \\ s^2 &= \frac{1}{\tau_L\sum_{i=1}^{L}\mu_{i,L}^2}\end{aligned} \tag{4.11}$$

$$\begin{aligned}\pi(\gamma_W|\underline{\mu},\gamma_L,\underline{\tau},\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) &\propto \\ \sqrt{1-\gamma_W^2}N_{\gamma_W}(\theta,s^2)I_{(0,1)}(\gamma_W), \\ \theta &= \frac{\sum_{i=1}^{L}\mu_{i,W}\mu_{i-1,W}}{\sum_{i=1}^{L}\mu_{i,W}^2}, \\ s^2 &= \frac{1}{\tau_W\sum_{i=1}^{L}\mu_{i,W}^2}\end{aligned} \tag{4.12}$$

$$\begin{aligned}\pi(\rho|\underline{\mu},\underline{\gamma},\underline{\tau},\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) &\propto \\ I_{[0,1]}(\rho)(1-\rho^2)^{-\frac{1}{2}\sum n_i}\exp(-\frac{1}{2}\sum_{ij}\underline{z_{ij}}'\Sigma^{-1}\underline{z_{ij}})\end{aligned} \tag{4.13}$$

$$\begin{aligned}\pi(\epsilon_L^*|\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\epsilon_W,\underline{z_{13}},\underline{z_{-13}}) &\propto \\ (1+\epsilon_L^*)^{-2}I_{[0,\infty)}(\epsilon_L^*)\times Ga(\frac{L+3}{2},s^2), \\ s^2 = \frac{\tau_L}{2}(\sum_{i=1}^{L}(\mu_{i,L}-\gamma_L\mu_{i-1,L})^2+(1-\gamma_L^2)\mu_{0,L}^2), \\ \epsilon_L &= \frac{\epsilon_L^*}{1+\epsilon_L^*}\end{aligned} \tag{4.14}$$

$$\begin{aligned}\pi(\epsilon^*_W|\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\epsilon_L,\underline{z_{13}},\underline{z_{-13}}) \propto \\ (1+\epsilon^*_W)^{-2}I_{[0,\infty)}(\epsilon^*_W)\times Ga(\frac{L+3}{2},s^2), \\ s^2=\frac{\tau_W}{2}(\sum_{i=1}^{L}(\mu_{i,W}-\gamma_W\mu_{i-1,W})^2+(1-\gamma_W^2)\mu_{0,W}^2), \\ \epsilon_W=\frac{\epsilon^*_W}{1+\epsilon^*_W}\end{aligned} \tag{4.15}$$

$$\begin{aligned}\pi(\tau_L|\underline{\mu},\underline{\gamma},\tau_W,\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) \propto \\ (1+\tau_L)^{-2}I_{[0,\infty)}(\tau_L)\exp(-\frac{s^2}{2}\tau_L) \\ \times\exp(\frac{\rho}{1-\rho^2}\sum_{ij}\ln(L_{ij})\ln(W_{ij})\tau_L^{\frac{1}{2}})\end{aligned} \tag{4.16}$$

$$\begin{aligned}\pi(\tau_W|\underline{\mu},\underline{\gamma},\tau_L,\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) \propto \\ (1+\tau_W)^{-2}I_{[0,\infty)}(\tau_W)\exp(-\frac{s^2}{2}\tau_W) \\ \times\exp(\frac{\rho}{1-\rho^2}\sum_{ij}\ln(L_{ij})\ln(W_{ij})\tau_W^{\frac{1}{2}})\end{aligned} \tag{4.17}$$

$$\pi(\mu_{0,L}|\underline{\mu_{-0}},\mu_{0,W},\underline{\gamma},\tau_L,\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) \propto N(\gamma_L\mu_{1,L},\tau^{-1}) \tag{4.18}$$

$$\pi(\mu_{0,W}|\underline{\mu_{-0}},\mu_{0,L},\underline{\gamma},\tau_W,\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) \propto N(\gamma_W\mu_{1,W},\tau^{-1}) \tag{4.19}$$

$$\begin{aligned}
\pi(\mu_{i,L}|\underline{\mu_{-i}}, \mu_{i,W}, \underline{\gamma}, \tau_L, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}) \propto N(\eta_{i,L}, \sigma_L^2(1+n_i(1-\rho^2)^{-1}+\gamma_L^2)^{-1}),\\
\eta_{i,L} = (\gamma_L(\mu_{i+1,L}+\mu_{i-1,L}) + (1-\rho^2)^{-1}\sum_j \ln(L_{i,j}) - \rho(1-\rho^2)^{-1}\\
\times\sigma_L\sigma_W^{-1}\sum_j(\ln(W_{i,j})-\mu_{i,W}))\\
\times\sigma_L^2(1+n_i(1-\rho^2)^{-1}+\gamma_L^2)^{-1}, 0<i<L
\end{aligned}\tag{4.20}$$

$$\begin{aligned}
\pi(\mu_{i,W}|\underline{\mu_{-i}}, \mu_{i,L}, \underline{\gamma}, \tau_W, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}) \propto N(\eta_{i,W}, \sigma_W^2(1+n_i(1-\rho^2)^{-1}+\gamma_W^2)^{-1}),\\
\eta_{i,W} = (\gamma_W(\mu_{i+1,W}+\mu_{i-1,W}) + (1-\rho^2)^{-1}\sum_j \ln(W_{i,j}) - \rho(1-\rho^2)^{-1}\\
\times\sigma_W\sigma_L^{-1}\sum_j(\ln(L_{i,j})-\mu_{i,W}))\\
\times\sigma_W^2(1+n_i(1-\rho^2)^{-1}+\gamma_W^2)^{-1}, 0<i<L
\end{aligned}\tag{4.21}$$

$$\begin{aligned}
\pi(\mu_{L,L}|\underline{\mu_{-L}}, \mu_{L,L}, \underline{\gamma}, \tau_L, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}) \propto N(\eta_{L,L}, \sigma_L^2(1+n_L(1-\rho^2)^{-1})^{-1}),\\
\eta_{L,L} = (\gamma_L\mu_{L-1,L} + (1-\rho^2)^{-1}\sum_j \ln(L_{i,j}) - \rho(1-\rho^2)^{-1}\sigma_L\sigma_W^{-1}\\
\times\sum_j(\ln(W_{L,j})-\mu_{L,W}))\\
\times\sigma_L^2(1+n_i(1-\rho^2)^{-1})^{-1}, 0<i<L
\end{aligned}\tag{4.22}$$

$$\begin{aligned}
\pi(\mu_{L,W}|\underline{\mu_{-L}}, \mu_{L,W}, \underline{\gamma}, \tau_W, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}) \propto N(\eta_{L,W}, \sigma_W^2(1+n_L(1-\rho^2)^{-1})^{-1}),\\
\eta_{L,W} = (\gamma_W\mu_{L-1,W} + (1-\rho^2)^{-1}\sum_j \ln(W_{i,j}) - \rho(1-\rho^2)^{-1}\sigma_W\sigma_L^{-1}\\
\times\sum_j(\ln(L_{L,j})-\mu_{L,L}))\\
\times\sigma_W^2(1+n_i(1-\rho^2)^{-1})^{-1}, 0<i<L
\end{aligned}\tag{4.23}$$

4.4 Simulation Results

In this section, we look at a simulation run of the Gibbs sampler for the imputation model, using 101000 iterations, 1000 burn-in terms, and sampling every 100 samples after burn-in (1000 samples total) to study the full posterior distribution, in particular the posterior mean of unobserved catfish lengths and weights.

In table 4.1 we list posterior quantities for the simulation. Our main interest here are the imputed (log-transformed) weights and lengths, so only these paramters will be given. Lag 1 autocorrelations seen in Table 4.2 indicate that convergence is a reasonable assumption.

	Mean	Std. Dev.	Num. Std. Err.	0.95 Cred. Int.
$lnL_{13,1}$	3.753	2.027	0.064	(3.638,3.942)
$lnL_{13,2}$	3.885	1.771	0.056	(3.761,3.932)
$lnL_{13,3}$	3.951	2.153	0.068	(3.882,3.988))
$lnL_{13,4}$	3.777	2.261	0.071	(3.696,3.824)
$lnL_{13,5}$	3.769	2.026	0.064	(3.761,3.811)
$lnL_{13,6}$	3.683	1.903	0.060	(3.653,3.697)
$lnW_{13,1}$	7.118	3.487	0.110	(6.659,7.244)
$lnW_{13,2}$	6.340	3.612	0.114	(6.194,6.791)
$lnW_{13,3}$	6.741	3.505	0.111	(6.397,6.908)
$lnW_{13,4}$	6.833	4.110	0.129	(6.522,6.802)
$lnW_{13,5}$	6.968	3.713	0.117	(5.991,7.783)
$lnW_{13,6}$	6.592	3.100	0.098	(6.297,6.820)

Table 4.1: Gibbs Sampler Posterior Quantities after burn-in

Parameter	AC
ρ	
$\ln L_{13,1}$	0.053
$\ln L_{13,2}$	0.041
$\ln L_{13,3}$	0.055
$\ln L_{13,4}$	-0.046
$\ln L_{13,5}$	0.050
$\ln L_{13,6}$	-0.047
$\ln W_{13,1}$	-0.062
$\ln W_{13,2}$	-0.058
$\ln W_{13,3}$	0.053
$\ln W_{13,4}$	-0.049
$\ln W_{13,5}$	0.057
$\ln W_{13,6}$	0.048

Table 4.2: Gibbs sampling autcorrelation after burn-in

4.5 Model Fitting with Completed Data

In this section, we use a naive prediction to fit the model discussed in chapter 2 (101000 iterates, 1000 burn-in, sampling every 100 after burn-in) to all locations between location 1 (TRM275) and location 14 (TRM340), including the unobserved location 14 (TRM335). Imputed lengths and weights of catfish for location 13 (TRM335) to compute the average length and weight of location 13. To determine the proportion of toxic catfish at location 13, we simply take the average of the proportions at location 12 and location 14, so that $x_{13,2} = 45.02, x_{13,3} = 894.4$ before standardization, and $y_{13} = \frac{3}{6}$. Again, we revert from ϕ's to $p = logit^{-1}(\phi)$.

The primary observation to make is that, as in chapter 2, locations with observed proportions of toxic catfish less than $\frac{4}{6}$ has 95 percent credible intervals which contain zero, where as other locations have $p > 0.5$. This implies that catfish are likely to be toxic when they are caught in regions of high observed proportions.

Lag 1 autocorrelations indicate that the samples after burn-in are approximately independent.

In chapter 5, we will do the prediction more optimally.

	Mean	Std. Dev.	Num. Std. Err.	0.95 Cred. Int.
γ	0.526	0.275	0.0087	(0.072, 0.993)
τ	1.687	0.790	0.025	(0.373, 4.592)
β_1	1.269	2.898	0.016	(-5.463, 8.795)
β_2	0.436	4.137	0.131	(-4.506, 5.412)
β_3	1.194	1.618	0.053	(-2.134, 4.439)
β_4	-0.872	0.794	0.0251	(-2.137, 0.311)
ϕ_0	2.552	1.429	0.045	(-0.033, 4.937)
p_1	0.940	0.100	0.0032	(0.616, 0.995)
p_2	0.984	0.045	0.0014	(0.935, 0.999)
p_3	0.862	0.029	0.0009	(0.643, 0.982)
p_4	0.808	0.110	0.0035	(0.448, 0.920)
p_5	0.870	0.091	0.0029	(0.645, 0.966)
p_6	0.933	0.073	0.0023	(0.497, 0.994)
p_7	0.796	0.114	0.0036	(0.509, 0.937)
p_8	0.778	0.147	0.0046	(0.474, 0.924)
p_9	0.679	0.138	0.0044	(0.485, 0.870)
p_{10}	0.940	0.077	0.0024	(0.630, 0.992)
p_{11}	0.708	0.190	0.0060	(0.326, 0.909)
p_{12}	0.666	0.140	0.0044	(0.358, 0.898)
p_{13}	0.721	0.195	0.0062	(0.427, 0.871)
p_{14}	0.691	0.180	0.0057	(0.414, 0.864)

Table 4.3: Gibbs Sampler Posterior Quantities after burn-in

Parameter	AC
γ	0.005
τ	-0.026
β_1	-0.043
β_2	0.025
β_3	-0.004
β_4	0.0137
ϕ_0	-0.078
p_1	-0.011
p_2	0.023
p_3	-0.031
p_4	0.055
p_5	-0.036
p_6	0.006
p_7	-0.0270
p_8	-0.0182
p_9	0.00811
p_{10}	-0.019
p_{11}	-0.022
p_{12}	0.024
p_{13}	-0.028
p_{14}	0.025

Table 4.4: Gibbs sampling autocorrelation after burn-in

Chapter 5

Combined Model for Imputation and Prediction

In this chapter, we combine results of chapters 3 and 4 to obtain a single model which imputes missing data for unobserved location 13 and predict probabilities of toxicity at once. The important simplifying assumption made here is that all catfish from the same location are toxic with the same probability. This simplification is justified because the preliminary models all suggest that weight and height are not significant in determining such probabilities. Allowing the probability to differ amongst catfish in a given location complicates implementation and analysis but may better accomodate reality.

5.1 Hierarchical Bayesian Model

Full specification of the model is given by

$$y_i|\nu_i, \underline{\beta}, n_i \sim Bin(n_i, logit^{-1}(\underline{x_i}'\underline{\beta} + \nu_i)), i = 1, \ldots, L \tag{5.1}$$

$$\nu_i|\nu_{i-1}, \gamma, \sigma^2 \sim N(\gamma\nu_{i-1}, \sigma^2), i = 1, \ldots, L, \nu_0|\gamma, \tau \sim N(0, \frac{\sigma^2}{1-\gamma^2}) \tag{5.2}$$

$$\underline{\beta} \sim N(\underline{\theta}, \Sigma) \tag{5.3}$$

$$\gamma, \rho \overset{iid}{\sim} U(0,1) \tag{5.4}$$

$$\pi(\sigma^2) \propto \frac{1}{(1+\sigma^2)^2}. \tag{5.5}$$

$$\underline{z_{ij}}|\underline{\mu_i}, \rho, \sigma_L^2, \sigma_W^2 \sim N(\underline{\mu_i}, \Sigma) \tag{5.6}$$

$$\begin{aligned} \mu_{i,L}|\mu_{i-1,L}, \gamma_L, \delta_L^2 \sim N(\gamma_L\nu_{i-1,L}, \delta_L^2), i = 1, \ldots, L, \\ \mu_{0,L}|\gamma_L, \delta_L^2 \sim N(0, \delta_L^2(1-\gamma_L^2)^{-1}) \end{aligned} \tag{5.7}$$

$$\begin{aligned} \mu_{i,W}|\mu_{i-1,W}, \gamma_W, \delta_W^2 \sim N(\gamma_W\nu_{i-1,W}, \delta_W^2), i = 1, \ldots, L, \\ \mu_{0,W}|\gamma_W, \delta_W^2 \sim N(0, \delta_W^2(1-\gamma_W^2)^{-1}) \end{aligned} \tag{5.8}$$

$$\delta_L^2 = \frac{1-\epsilon_L}{\epsilon_L}\sigma_L^2 \tag{5.9}$$

$$\delta_W^2 = \frac{1-\epsilon_W}{\epsilon_W}\sigma_W^2 \tag{5.10}$$

$$\sigma_L^2, \sigma_W^2 \overset{iid}{\sim} (1+\sigma^2)^{-2} I_{[0,\infty)}(\sigma^2) \tag{5.11}$$

$$\rho, \epsilon_L, \epsilon_W, \gamma_L, \gamma_W \overset{iid}{\sim} U(0,1) \tag{5.12}$$

where

$$\Sigma = \begin{bmatrix} \sigma_L^2 & \rho\sigma_L\sigma_W \\ \rho\sigma_L\sigma_W & \sigma_W^2 \end{bmatrix} \tag{5.13}$$

and

$$\underline{x_{13}} = \frac{1}{n_{13}} \sum_{j=1}^{n_13} \beta_1 + L_{13,j}\beta_2 + W_{13,j}\beta_3 + 13\beta_4 \tag{5.14}$$

5.2 Full Posterior Density for Hierarchical Model

$$\begin{aligned} \tau_L = \sigma_L^{-2}, \tau_W &= \sigma_W^{-2} \\ \phi_i &= \underline{x_i}'\underline{\beta} + \nu_i, i = 0, \ldots, L \\ \tau &= \frac{1}{\sigma^2} \end{aligned} \tag{5.15}$$

where $\underline{x_0} = \underline{x_1}$ for convenience. The absolute value of the determinant of the Jacobian for this transformation is 1.

Then the full posterior density of the model is given by

$$
\begin{aligned}
\pi(\underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{\mu}, \underline{\gamma}, \underline{\tau}, \rho, \underline{\epsilon}, \underline{z_{13}} | \underline{z_{-13}, \underline{y}}) \propto \\
\prod_{i,j} N_{z_{ij}}(\underline{\mu_i}, \Sigma) \\
\times \prod_{i=1}^{L} \delta_L^{-1} \exp(-\frac{1}{2\delta_L^2}(\mu_{i,L} - \gamma_L \mu_{i,L})^2) \\
\times((1-\gamma_L^2)\delta_L^{-2})^{\frac{1}{2}} \exp(-\frac{1-\gamma_L^2}{2\delta_L^2}\mu_{0,L}^2) \\
\times \prod_{i=1}^{L} \delta_W^{-1} \exp(-\frac{1}{2\delta_W^2}(\mu_{i,W} - \gamma_W \mu_{i,W})^2) \\
\times((1-\gamma_W^2)\delta_W^{-2})^{\frac{1}{2}} \exp(-\frac{1-\gamma_W^2}{2\delta_W^2}\mu_{0,W}^2) \\
\times(1+\sigma_L^2)^{-2} I_{[0,\infty)}(\sigma_L^2)(1+\sigma_W^2)^{-2} I_{[0,\infty)}(\sigma_W^2) \\
\times U_\rho(0,1) U_{\gamma_L}(0,1) U_{\gamma_W}(0,1) U_{\epsilon_L}(0,1) U_{\epsilon_W}(0,1) \\
\times \prod_{i=1}^{L} Bin_{y_i}(n_i, logit^{-1}(\underline{x_i}'\underline{\beta})) \\
\times \prod_{i=1}^{L} \tau^{\frac{1}{2}} \exp(-\frac{\tau}{2}((\phi_i - \underline{x_i}'\underline{\beta}) - (\phi_{i-1} - \underline{x_{i-1}}'\underline{\beta}))^2) \\
\times((1-\gamma^2)\tau)^{\frac{1}{2}} \exp(-\frac{(1-\gamma^2)\tau}{2}((\phi_0 - \underline{x_0}'\underline{\beta})^2)) \\
\times N_{\underline{\beta}}(\underline{\theta}, \Sigma) \frac{1}{(1+\tau)^2} U_\gamma(0,1) U_\rho(0,1)
\end{aligned}
\qquad (5.16)
$$

5.3 Gibbs Sampler for Hierarchical Model

The Gibbs sampler for the complete model is then

$$\begin{aligned}\pi(\underline{z_{13,j}}|\underline{\beta},\underline{\phi},\gamma,\tau,\underline{y},\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\underline{\epsilon},\underline{z_{-13}}) &\overset{iid}{\propto} N(\underline{\mu_{13}},\Sigma)\\ \times Bin_{y_{13}}(\frac{1}{n_{13}}\sum_{j=1}^{n_13}\beta_1+L_{13,j}\beta_2+W_{13,j}\beta_3+13\beta_4)\end{aligned}\tag{5.17}$$

$$\begin{aligned}\pi(y_{13}|\underline{\beta},\underline{\phi},\gamma,\tau,\underline{y_{-13}},\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\underline{\epsilon},\underline{z}) \propto\\ Bin(\frac{1}{n_{13}}\sum_{j=1}^{n_13}\beta_1+L_{13,j}\beta_2+W_{13,j}\beta_3+13\beta_4)\end{aligned}\tag{5.18}$$

$$\begin{aligned}\pi(\underline{\beta}|\underline{\phi},\gamma,\tau,\underline{y},\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}})\\ = N((\Sigma^{-1}+\tau\sum(\underline{x_i}-\gamma\underline{x_{i-1}})\otimes(\underline{x_i}-\gamma\underline{x_{i-1}})),\\ (\Sigma^{-1}+\tau\sum(\underline{x_i}-\gamma\underline{x_{i-1}})\otimes(\underline{x_i}-\gamma\underline{x_{i-1}}))\\ (\Sigma^{-1}\underline{\theta}+\tau\sum(\phi_i-\gamma\phi_{i-1})(\underline{x_i}-\gamma\underline{x_{i-1}})))\end{aligned}\tag{5.19}$$

$$\begin{aligned}\pi(\tau|\underline{\beta},\underline{\phi},\gamma,\underline{y},\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) = Ga(\frac{L+3}{2},\frac{s^2}{2}),\\ s^2=\sum(\nu_i-\gamma\nu_{i-1})^2+(1-\gamma^2)\nu_0^2\end{aligned}\tag{5.20}$$

$$\pi(\phi_0|\underline{\beta},\underline{\phi_{-0}},\gamma,\underline{y},\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) = N(\gamma\phi_1,\tau^{-1})\tag{5.21}$$

$$\begin{aligned}\pi(\phi_i|\underline{\beta},\underline{\phi_{-i}},\gamma,\underline{y},\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) \propto\\ Bin_{y_i}(n_i,logit^{-1}(\underline{x_i}'\underline{\beta}))\\ \times N(\underline{x_i}'\underline{\beta}+\frac{\gamma}{1+\gamma^2}(\phi_{i+1}+\phi_{i-1}-\underline{x_{i+1}}'\underline{\beta}-\underline{x_{i-1}}'\underline{\beta},\tau^{-1})),i=1,\ldots,L-1\end{aligned}\tag{5.22}$$

$$\begin{aligned}\pi(\phi_L|\underline{\beta},\underline{\phi_{-L}},\gamma,\underline{y},\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}}) \propto\\ Bin_{y_i}(n_L,logit^{-1}(\underline{x_L}'\underline{\beta}))N(\gamma\phi_{L-1},\tau^{-1})\end{aligned}\tag{5.23}$$

$$\begin{aligned}\pi(\gamma|\underline{\beta}, \underline{\phi}, \tau, \underline{y}, \underline{\mu}, \underline{\gamma}, \underline{\tau}, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}) \propto \sqrt{1-\gamma^2}N_{\gamma}(\mu, s^2)I_{(0,1)}(\gamma),\\ \mu = \frac{\sum_{i=1}^{L} \nu_i \nu_{i-1}}{\sum_{i=1}^{L} \nu_i^2},\\ s^2 = \frac{1}{\tau \sum_{i=1}^{L} \nu_i^2}\end{aligned} \tag{5.24}$$

$$\begin{aligned}\pi(\gamma_L|\underline{\mu}, \gamma_W, \underline{\tau}, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y}) \propto\\ \sqrt{1-\gamma_L^2}N_{\gamma_L}(\theta, s^2)I_{(0,1)}(\gamma_L),\\ \theta = \frac{\sum_{i=1}^{L} \mu_{i,L}\mu_{i-1,L}}{\sum_{i=1}^{L} \mu_{i,L}^2},\\ s^2 = \frac{1}{\tau_L \sum_{i=1}^{L} \mu_{i,L}^2}\end{aligned} \tag{5.25}$$

$$\begin{aligned}\pi(\gamma_W|\underline{\mu}, \gamma_L, \underline{\tau}, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y}) \propto\\ \sqrt{1-\gamma_W^2}N_{\gamma_W}(\theta, s^2)I_{(0,1)}(\gamma_W),\\ \theta = \frac{\sum_{i=1}^{L} \mu_{i,W}\mu_{i-1,W}}{\sum_{i=1}^{L} \mu_{i,W}^2},\\ s^2 = \frac{1}{\tau_W \sum_{i=1}^{L} \mu_{i,W}^2}\end{aligned} \tag{5.26}$$

$$\begin{aligned}\pi(\rho|\underline{\mu}, \underline{\gamma}, \underline{\tau}, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y}) \propto\\ I_{[0,1]}(\rho)(1-\rho^2)^{-\frac{1}{2}\sum n_i} \exp(-\frac{1}{2}\sum_{ij} \underline{z_{ij}}'\Sigma^{-1}\underline{z_{ij}})\end{aligned} \tag{5.27}$$

$$\begin{aligned}\pi(\epsilon_L^*|\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\epsilon_W,\underline{z_{13}},\underline{z_{-13}},\underline{\beta},\underline{\phi},\gamma,\tau,\underline{y}) &\propto \\ (1+\epsilon_L^*)^{-2}I_{[0,\infty)}(\epsilon_L^*) &\times Ga(\frac{L+3}{2},s^2), \\ s^2=\frac{\tau_L}{2}(\sum_{i=1}^{L}(\mu_{i,L}-\gamma_L\mu_{i-1,L})^2 &+ (1-\gamma_L^2)\mu_{0,L}^2), \\ \epsilon_L &= \frac{\epsilon_L^*}{1+\epsilon_L^*}\end{aligned} \tag{5.28}$$

$$\begin{aligned}\pi(\epsilon_W^*|\underline{\mu},\underline{\gamma},\underline{\tau},\rho,\epsilon_L,\underline{z_{13}},\underline{z_{-13}},\underline{\beta},\underline{\phi},\gamma,\tau,\underline{y}) &\propto \\ (1+\epsilon_W^*)^{-2}I_{[0,\infty)}(\epsilon_W^*) &\times Ga(\frac{L+3}{2},s^2), \\ s^2=\frac{\tau_W}{2}(\sum_{i=1}^{L}(\mu_{i,W}-\gamma_W\mu_{i-1,W})^2 &+ (1-\gamma_W^2)\mu_{0,W}^2), \\ \epsilon_W &= \frac{\epsilon_W^*}{1+\epsilon_W^*}\end{aligned} \tag{5.29}$$

$$\begin{aligned}\pi(\tau_L|\underline{\mu},\underline{\gamma},\tau_W,\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}},\underline{\beta},\underline{\phi},\gamma,\tau,\underline{y}) &\propto \\ (1+\tau_L)^{-2}I_{[0,\infty)}(\tau_L)\exp(-\frac{s^2}{2}\tau_L) \\ \times\exp(\frac{\rho}{1-\rho^2}\sum_{ij}\ln(L_{ij})\ln(W_{ij})\tau_L^{\frac{1}{2}})\end{aligned} \tag{5.30}$$

$$\begin{aligned}\pi(\tau_W|\underline{\mu},\underline{\gamma},\tau_L,\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}},\underline{\beta},\underline{\phi},\gamma,\tau,\underline{y}) &\propto \\ (1+\tau_W)^{-2}I_{[0,\infty)}(\tau_W)\exp(-\frac{s^2}{2}\tau_W) \\ \times\exp(\frac{\rho}{1-\rho^2}\sum_{ij}\ln(L_{ij})\ln(W_{ij})\tau_W^{\frac{1}{2}})\end{aligned} \tag{5.31}$$

$$\pi(\mu_{0,L}|\underline{\mu_{-0}},\mu_{0,W},\underline{\gamma},\tau_L,\rho,\underline{\epsilon},\underline{z_{13}},\underline{z_{-13}},\underline{\beta},\underline{\phi},\gamma,\tau,\underline{y}) \propto N(\gamma_L\mu_{1,L},\tau^{-1}) \tag{5.32}$$

$$\pi(\mu_{0,W}|\underline{\mu_{-0}}, \mu_{0,L}, \underline{\gamma}, \tau_W, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y}) \propto N(\gamma_W \mu_{1,W}, \tau^{-1}) \tag{5.33}$$

$$\begin{aligned}
\pi(\mu_{i,L}|\underline{\mu_{-i}}, \mu_{i,W}, \underline{\gamma}, \tau_L, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y})& \\
\propto N(\eta_{i,L}, \sigma_L^2(1 + n_i(1-\rho^2)^{-1} + \gamma_L^2)^{-1}),& \\
\eta_{i,L} = (\gamma_L(\mu_{i+1,L} + \mu_{i-1,L}) + (1-\rho^2)^{-1}\sum_j \ln(L_{i,j}) - \rho(1-\rho^2)^{-1}& \\
\times \sigma_L \sigma_W^{-1} \sum_j (\ln(W_{i,j}) - \mu_{i,W}))& \\
\times \sigma_L^2(1 + n_i(1-\rho^2)^{-1} + \gamma_L^2)^{-1}, 0 < i < L&
\end{aligned} \tag{5.34}$$

$$\begin{aligned}
\pi(\mu_{i,W}|\underline{\mu_{-i}}, \mu_{i,L}, \underline{\gamma}, \tau_W, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y})& \\
\propto N(\eta_{i,W}, \sigma_W^2(1 + n_i(1-\rho^2)^{-1} + \gamma_W^2)^{-1}),& \\
\eta_{i,W} = (\gamma_W(\mu_{i+1,W} + \mu_{i-1,W}) + (1-\rho^2)^{-1}\sum_j \ln(W_{i,j}) - \rho(1-\rho^2)^{-1}& \\
\times \sigma_W \sigma_L^{-1} \sum_j (\ln(L_{i,j}) - \mu_{i,W}))& \\
\times \sigma_W^2(1 + n_i(1-\rho^2)^{-1} + \gamma_W^2)^{-1}, 0 < i < L&
\end{aligned} \tag{5.35}$$

$$\begin{aligned}
\pi(\mu_{L,L}|\underline{\mu_{-L}}, \mu_{L,L}, \underline{\gamma}, \tau_L, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y})& \\
\propto N(\eta_{L,L}, \sigma_L^2(1 + n_L(1-\rho^2)^{-1})^{-1}),& \\
\eta_{L,L} = (\gamma_L \mu_{L-1,L} + (1-\rho^2)^{-1}\sum_j \ln(L_{i,j}) - \rho(1-\rho^2)^{-1}\sigma_L\sigma_W^{-1}& \\
\times \sum_j (\ln(W_{L,j}) - \mu_{L,W}))& \\
\times \sigma_L^2(1 + n_i(1-\rho^2)^{-1})^{-1}, 0 < i < L&
\end{aligned} \tag{5.36}$$

$$
\begin{aligned}
\pi(\mu_{L,W} | \underline{\mu_{-L}}, \mu_{L,W}, \underline{\gamma}, \tau_W, \rho, \underline{\epsilon}, \underline{z_{13}}, \underline{z_{-13}}, \underline{\beta}, \underline{\phi}, \gamma, \tau, \underline{y}) \\
\propto N(\eta_{L,W}, \sigma_W^2(1 + n_L(1-\rho^2)^{-1})^{-1}), \\
\eta_{L,W} = (\gamma_W \mu_{L-1,W} + (1-\rho^2)^{-1} \sum_j \ln(W_{i,j}) - \rho(1-\rho^2)^{-1}\sigma_W \sigma_L^{-1} \\
\times \sum_j (\ln(L_{L,j}) - \mu_{L,L})) \\
\times \sigma_W^2 (1 + n_i(1-\rho^2)^{-1})^{-1}, 0 < i < L
\end{aligned}
\tag{5.37}
$$

5.4 Simulation Results

We now run the complete model with 101000 iterates, 1000 burn-in, sampling every 100 after burn-in. We list posterior quantities for the more important parameters, especially $p = logit^{-1}(\phi)$.

Once again, lag 1 autocorrelations show that the Gibbs sampler is converging, albeit slowly. τ, τ_L, and τ_W can in theory be unbounded, and sampling such quantities may cause problems. However, the time series plots of the post-burn-in iterates that are sampled show that those quantities behave reasonably well.

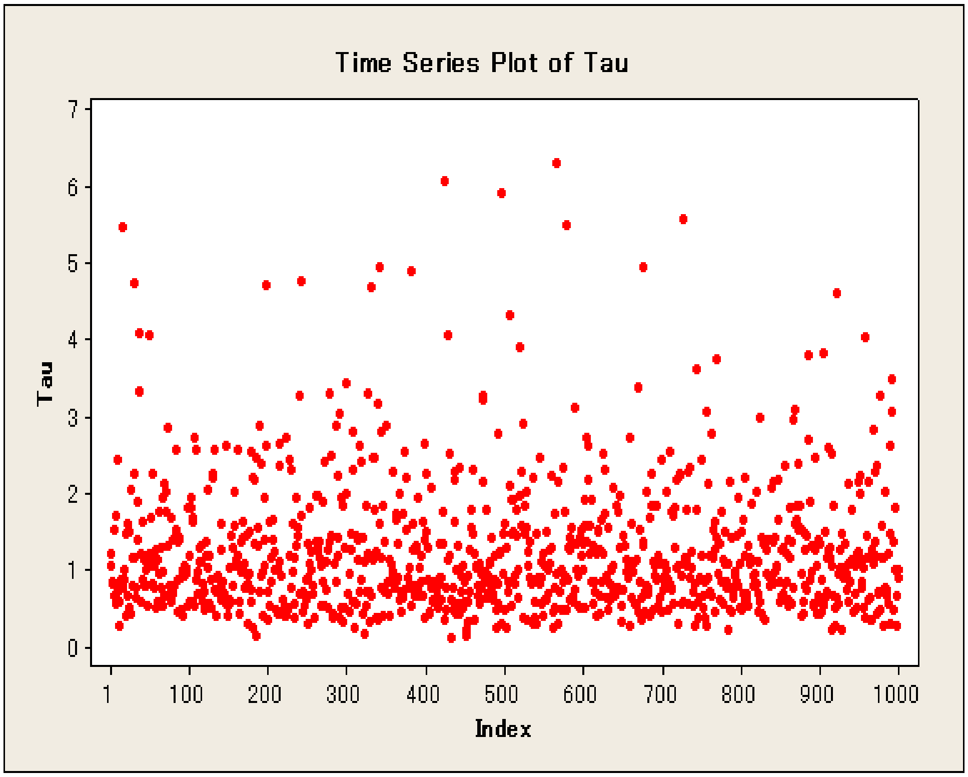

Figure 5.1: Time Series Plot of Sampled Tau

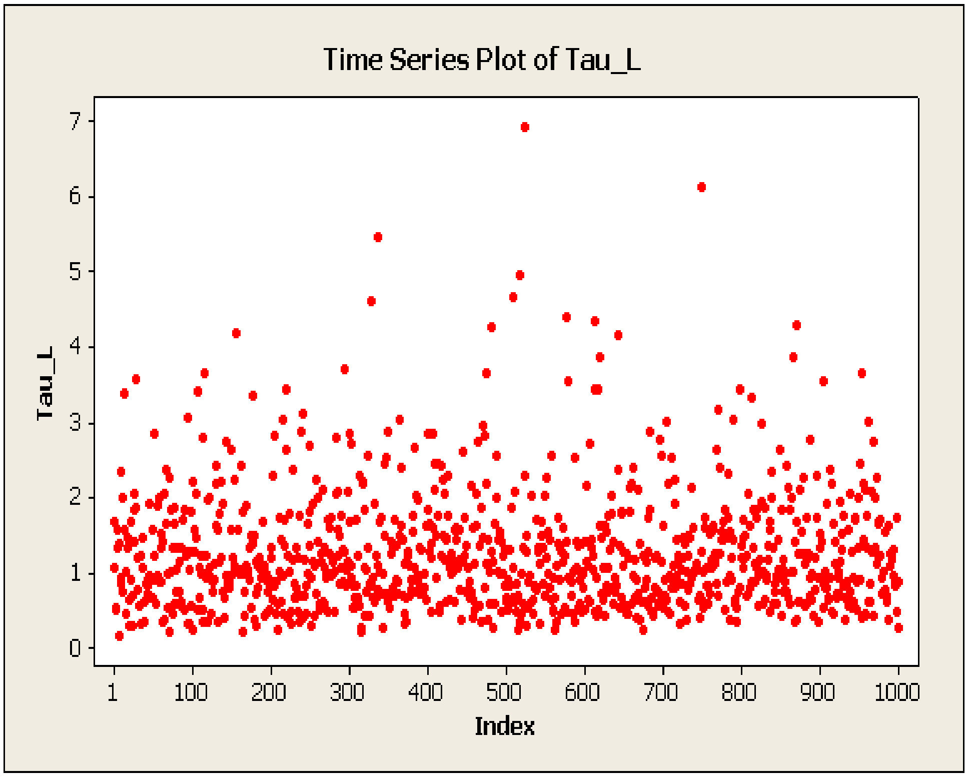

Figure 5.2: Time Series Plot of Sampled Tau L

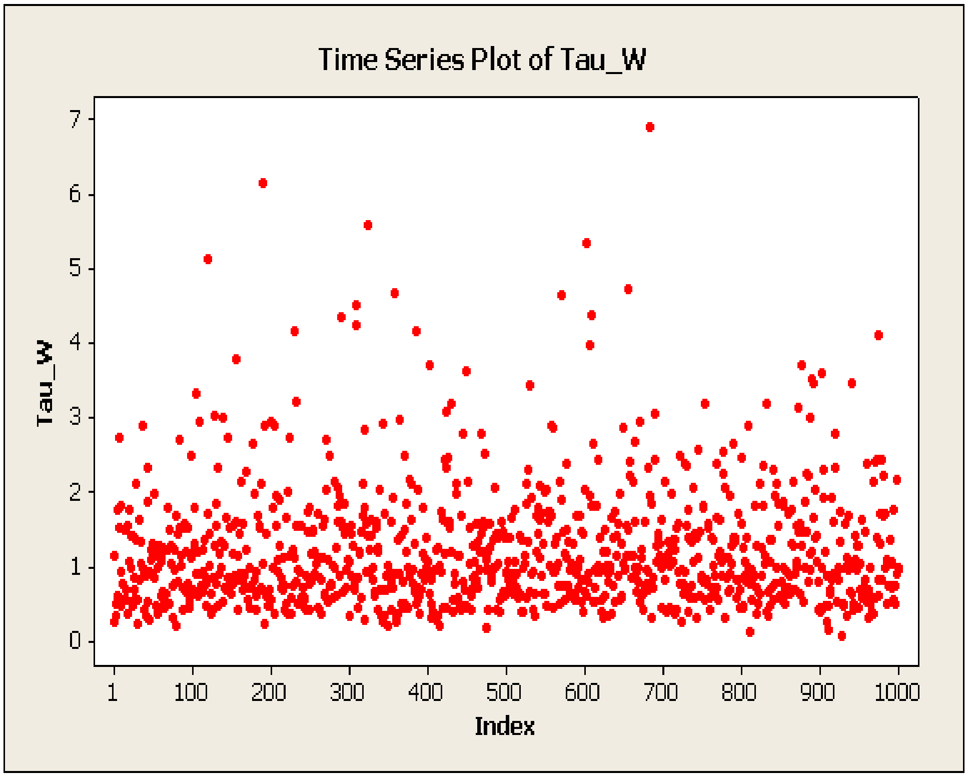

Figure 5.3: Time Series Plot of Sampled Tau W

	Mean	Std. Dev.	Num. Std. Err.	0.95 Cred. Int.
τ	1.24	0.857	0.0271	(0.300, 3.48)
β_1	2.06	4.31	0.136	(-6.98, 11.4)
β_2	0.249	1.62	0.0512	(-3.03, 3.42)
β_3	1.50	1.53	0.0484	(-1.60, 4.49)
β_4	-0.95	1.03	0.0326	(-2.37, -0.038)
p_1	0.940	0.100	0.0032	(0.616, 0.995)
p_2	0.984	0.045	0.0014	(0.935, 0.999)
p_3	0.862	0.029	0.0009	(0.643, 0.982)
p_4	0.808	0.110	0.0035	(0.448, 0.920)
p_5	0.870	0.091	0.0029	(0.645, 0.966)
p_6	0.933	0.073	0.0023	(0.497, 0.994)
p_7	0.796	0.114	0.0036	(0.509, 0.937)
p_8	0.778	0.147	0.0046	(0.474, 0.924)
p_9	0.679	0.138	0.0044	(0.485, 0.870)
p_{10}	0.940	0.077	0.0024	(0.630, 0.992)
p_{11}	0.708	0.190	0.0060	(0.326, 0.909)
p_{12}	0.666	0.140	0.0044	(0.358, 0.898)
p_{13}	0.721	0.195	0.0062	(0.427, 0.871)
p_{14}	0.691	0.180	0.0057	(0.414, 0.864)
τ_L	1.26	0.797	.0252	(0.423,5.933)
$lnL_{13,1}$	3.753	2.027	0.064	(3.638,3.942)
$lnL_{13,2}$	3.885	1.771	0.056	(3.761,3.932)
$lnL_{13,3}$	3.951	2.153	0.068	(3.882,3.988))
$lnL_{13,4}$	3.777	2.261	0.071	(3.696,3.824)
$lnL_{13,5}$	3.769	2.026	0.064	(3.761,3.811)
$lnL_{13,6}$	3.683	1.903	0.060	(3.653,3.697)
τ_W	1.22	0.814	.0257	(0.713,6.217)
$lnW_{13,1}$	7.118	3.487	0.110	(6.659,7.244)
$lnW_{13,2}$	6.340	3.612	0.114	(6.194,6.791)
$lnW_{13,3}$	6.741	3.505	0.111	(6.397,6.908)
$lnW_{13,4}$	6.833	4.110	0.129	(6.522,6.802)
$lnW_{13,5}$	6.968	3.713	0.117	(5.991,7.783)
$lnW_{13,6}$	6.592	3.100	0.098	(6.297,6.820)

Table 5.1: Gibbs Sampler Posterior Quantities after burn-in

Parameter	AC
τ	0.0411
β_1	-0.0228
β_2	0.0025
β_3	-0.00398
β_4	-0.0523
p_1	-0.0086
p_2	-0.0325
p_3	-0.0067
p_4	0.0291
p_5	0.0451
p_6	-0.103
p_7	-0.0270
p_8	0.0183
p_9	-0.00811
p_{10}	-0.0139
p_{11}	-0.0157
p_{12}	0.0304
p_{13}	-0.095
p_{14}	0.0082
τ_L	-0.100
$\ln L_{13,1}$	-0.114
$\ln L_{13,2}$	0.0421
$\ln L_{13,3}$	0.0069
$\ln L_{13,4}$	-0.0053
$\ln L_{13,5}$	-0.0028
$\ln L_{13,6}$	0.0019
τ_W	0.0034
$\ln W_{13,1}$	-0.203
$\ln W_{13,2}$	-0.091
$\ln W_{13,3}$	-0.0072
$\ln W_{13,4}$	0.0036
$\ln W_{13,5}$	0.0012
$\ln W_{13,6}$	0.099

Table 5.2: Gibbs sampling autcorrelation after burn-in

Chapter 6

Conclusions

In this paper, we have presented a parsimonious Bayesian logistic regression model which incorporates spatial effects, following the spirit of generalized linear models. The count of toxic catfish at each location is binomial given the parameters, where the probability that a catfish is toxic depends, through the logistic link function, on an affine combination of the covariates (average length and weight of catfish, and its location) and on the spatial effect specific to the location. The spatial effects are modeled by an AR(1) process with positive correlation. A Gibbs sampler was constructed from the full posterior density specified by the model to estimate posterior quantities of interest. This model offers the advantage that it is easier to interpret than a frequentist logistic regression model and describes the data better than a Bayesian logistic regression model with uncorrelated spatial random effects.

To extend the utility of this model, we developed a model to impute length and weight of catfish from locations where data was not collected. The imputation model treats unobserved (log-transformed) lengths and weights as parameters on their own right that are generated by the same process as observed lengths and weights. The length and weight of a catfish of a given location is assumed to be from a bivariate normal distribution, and the mean value of log-transformed length and weight, respectively, are treated as AR(1) processes, with the length- and weigh- means independent of each other. This model is then used to impute weight and length measurements for 6 unobserved fish from location TRM335. Using the imputed values, we fit our Bayesian spatial effects to the augmented dataset.

Finally, we combined the models of chapter 3 and 4 so as to treat imputation and prediction simultaneously and without any other ad hoc procedures. Using the complete model, we have point and interval estimates for the probability that a catfish at any location between TRM275 and TRM340 contains 5 ppm or more of DDT in their filet.

	Mean	Std. Dev.	Num. Std. Err.	0.95 Cred. Int.
p_1	0.940	0.100	0.0032	(0.616, 0.995)
p_2	0.984	0.045	0.0014	(0.935, 0.999)
p_3	0.862	0.029	0.0009	(0.643, 0.982)
p_4	0.808	0.110	0.0035	(0.448, 0.920)
p_5	0.870	0.091	0.0029	(0.645, 0.966)
p_6	0.933	0.073	0.0023	(0.497, 0.994)
p_7	0.796	0.114	0.0036	(0.509, 0.937)
p_8	0.778	0.147	0.0046	(0.474, 0.924)
p_9	0.679	0.138	0.0044	(0.485, 0.870)
p_{10}	0.940	0.077	0.0024	(0.630, 0.992)
p_{11}	0.708	0.190	0.0060	(0.326, 0.909)
p_{12}	0.666	0.140	0.0044	(0.358, 0.898)
p_{13}	0.721	0.195	0.0062	(0.427, 0.871)
p_{14}	0.691	0.180	0.0057	(0.414, 0.864)

Table 6.1: Posterior probabilities that a catfish from a location is toxic

The major finding is that including the imputed values does not alter the inference significantly. Under the Bayesian statistical models considered, the 95 percent credible intervals of all components of the regression vector $\underline{\beta}$ besides location contain zero and are not important. Even the frequentist logistic regression models gave $\underline{\beta}$ confidence intervals containing zero. Thus, under all of these circumstances, catfish data indicates that only location is a significant factor in determining proportion of toxic catfish.

Many opportunities exist to extend the results of this paper. We can explore the behavior of catfish farther downstream. There may be ways to treat the spatial effect of location zero, which corresponds to TRM270 (for which we have no data), from the effects of catfish that actually live there. What happens when there is no longer a unidirectional flow as in a river? Can these results be extended to model such situations? As with any observational study, we may want to do sensitivity analysis to determine if there are hidden covariates that might improve our understanding of the data (Rosenbaum, 2002).

Appendix A

Dataset

Table A.1 and A.2 contain information for each fish caught: toxic or not (5 ppm or more of DDT in filet), length, and weight. The data was collected along the Tennessee River, Alabama, during the summer of 1980.

Location	1	2	3	4	5	6	7
y_1	1	1	1	1	1	1	1
L_1	48.0	48.0	44.0	41.0	36.0	36.0	35.0
W_1	986	1048	936	961	980	847	613
y_2	1	1	1	1	1	1	1
L_2	45.0	51.0	42.0	44.0	47.5	37.0	51.0
W_2	1023	1641	1058	886	1176	876	353
y_3	1	1	1	1	0	0	1
L_3	49.0	48.5	42.5	41.0	41.5	35.0	42.5
W_3	1266	1331	800	678	989	844	909
y_4	1	1	1	0	1	1	1
L_4	50.0	51.0	45.5	42.0	49.5	36.0	38.0
W_4	1086	1728	1087	1011	1084	908	886
y_5	1	1	1	1	1	1	1
L_5	46.0	44.0	48.0	42.5	46.0	48.0	41.0
W_5	1044	917	1329	947	1115	1358	890
y_6	1	1	0	1	1	1	1
L_6	52.0	51.0	44.0	44.0	46.5	49.0	47.0
W_6	1770	1398	897	989	724	1019	1031

Table A.1: Channel Catfish Data (Locations 1 through 7)

Location	8	9	10	11	12	13	14
y_1	1	1	1	1	1		1
L_1	45.0	48.0	47.5	46.0	36.0		50.0
W_1	1083	476	983	863	556		1207
y_2	0	0	1	1	1		1
L_2	45.5	29.5	51.5	40.0	42.0		45.0
W_2	864	743	1251	549	659		911
y_3	1	0	1	0	1		0
L_3	45.0	42.0	49.5	43.5	40.5		49.0
W_3	886	1128	1255	810	1229		1498
y_4	0	1	1	0	0		0
L_4	45.0	47.5	47.0	46.5	51.5		39.5
W_4	965	848	1152	908	1050		1496
y_5	1	1	1	0	1		0
L_5	39.0	47.5	47.5	43.0	47.0		50.0
W_5	537	1091	1085	804	952		1142
y_6	1	1	1	1	0		0
L_6	40.5	43.5	47.0	47.5	41.0		45.0
W_6	630	715	1118	1179	826		879

Table A.2: Channel Catfish Data (Locations 8 through 14)

Appendix B

Accept-Reject Sampling

The accept-reject sampling method allows one to draw samples from density functions for which sampling procedure is not known (Casella and Berger (2002)).

Let $f_Y(y)$ be the target density from which samples are to be drawn, and let $f_V(v)$ have the same support as the target density and such that $M = \sup \frac{f_Y(y)}{f_V(y)} < \infty$.

```
Drawing $Y \sim f_Y(y)$
(1) Sample $U \sim U(0,1)$, $V \sim f_V$.
(2) Calculate $prob = \frac{1}{M} \frac{f_Y(V)}{f_V(V)}$
(3) If $U > prob$
    then return $V$
    else return to (1)
```

Figure B.1: Sampling with Accept-Reject method

Appendix C

Gibbs Sampling

Geman and Geman (1984) made the Gibbs sampler widely known in the context of image restoration, but the Gibbs sampler had been implemented in various guises during the 1970s in fields including statistical mechanics and spatial statistics. Gelfand and Smith (1990) caused an explosion in applications of Bayesian statistics by demonstrating how the algorithm of Geman and Geman can be used to sample from the general continuous posterior distributions typically found in Bayesian models.

Let $\underline{\theta} = (\underline{\theta_1}, \ldots, \underline{\theta_p})$ be the vector of parameters organized into p blocks. We would like to draw from $\pi(\underline{\theta}|\underline{y})$. The Gibbs sampler algorithm is given as follows (Chen, Shao and Ibrahim (2000)):

Tierney (1994) develops the relationships between Markov chains and Monte Carlo methods; See Cowles and Carlin (1996) for a comprehensive review of convergence diagnostics of Gibbs samplers and other Markov chain Monte Carlo methods.

```
GIBBS SAMPLING
(1) SET initial values $\underline{\theta_0}$.
(2) FOR each i=1,2,..., repeat the following.
    SAMPLE $\underline{\theta_1}^{(i)} \sim \pi(\underline{\theta_1}^{(i)}|\underline{\theta_2}^{(i-1)}, \underline{\theta_2}^{(i-1)}, \ldots, \underline{\theta_p}^{(i-1)}, \underline{y})$
$\underline{\theta_2}^{(i)} \sim \pi(\underline{\theta_2}^{(i)}|\underline{\theta_1}^{(i)}, \underline{\theta_3}^{(i-1)}, \ldots, \underline{\theta_p}^{(i-1)}, \underline{y})$
$\underline{\theta_j}^{(i)} \sim \pi(\underline{\theta_j}^{(i)}|, \ldots, \underline{\theta_{j-1}}^{(i)}, \underline{\theta_{j+1}}^{(i-1)}, \ldots, \underline{\theta_p}^{(i-1)}, \underline{y})$
$\underline{\theta_p}^{(i)} \sim \pi(\underline{\theta_p}^{(i)}|\underline{\theta_1}^{(i)}, \ldots, \underline{\theta_{p-1}}^{(i)}, \underline{y})$
(3) RECORD $\underline{\theta}^{(i)}$, $i = N, N+1, \ldots$ for N large.
```

Figure C.1: Construction of Gibbs Samplers

Appendix D

AR(1) Models

A discrete time stochastic process $\{y_t\}, t = 1, \dots, n$ is called an AR(1) process, AR standing for autoregressive, if

$$\begin{aligned} y_t = \phi y_{t-1} + \epsilon_t, \epsilon_t \sim i.i.d.N(0, \sigma^2), t = 1, \dots, n \\ y_0 \sim N(0, \frac{\sigma^2}{1-\phi^2}), |\phi| < 1. \end{aligned} \tag{D.1}$$

Because all stochastic processes are characterized by its finite-dimensional distributions according Kolmogorov's theorem (see Billingsley (1994) for theory), one may be interested in the joint density of $(y_0, \dots, y_n)$, i.e., the joint density of y up through time $t = n$. This density is multivariate normal, given by

$$f(y_0, \dots, y_n) = N(\underline{\mu}, \Sigma) \tag{D.2}$$

where $(\Sigma_{ij}) = \sigma^{-2}(1 + \delta^i_j \phi^2)$ (δ is Kronecker's delta), and $\mu = \phi \Sigma^{-1}(y_1, \dots, y_n, 0)'$. Under this model, given ϕ, $Cor(y_s, y_t) = \phi^{|s-t|}$.

In the Bayesian setting, a prior distribution for ϕ and σ^2, $\pi(\phi, \sigma^2)$, needs to be specified.

Bibliography

[1] Banerjee, S., Carlin, B., and Gelfand, A. (2004), Hierarchical modeling and analysis for spatial data: Chapman and Hall.

[2] Billingsley, P. (1994), Probability and Measure (3rd ed.), New York: Wiley-Interscience

[3] Box, G., and Tiao, G. (1973), Bayesian Inference in Statistical Analysis, Reading: Addison Wesley.

[4] Casella, G., and Berger, R.L. (2002), Statistical Inference (2nd ed.), Duxbury.

[5] Chen, M.H., Shao, Q.M., and Ibrahim, J.G. (2000), Monte Carlo Methods in Bayesian Computation, New York: Springer.

[6] Cowles, M.K., and Carlin, B.P. (1995), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," Journal of the American Statistical Association, 91, 434, 883 - 904.

[7] Cressie, N. (1993), Statistics for Spatial Data (Revised ed.), New York: Wiley-Interscience.

[8] Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," Bayesian Analysis, 1(3), 515-534.

[9] Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images", IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(6), 721-741.

[10] Gelman, A., Roberts, G., and Gilks, W. (1995). "Efficient Metropolis jumping rules." In Bayesian Statistics 5, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford University Press.

[11] Lemeshow, S., and Hosmer, D., (1989), Applied Logistic Regression, New York: John Wiley and Sons.

[12] Jank, W., and Kannan, P.K. (2005), "Understanding Geographical Markets of Online Firms Using Spatial Models of Customer Choice," Marketing Science, 24, 4, 623-634.

[13] Rosenbaum, P. R. (2002), Observational Studies, New York: Springer.

[14] Shumway, R.H., and Stoffer, D.S. (2006), Time Series Analysis and Its Applications, New York: Springer.

[15] Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," Annals of Statistics, 22, 1701 - 1762.

[16] US Army Corps of Engineers Home Page, http://www.orn.usace.army.mil/.